Title
What to Maximize If You Must

Authors
Heifetz, Aviad
Shannon, Chris
Spiegel, Yossi

Publication Date
2002-12-01

Peer reviewed

UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720-3880

Working Paper No. E02-326

What to Maximize If You Must

Aviad Heifetz
Tel Aviv University

Chris Shannon
University of California, Berkeley

Yossi Spiegel
Tel Aviv University

December 2002

JEL Classification: C7

Abstract

The assumption that decision makers choose actions to maximize their preferences is a central tenet in economics. This assumption is often justified either formally or informally by appealing to evolutionary arguments. In contrast, this paper shows that in almost every game, payoff. maximization cannot be justified by appealing to such arguments. We show that in almost every game, for almost every distortion of a player's actual payoffs, some extent of this distortion is beneficial to the player because of the resulting effect on opponents' play. Consequently, such distortions will not be driven out by any evolutionary process involving payoff.- monotonic selection dynamics, in which agents with higher actual payoffs proliferate at the expense of less successful agents. In particular, under any such selection dynamics, the population will not converge to payoff-maximizing behavior. We also show that payoff-maximizing behavior need not prevail even when preferences are imperfectly observed.

What to Maximize if You Must*

Aviad Heifetz† Chris Shannon‡ Yossi Spiegel§

October 28, 2002

Abstract

The assumption that decision makers choose actions to maximize their preferences is a central tenet in economics. This assumption is often justified either formally or informally by appealing to evolutionary arguments. In contrast, this paper shows that in almost every game, payoff maximization cannot be justified by appealing to such arguments. We show that in almost every game, for almost every distortion of a player's actual payoffs, some extent of this distortion is beneficial to the player because of the resulting effect on opponents' play. Consequently, such distortions will not be driven out by any evolutionary process involving payoff-monotonic selection dynamics, in which agents with higher actual payoffs proliferate at the expense of less successful agents. In particular, under any such selection dynamics, the population will *not* converge to payoff-maximizing behavior. We also show that payoff-maximizing behavior need not prevail even when preferences are imperfectly observed.

1 Introduction

The assumption that decision makers choose actions to maximize their preferences is a central tenet in economics. 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 only firms behaving in a manner consistent with profit maximization will survive in the long run. Under this argument, firms failing to act so as to maximize profits will be driven out of the market

*We are grateful for valuable comments from Bill Zame, Eddie Dekel, Menachem Yaari, and participants of the 11th European Workshop in General Equilibrium Theory.

†The School of Economics, Tel Aviv University, heifetz@post.tau.ac.il

‡Department of Economics, University of California, Berkeley, cshannon@econ.berkeley.edu

§The Faculty of Management, Tel Aviv University, spiegel@post.tau.ac.il

by more profitable rivals, even if none of these firms deliberately maximizes profits or is even aware of its cost or revenue functions. Similar arguments that consumers behave "as if" maximizing preferences due to myriad market forces that exploit non-optimal behavior are pervasive. More recently, Sandroni (2000) gives such a justification for rational expectations equilibria, showing that a market populated by agents who initially differ in the accuracy of their predictions will nonetheless converge to a competitive rational expectations equilibrium as those agents who make inaccurate predictions are driven out of the market by those who are more accurate.

In contrast, this paper shows that in *almost every* strategic interaction, payoff maximization cannot be justified by appealing to evolutionary arguments. Specifically, we show that in almost every game, for almost every type of distortion of a player's actual payoffs, some extent of this distortion is beneficial to the player because of the resulting effect on opponents' play. Consequently, we show that such distortions will not be driven out by any evolutionary process involving payoff-monotonic selection dynamics, in which agents with higher actual payoffs proliferate at the expense of less successful agents. In particular, under any such selection dynamics, the population will *not* converge to payoff maximizing behavior.

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. Related ideas run through work ranging from Stackelberg's (1934) classic work on timing in oligopoly to the theories of reputation in Kreps and Wilson (1982), and Milgrom and Roberts (1982). For similar reasons, Frank (1987, 1988) argues that emotions may be a beneficial commitment device. Recently, a large and growing literature has emerged that formalizes some of these ideas by explicitly studying the evolution of preferences. This work shows that in strategic interactions, a wide array of distortions of actual payoffs, representing features such as altruism, spite, overconfidence, fairness, and reciprocity, that bias individuals' objectives away from actual payoff maximization, may be evolutionarily stable.[1]

Unlike most standard evolutionary game theory, in which individuals are essentially treated as "machines" programmed to play a specific action, the work on the evolution of preferences treats individuals as decision makers who choose actions to maximize their preferences, and then studies how the distribution of these preferences evolves over time. Preferences that are distortions of true payoffs – or "dispositions" – drive a wedge between an individual's objectives and actual payoffs. Dispositions may nonetheless be evolutionarily stable because the resulting bias in a player's objectives may induce favorable behavior in rivals that may more than compensate for the loss stemming from

[1]For a brief overview of this literature, see Samuelson (2001). Examples include Güth and Yaari (1992), Huck and Oechssler (1998), Fershtman and Weiss (1997, 1998), Rotemberg (1994), Bester and Güth (1998), Possajennikov (2000), Bolle (2000), Bergman and Bergman (2000), Koçkesen, Ok, and Sethi (2000a, 2000b), Guttmann (2000), Sethi and Somanathan (2001), Kyle and Wang (1997), Benos (1998), and Heifetz and Segev (2001).

departures from actual payoff maximization. Thus the literature on the evolution of preferences illustrates the point that in a variety of strategic interactions, individuals who fail to maximize their true payoffs due to the bias created by various dispositions may actually end up with higher payoffs than individuals who are unbiased. Such beneficial dispositions would then not be weeded out by any selection dynamics in which more successful behavior proliferates at the expense of less successful behavior, where success is measured in terms of actual payoffs.

Much of the work on the evolution of preferences, however, focuses on *specific* kinds of dispositions, such as altruism or reciprocity, and addresses these questions using *specific* functional forms for both the individuals' payoffs and dispositions. Such results then provide conditions on the parameters of the particular model at hand that guarantee that some non-zero degree of this disposition will survive evolutionary pressures. Our results generalize this work in an important way by showing that the evolutionary emergence of dispositions is in fact *generic*. In particular, we show that in almost any kind of strategic interaction and for almost any kind of disposition, having some degree of this disposition is better for a player than having no disposition at all. That is, in almost any game, some degree of almost any kind of disposition results in a higher equilibrium payoff than could be attained without the disposition. Any such disposition will not become extinct under any payoff-monotonic selection dynamics, and in any such setting the population will not converge to payoff-maximizing behavior under these selection dynamics.

Our genericity results are fairly intuitive. Having a disposition affects a player's payoff in two ways: directly, through the player's own actions, and indirectly, by influencing other players' actions. A crucial observation is that a small degree of disposition leads to a slight deviation from payoff-optimizing behavior, and therefore has only a negligible direct effect on the player's payoff. The crux of our argument is that for generic combinations of games and dispositions, the indirect effect on the player's payoff resulting from such a small degree of the disposition is *not* negligible. We first prove this result for a class of finite-dimensional manifolds of payoff and disposition functions, and then generalize it to the infinite-dimensional families of all payoff and disposition functions.

Our main results are derived under the assumption that players' preferences are perfectly observable. We then show that dispositions may remain evolutionarily viable even when the players' preferences can be only imperfectly observed. Here the natural solution concept given the imperfect observability of preferences is Bayesian equilibrium. This highlights a technical obstacle surrounding results about the evolutionary viability of dispositions. Unlike Nash equilibria with perfect observability, Bayesian equilibria are typically not locally unique (see, e.g., Leininger, Linhart, and Radner, 1989). In such cases an equilibrium selection is not well-defined even locally, and different selections from the equilibrium correspondence may result in contradictory conclusions regarding the effects of dispositions. While this precludes a general analysis of imperfect observability, in the context of an example with a unique Bayesian equilibrium we show that dispositions may remain evolutionarily viable even in the presence of imperfect observability. We con-

sider three alternative settings: first, that preferences are perfectly unobserved in some fraction of interactions while unobserved in others; second, that preferences are observed but with noise; and third, that the model involves costly signaling of preferences. We show that in all three settings, dispositions will not be eliminated for almost all parameter values in the model, and in the first and third settings we completely characterize the limiting distribution.

The paper proceeds as follows. Section 2 contains the development of our framework and our main results, showing generically that dispositions do not become asymptotically extinct under payoff-monotonic selection dynamics. We prove this result both in the case where the payoff and disposition functions vary over a particular class of finite-dimensional sets, as well as for the case where they vary over the infinite-dimensional set of all payoff and disposition functions. In Section 3 we illustrate these results by means of a specific example. In this example we can derive sharper predictions than under our general results. Here we fully characterize the asymptotic distribution of types, and show that in the limit distribution players will have dispositions. In Section 4 we relax the assumption that types are perfectly observed and consider the three alternatives involving imperfect observability outlined above. In each case we show that our main results carry over to these settings. All proofs are collected in the Appendix.

2 The genericity of dispositions

2.1 Payoffs and dispositions

Two players, i and j, engage in strategic interaction. The strategy spaces of the two players, X^i and X^j, are open subsets of $\mathbf{R}^M$ and $\mathbf{R}^N$, respectively, where, without loss of generality, $M \leq N$.[2] Typical strategies are denoted $x^i = (x^i_1, \ldots, x^i_M)$ and $x^j = (x^j_1, \ldots, x^j_N)$. The payoffs of the two players are given by the C^3 functions

$$\Pi^i, \Pi^j : X^i \times X^j \to \mathbf{R}.$$

In what follows we denote the partial derivatives of Π^i by

$$\Pi^i_i \equiv D_i \Pi^i = \left(\frac{\partial \Pi^i}{\partial x^i_1}, \ldots, \frac{\partial \Pi^i}{\partial x^i_M} \right) \quad \text{and} \quad \Pi^i_{ij} \equiv D_j \Pi^i_i = \begin{pmatrix} \frac{\partial^2 \Pi^i}{\partial x^i_1 \partial x^j_1} & \cdots & \frac{\partial^2 \Pi^i}{\partial x^i_1 \partial x^j_N} \\ & \ddots & \\ \frac{\partial^2 \Pi^i}{\partial x^i_M \partial x^j_1} & \cdots & \frac{\partial^2 \Pi^i}{\partial x^i_M \partial x^j_N} \end{pmatrix}.$$

The partial derivatives of Π^j and of other functions are denoted similarly.

[2]The restriction to two players is just for notational convenience; all of our results carry over directly for games with an arbitrary number of players. For games with more players and more general strategy sets, see Remarks 1 and 2 below.

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

$$\begin{aligned} U^i(x^i, x^j, \tau) &\equiv \Pi^i(x^i, x^j) + B^i(x^i, x^j, \tau), \\ U^j(x^i, x^j, \theta) &\equiv \Pi^j(x^i, x^j) + B^j(x^i, x^j, \theta), \end{aligned} \tag{2.1}$$

where

$$B^i, B^j : X^i \times X^j \times \mathbf{R} \to \mathbf{R}$$

are the dispositions of players i and j and τ and θ are the players' (one-dimensional) types. The introduction of dispositions then drives 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. We assume that B^i and B^j are C^3. Moreover, as a normalization we assume that when τ or θ is zero, the players' perceived payoffs coincide with their actual payoffs:

$$B^i(x^i, x^j, 0) \equiv B^j(x^i, x^j, 0) \equiv 0. \tag{2.2}$$

That is, a type 0 player has no disposition and simply chooses actions to maximize his actual payoff.[3]

Our framework captures a wide range of situations. For instance, the players might be altruistic or spiteful, thus care not only about their own payoffs but also about their rival's payoffs. To model this idea we can, as in Bester and Güth (1998) and Possajennikov (2000), write the players' dispositions as $B^i(x^i, x^j, \tau) = \tau\Pi^j(x^i, x^j)$ and $B^j(x^i, x^j, \theta) = \theta\Pi^i(x^i, x^j)$. When τ and θ are positive, the players are altruistic as they attach positive weights to their rival's payoff, while when τ and θ are negative the players are spiteful.

Another example of this framework is concern about social status. Here suppose that $M = N = 1$ (the strategies of the two players are one-dimensional) and let Π^i and Π^j represent the monetary payoffs of the two players. Then, as in Fershtman and Weiss (1998), we can write the dispositions as $B^i(x^i, x^j, \tau) = \tau\sigma(x^i - x^e)$ and $B^j(x^i, x^j, \theta) = \theta\sigma(x^j - x^e)$, where σ is either a positive or a negative parameter and x^e is the average action in the population. Here the revealed preferences of the players are to maximize the sum of their monetary payoffs and their social status, where the latter is linked to the gap between the players' own actions and the average action in the population. The players' types, τ and θ, represent the weights that the players attach to social status.

2.2 The evolution of dispositions

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 from X^i and X^j, respectively, to maximize their perceived payoffs, $U^i(\cdot, \tau)$ and $U^j(\cdot, \theta)$, and obtain true payoffs Π^i and Π^j. For each (τ, θ), let $(y^i(\tau, \theta), y^j(\tau, \theta))$ be a Nash

[3]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 $B^i(x^i, x^j, \tau) \equiv U^i(x^i, x^j, \tau) - \Pi^i(x^i, x^j)$.

equilibrium of this game. Since the strategy spaces X^i and X^j are open, the Nash equilibria of this game are interior. We assume for this discussion that the selection $(y^i(\tau,\theta), y^j(\tau,\theta))$ from the Nash equilibrium correspondence is continuously differentiable at $(\tau,\theta) = (0,0)$.[4] The true payoffs of players i and j in this Nash equilibrium are

$$f^i(\tau,\theta) \equiv \Pi^i\left(y^i(\tau,\theta), y^j(\tau,\theta)\right) \quad \text{and} \quad f^j(\tau,\theta) \equiv \Pi^j\left(y^i(\tau,\theta), y^j(\tau,\theta)\right). \tag{2.3}$$

Since we cast our analysis in an evolutionary setting, these equilibrium payoffs, f^i and f^j, will represent fitness. This formulation leads directly to a natural selection process among different types in the population.

To assess the evolutionary viability of various dispositions, we begin by asking which dispositions are beneficial to a player. Since we are interested in characterizing whether having no disposition (i.e., maximizing true payoffs) can survive evolutionary pressures, we introduce the following notion:

Definition 1 *(Unilaterally beneficial dispositions) The disposition B^i (B^j) is said to be unilaterally beneficial for player i (player j) in the game Γ if there exists $\tau \neq 0$ ($\theta \neq 0$) such that $f^i(\tau,0) > f^i(0,0)$ ($f^j(0,\theta) > f^j(0,0)$).*

It is important to note that Definition 1 says that a disposition is unilaterally beneficial for player i if, given that player j has no disposition (i.e., $\theta = 0$), there exists *some* non-zero type of player i whose fitness is higher than the fitness of type 0. In particular, the definition does not require this property to hold for *all* of types of player i: a unilaterally beneficial disposition might be beneficial for some types of player i but harmful for others.[5]

To study how dispositions evolve, suppose that there is a large population of individuals. At each point $t \geq 0$ in time, the population is characterized by the (possibly correlated) distributions $(\mathcal{T}_t, \Theta_t) \in \Delta(\mathbf{R}) \times \Delta(\mathbf{R})$ of (τ,θ), where $\Delta(\mathbf{R})$ denotes the set of Borel probability distributions over $\mathbf{R}$. At each point in time, a pair of individuals is selected at random from the population. These individuals assume the roles i and j with equal probabilities, and then play the game Γ against one another. Thus, the average fitnesses of types τ and θ at time t are given by

$$\int f^i(\tau,\theta)d\Theta_t \quad \text{and} \quad \int f^j(\tau,\theta)d\mathcal{T}_t. \tag{2.4}$$

[4] We show in the appendix that such a selection is feasible for generic games.

[5] Consider for instance the altruism/spite example mentioned above. Suppose that $f^i_\tau(0,0) \neq 0$. Then if a small degree of altruism ($\tau > 0$) is beneficial, a small degree of spite ($\tau < 0$) would be harmful and vice versa.

We assume that the selection dynamics are monotonically increasing in average fitness. That is, we assume that the distributions of types evolve as follows:

$$\begin{array}{ll} \frac{d}{dt}\mathcal{T}_t(A^i) = \int_{A^i} g^i(\tau, \Theta_t) d\mathcal{T}_t, & A^i \subseteq \mathbf{R} \text{ Borel measurable,} \\ \frac{d}{dt}\Theta_t(A^j) = \int_{A^j} g^j(\mathcal{T}_t, \theta) d\Theta_t, & A^j \subseteq \mathbf{R} \text{ Borel measurable,} \end{array} \tag{2.5}$$

where g^i and g^j are continuous growth-rate functions that satisfy

$$\begin{aligned} g^i(\tau, \Theta_t) > g^i(\tilde{\tau}, \Theta_t) &\iff \int f^i(\tau, \theta) d\Theta_t > \int f^i(\tilde{\tau}, \theta) d\Theta_t, \\ g^j(\mathcal{T}_t, \theta) > g^j(\mathcal{T}_t, \tilde{\theta}) &\iff \int f^j(\tau, \theta) d\mathcal{T}_t > \int f^j(\tau, \tilde{\theta}) d\mathcal{T}_t. \end{aligned} \tag{2.6}$$

To ensure that $\mathcal{T}_t$ and Θ_t remain probability measures for each t, we also assume that g^i and g^j satisfy

$$\int g^i(\tau, \Theta_t) d\mathcal{T}_t = 0, \quad \text{and} \quad \int g^j(\mathcal{T}_t, \theta) d\Theta_t = 0 \quad \text{for each } t. \tag{2.7}$$

Equations (2.5)-(2.7) reflect the idea that the proportion of more successful types in the population increases at the expense of less successful types. This may be due to the fact that more successful individuals have more supportable 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, more successful individuals take part in a larger share of the economic interactions.

To guarantee that the system of differential equations (2.5) has a well-defined solution, we require some additional regularity conditions on the selection dynamics as follows.

Definition 2 *(Regular dynamics) Payoff-monotonic selection dynamics are called regular if g^i and g^j can be extended to the domain $\mathbf{R} \times Y$, where Y is the set of signed Borel measures with variational norm smaller than 2, and on this extended domain, g^i and g^j are uniformly bounded and uniformly Lipschitz continuous. That is,*

$$\begin{aligned} \sup_{\tau \in \mathbf{R}} \left| g^i(\tau, \Theta_t) \right| &< \infty, \quad \sup_{\tau \in \mathbf{R}} \left| g^i(\tau, \Theta_t) - g^i(\tau, \widetilde{\Theta}_t) \right| < K^i \left\| \Theta_t - \widetilde{\Theta}_t \right\|, \qquad \Theta_t, \widetilde{\Theta_t} \in Y, \\ \sup_{\theta \in \mathbf{R}} \left| g^j(\mathcal{T}_t, \theta) \right| &< \infty, \quad \sup_{\theta \in \mathbf{R}} \left| g^j(\mathcal{T}_t, \theta) - g^j(\widetilde{\mathcal{T}}_t, \theta) \right| < K^j \left\| \mathcal{T}_t - \widetilde{\mathcal{T}}_t \right\|, \qquad \mathcal{T}_t, \widetilde{\mathcal{T}}_t \in Y, \end{aligned}$$

for some constants $K^i, K^j > 0$, where $\|\mu\| = \sup_{|h| \leq 1} \left| \int_{\mathbf{R}} h d\mu \right|$ is the variational norm of the signed measure μ.

Oechssler and Riedel (2001, Lemma 3) show that regularity of the dynamics guarantees that the map $(\mathcal{T}_t, \Theta_t) \mapsto \left(\int g^i(\tau, \Theta_t) d\mathcal{T}_t,\ \int g^j(\mathcal{T}_t, \theta) d\Theta_t\right)$ is bounded and Lipschitz continuous in the variational norm, which implies that for any initial distributions $(\mathcal{T}_0, \Theta_0)$, the system of differential equations (2.5) has a unique solution.

To characterize the asymptotic properties of the distributions $(\mathcal{T}_t, \Theta_t)$ we will use the following notion.

Definition 3 *(Asymptotic extinction) The dispositions (B^i, B^j) become asymptotically extinct in the game Γ if $(\mathcal{T}_t, \Theta_t)$ converges weakly to a unit mass at $(\tau, \theta) = (0, 0)$ as $t \to \infty$.*

Theorems 1 and 2 below show that *generically* dispositions do not become asymptotically extinct under any regular payoff-monotonic selection dynamics. Theorem 1 applies to finite-dimensional manifolds of payoff and disposition functions. Here we allow payoff and disposition functions to vary over an arbitrary finite-dimensional manifold provided it contains a sufficiently rich class of functions. We use these finite-dimensional results to show in Theorem 2 that the same result holds when varying over the entire infinite-dimensional families of all thrice continuously differentiable payoff and disposition functions.

2.3 Finite-dimensional manifolds

Let $\tilde{\mathcal{G}}$ denote the space of all pairs of C^3 payoff functions (Π^i, Π^j), and let $\tilde{\mathcal{B}}$ denote the space of all pairs of C^3 disposition functions (B^i, B^j). We endow $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{B}}$ with the Whitney C^3 topology, and $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$ with the natural product topology.[6]

We start by considering a finite-dimensional, boundaryless submanifold $\mathcal{G}$ of $\tilde{\mathcal{G}}$ that is rich enough to allow us to perturb each payoff function in each of the directions x_m^i, x_n^j and $x_m^i x_m^j$ independently and obtain a new pair of payoff functions in $\mathcal{G}$. To formalize this idea, let

$$\begin{aligned} p &= \left(p^1, p^2, p^3\right) = \left(\left(p_1^1, \ldots, p_M^1\right), \left(p_1^2, \ldots, p_N^2\right), \left(p_1^3, \ldots, p_M^3\right)\right) \in \mathbf{R}^{M+N+M}, \\ q &= \left(q^1, q^2, q^3\right) = \left(\left(q_1^1, \ldots, q_M^1\right), \left(q_1^2, \ldots, q_N^2\right), \left(q_1^3, \ldots, q_M^3\right)\right) \in \mathbf{R}^{M+N+M}. \end{aligned}$$

[6]Roughly, the Whitney C^k topology is the topology in which two C^k functions are close if their values, and the values of all of their derivatives of orders up to and including k, are uniformly close. For a formal description and discussion, see e.g. Golubitsky and Guillemin (1973). This is the appropriate topology for our problem because it guarantees that all of the maps we work with, such as the first order conditions for Nash equilibria, are continuous as we vary the payoff and disposition functions.

Given a pair of payoff functions (Π^i, Π^j), define

$$\begin{aligned}\bar{\Pi}^i(x^i, x^j, p) &\equiv \Pi^i\left(x^i, x^j\right) + \sum_{m=1}^{M} p_m^1 x_m^i + \sum_{n=1}^{N} p_n^2 x_n^j + \sum_{m=1}^{M} p_m^3 x_m^i x_m^j, \\ \bar{\Pi}^j(x^i, x^j, q) &\equiv \Pi^j\left(x^i, x^j\right) + \sum_{m=1}^{M} q_m^1 x_m^i + \sum_{n=1}^{N} q_n^2 x_n^j + \sum_{m=1}^{M} q_m^3 x_m^i x_m^j.\end{aligned} \tag{2.8}$$

Using this notation, we assume that the manifold $\mathcal{G}$ is such that for every pair of payoff functions $(\Pi^i, \Pi^j) \in \mathcal{G}$ there exist open neighborhoods $P, Q \subseteq \mathbf{R}^{M+N+M}$ of zero such that $(\bar{\Pi}^i(\cdot,\cdot,p), \bar{\Pi}^j(\cdot,\cdot,q)) \in \mathcal{G}$ for every $(p,q) \in P \times Q$.

Similarly, let $v = (v_1, \ldots, v_M) \in \mathbf{R}^M$ and $w = (w_1, \ldots, w_N) \in \mathbf{R}^N$. Given a pair of dispositions (B^i, B^j), define

$$\begin{aligned}\bar{B}^i(x^i, x^j, \tau, v) &\equiv B^i\left(x^i, x^j, \tau\right) + \tau \sum_{m=1}^{M} v_m x_m^i, \\ \bar{B}^j(x^i, x^j, \theta, w) &\equiv B^j\left(x^i, x^j, \theta\right) + \theta \sum_{n=1}^{N} w_n x_n^j.\end{aligned} \tag{2.9}$$

We consider a finite-dimensional submanifold $\mathcal{B}$ of $\tilde{\mathcal{B}}$ such that for every $(B^i, B^j) \in \mathcal{B}$, there exist neighborhoods $V \subseteq \mathbf{R}^M$, $W \subseteq \mathbf{R}^N$ of zero such that for every $(v, w) \in V \times W$, $(\bar{B}^i(\cdot,\cdot,\cdot,v), \bar{B}^j(\cdot,\cdot,\cdot,w)) \in \mathcal{B}$.

In this finite-dimensional setting, the natural notion of genericity is as follows.

Definition 4 *(Genericity) A property is said to hold for generic combinations of pairs of payoff functions in $\mathcal{G}$ and dispositions in $\mathcal{B}$ if there is an open, full-measure subset A of the product manifold $\mathcal{G} \times \mathcal{B}$ such that the property obtains for all $(\Pi^i, \Pi^j, B^i, B^j) \in A$.*

We can now state the first version of our main result.

Theorem 1 *For generic combinations of pairs of payoff functions $(\Pi^i, \Pi^j) \in \mathcal{G}$ and dispositions $(B^i, B^j) \in \mathcal{B}$:*

(i) The disposition B^i is unilaterally beneficial for player i.

(ii) The dispositions (B^i, B^j) do not asymptotically become extinct under any regular payoff-monotonic selection dynamics.

The basic idea behind this result is can be summarized as follows. Suppose that both players do not have dispositions, so that $\tau = \theta = 0$. The resulting Nash equilibrium of the game Γ is therefore $(y^i(0,0), y^j(0,0))$. Introducing a slight disposition for player i would then change the player's fitness at the rate

$$f^i_\tau(0,0) = \Pi^i_i\left(y^i(0,0), y^j(0,0)\right) y^i_\tau(0,0) + \Pi^i_j\left(y^i(0,0), y^j(0,0)\right) y^j_\tau(0,0). \tag{2.10}$$

The first term is the direct effect on i's equilibrium payoff due to the change in i's own behavior. The second term is the indirect effect caused by the change in j's equilibrium behavior. For generic pairs of payoffs and dispositions, $y^i_\tau(0,0)$ and $y^j_\tau(0,0)$ are well-defined. As $(y^i(0,0), y^j(0,0))$ is an interior Nash equilibrium of Γ, it follows that

$$\Pi^i_i(y^i(0,0), y^j(0,0)) = 0. \tag{2.11}$$

Therefore the first, direct effect vanishes. The essence of the proof is then to show that for generic combinations of payoff and disposition functions, a perturbation in i's disposition ensures that the second, indirect effect does not vanish. That is,

$$f^i_\tau(0,0) = \Pi^i_j\left(y^i(0,0), y^j(0,0)\right) y^j_\tau(0,0) \neq 0. \tag{2.12}$$

This implies in turn that payoff-monotonic selection dynamics cannot converge to a unit mass at $(\tau, \theta) = (0,0)$. If instead the distribution of player j's type were to become concentrated around $\theta = 0$, the fact that $f^i_\tau(0,0) \neq 0$ means that some small nonzero value of τ (positive or negative, depending on the sign of $f^i_\tau(0,0)$) increases the fitness of player i. This in turn implies that a non-zero type of player i would fare better than a type zero player i, and would therefore increase in number at the expense of the type zero player. Thus in the limit the dispositions will not become extinct.[7]

Several remarks about Theorem 1 are now in order.

Remark 1: Theorem 1 can be easily generalized to games with finitely many players. In that case, the proof of the theorem applies verbatim with the index j being interpreted as the vector of all players but i, and with N being the dimension of the product of the strategy spaces of all players but i.

Remark 2: The proof of Theorem 1 relies on the first-order necessary conditions that obtain at interior Nash equilibria of Γ. If we allow the strategy spaces of the players, X^i and X^j, to be closed subsets of $\mathbf{R}^M$ and $\mathbf{R}^N$, then some Nash equilibria may be on the boundary. In such a case, the analysis carries over when restricting attention to the set of directions for which the first-order conditions do hold at equilibrium.[8] No

[7] For symmetric games, Güth and Peleg (2001) identified the analogue of (2.12) as a necessary condition for evolutionary stability (in contrast with the fully dynamic analysis of the current paper). However, Güth and Peleg did not investigate the genericity of this condition.

[8] Dubey (1986) and Anderson and Zame (2001) employ a similar approach to demonstrate the generic Pareto-inefficiency of "non-vertex" Nash equilibria.

first-order conditions need to hold at Nash equilibrium strategies that are extreme points in the strategy sets X^i and X^j, however. This will be the case for instance for pure-strategy Nash equilibria when X^i and X^j are simplices of mixed strategies. Such extreme equilibria are not perturbed when the game is perturbed with a slight disposition, so the marginal analysis in the proof does not apply in this case. In such games, types with small dispositions may have the same fitness as zero types with no disposition.

Our genericity analysis is also inappropriate for pure-strategy Nash equilibria in games with finitely many pure strategies. For such games a global analysis rather than a marginal one is appropriate for characterizing equilibria. Nonetheless, similar results may hold in some such games. For example, in symmetric games with finitely many pure strategies, Dekel et al. (1998) show that for any symmetric Nash equilibrium different from the payoff-maximizing symmetric outcome (as, for example, in the prisoners' dilemma), the lack of dispositions is not evolutionarily viable.

Remark 3: A similar result holds when the strategy spaces X^i and X^j are infinite-dimensional. Unfortunately, in the most obvious examples of such games, such as infinitely repeated games or games with incomplete information, Nash equilibria are typically not locally unique. For infinitely repeated games this follows from the Folk Theorem, while incomplete information games typically have a continuum of Bayesian-Nash equilibria (see e.g., Leininger, Linhart, and Radner, 1989). In such cases, an equilibrium selection is not well-defined even locally, so when small dispositions are introduced it is unclear which equilibrium to consider. Different selections from the equilibrium correspondence may result in contradictory conclusions regarding the effects of the dispositions.[9] We wish to emphasize however that this problem arises not from any inherent limitation of the argument itself; rather, the evolutionary analysis ceases to be predictive because the equilibrium is not locally unique.

2.4 All games and dispositions

The genericity result established in the previous subsection might appear to be somewhat limited in scope because of its restriction to certain finite-dimensional submanifolds $\mathcal{G}$ and $\mathcal{B}$. Next we show that an analogous result holds when we vary over the infinite-dimensional sets of all possible pairs of payoff functions and dispositions.

To extend our genericity results to the space of all payoff and distribution functions, we will need a notion of genericity that is suitable in an infinite-dimensional setting. Unfortunately, there is no natural analogue of Lebesgue measure in an infinite-dimensional space, and standard topological notions of "almost all" such as open and dense or residual are not entirely satisfactory, particularly in problems like ours in which "almost all" is

[9] In specific cases, however, there may be more natural candidates for such selections; see for example the analysis in Section 4 below and in Heifetz and Segev (2001).

loosely interpreted in a probabilistic sense as a statement about the likelihood of particular events. For example, open and dense sets in $\mathbf{R}^n$ can have arbitrarily small measure, and residual sets can have measure 0. Nevertheless, Christensen (1974) and Hunt, Sauer, and Yorke (1992) have developed measure-theoretic analogues of Lebesgue measure 0 and full Lebesgue measure for infinite-dimensional spaces, called shyness and prevalence.

Definition 5 *(Shyness and prevalence) Let Y be a topological vector space. A universally measurable subset $E \subset Y$ is* shy *if there is a regular Borel probability measure μ on Y with compact support such that $\mu(E+y) = 0$ for every $y \in Y$.*[10] *A (not necessarily universally measurable) subset $F \subset Y$ is* shy *if it is contained in a shy universally measurable set. A subset $E \subset Y$ is* prevalent *if its complement $Y \setminus E$ is shy.*

Christensen (1974) and Hunt, Sauer and Yorke (1992) show that shyness and prevalence have the properties we ought to require of measure-theoretic notions of "smallness" and "largeness." In particular, the countable union of shy sets is shy, no relatively open subset is shy, and a subset of $\mathbf{R}^n$ is shy in $\mathbf{R}^n$ if and only if it has Lebesgue measure 0. Hunt, Sauer, and Yorke (1992) also provide simple sufficient conditions for their notions of shyness and prevalence (here we adopt the somewhat more descriptive terminology from Anderson and Zame, 2001).[11]

Definition 6 *(Finite shyness and finite prevalence) Let Y be a topological vector space. A universally measurable set $E \subset Y$ is* finitely shy *if there is a finite dimensional subspace $V \subset Y$ such that $(E-y) \cap V$ has Lebesgue measure 0 in V for every $y \in Y$. A universally measurable set $E \subset Y$ is* finitely prevalent *if its complement $Y \setminus E$ is finitely shy.*

Sets that are finitely shy are shy, hence sets that are finitely prevalent are prevalent. Using this fact together with the results we established for finite-dimensional submanifolds will yield a general version of our results when payoffs and dispositions vary over the entire infinite-dimensional spaces $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{B}}$.

We can now state a second version of our main result:

Theorem 2 *There exists an open, prevalent subset $\mathcal{P}$ of $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$ such that for each $(\Pi^i, \Pi^j, B^i, B^j) \in \mathcal{P}$,*

[10] A set $E \subset Y$ is universally measurable if for every Borel measure η on Y, E belongs to the completion with respect to η of the sigma algebra of Borel sets.

[11] Anderson and Zame (2001) have extended the work of Hunt, Sauer and Yorke (1992) and Christensen (1974) by defining prevalence and shyness relative to a convex subset that may be a small subset of the ambient space. Their extension is useful in many applications, particularly in economics, in which the relevant parameters are drawn not from the whole space but from some subset, such as a convex cone or an order interval, that may itself be a shy subset of the ambient space. Here we use the original notion as formulated in Hunt, Sauer and Yorke (1992).

(i) The disposition B^i is beneficial for player i.

(ii) The dispositions (B^i, B^j) do not asymptotically become extinct under any regular payoff-monotonic dynamics.

3 An example

In this section we illustrate our main results by means of an example. In this example we are able to completely characterize the asymptotic distribution in the population by making use of the special structure of payoffs and dispositions. The example yields stronger results than our general results in Theorems 1 and 2 since it shows that the evolutionary process converges to a unique, positive type, whereas the two general theorems merely show that the evolutionary process cannot converge to zero types who lack dispositions.

For this example, suppose that the strategy spaces of the players are $X^i = X^j = \mathbf{R}$, and the actual payoff functions are

$$\Pi^i(x^i, x^j) = (\alpha - bx^j - x^i)x^i, \qquad \Pi^j(x^i, x^j) = (\alpha - bx^i - x^j)x^j, \tag{3.1}$$

where $\alpha > 0$, and $b \in (-1, 1)$. Moreover, suppose that the dispositions of the players are given by:

$$B^i(x^i, x^j, \tau) = \tau x^i, \qquad B^j(x^i, x^j, \theta) = \theta x^j, \qquad \tau, \theta \in T, \tag{3.2}$$

where $T \subset \mathbf{R}$ is a compact interval that contains 0. The example differs from the more general analysis in Section 2 in that the set of types is now a compact interval rather than $\mathbf{R}$. This assumption will enable us to characterize the asymptotic distribution of types.

Using these payoff and disposition functions, the perceived payoff functions are given by

$$\begin{aligned} U^i(x^i, x^j, \tau) &= \Pi^i(x^i, x^j) + B^i(x^i, x^j, \tau) = (\alpha + \tau - bx^j - x^i)x^i, \\ U^j(x^i, x^j, \theta) &= \Pi^j(x^i, x^j) + B^i(x^i, x^j, \theta) = (\alpha + \theta - bx^i - x^j)x^j. \end{aligned} \tag{3.3}$$

These dispositions can be interpreted as "self-esteem" biases reflecting over- and under-confidence. Here the players either overestimate the return to their own actions, if τ and θ are positive, or underestimate these returns, if τ and θ are negative.

The unique Nash equilibrium when players i and j choose x^i and x^j to maximize their perceived payoffs given (τ, θ) is

$$y^i(\tau, \theta) = \frac{2(\alpha + \tau) - b(\alpha + \theta)}{4 - b^2}, \qquad y^j(\tau, \theta) = \frac{2(\alpha + \theta) - b(\alpha + \tau)}{4 - b^2}. \tag{3.4}$$

The resulting fitnesses of the players are given by

$$\begin{aligned} f^i(\tau, \theta) &\equiv \Pi^i(y^i(\tau, \theta), y^j(\tau, \theta)) = \frac{(2(\alpha+\tau) - b(\alpha+\theta))\big(2\alpha - (2-b^2)\tau - b(\alpha+\theta)\big)}{(4-b^2)^2}, \\ f^j(\tau, \theta) &\equiv \Pi^j(y^i(\tau, \theta), y^j(\tau, \theta)) = \frac{(2(\alpha+\theta) - b(\alpha+\tau))\big(2\alpha - (2-b^2)\theta - b(\alpha+\tau)\big)}{(4-b^2)^2}. \end{aligned} \tag{3.5}$$

Note that the dispositions B^i and B^j are unilaterally beneficial for players i and j, because

$$f^i_\tau(0,0) = f^j_\theta(0,0) = \frac{\alpha b^2}{(2+b)^2(2-b)} > 0. \tag{3.6}$$

We are now interested in characterizing the asymptotic distribution of types. To simplify the discussion and notation, we assume that $\mathcal{T}_t = \Theta_t$, i.e., the types of both players are drawn from the same distribution on the support T.[12]

Proposition 1: *Suppose that* $\frac{b^2\alpha}{4+2b-b^2} \in T$. *Then for any initial distribution* $\mathcal{T}_0$, $\mathcal{T}_t$ *converges in distribution to a unit mass at* $\frac{b^2\alpha}{4+2b-b^2}$ *under any regular payoff-monotonic selection dynamics.*

Proposition 1 shows that as long as $b \neq 0$, the dispositions B^i and B^j do not asymptotically become extinct, and in fact the population converges to a unit mass at $\frac{b^2\alpha}{4+2b-b^2}$. Since $\frac{b^2\alpha}{4+2b-b^2} > 0$, it follows that aside from the case where $b = 0$, in which the payoff of each player is independent of the other player's actions so there is no strategic interaction, 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, asymptotically become extinct. In this case we have a sharper result than obtains in Theorem 1, in that for a generic set of parameter values, in particular for any values other than $b = 0$, a unique positive type of the disposition survives in the limit.

4 Imperfect observability of dispositions

Thus far, we have assumed that players i and j 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 this section, we pursue further the possibility that preferences may not be perfectly observed. We consider three different settings. First, we consider settings in which preferences are observed in only a fraction of all encounters. Second, we consider the case in which players' types are observed with some noise. Third, we consider the case in which players may be engaged in costly signaling regarding their preferences.

The natural solution concept for each of these settings is Bayesian equilibrium. Unfortunately, as we discussed above, Bayesian equilibria are typically not locally unique;

[12]The results would be identical even if the players' types were drawn from different distributions.

consequently, it is impossible to generalize Theorems 1 and 2 to these settings. Nonetheless, we can show that in the absence of this technical obstacle, the evolutionary viability of dispositions is maintained under imperfect observability. We use the example of Section 3 as our vehicle, since in all three settings it gives rise to a unique Bayesian equilibrium for any given distribution $(\mathcal{T}, \Theta)$ of types. Qualitatively similar results would obtain for any other example that admits a unique Bayesian equilibrium at least in some weak neighborhood of the unit mass at $(\tau, \theta) = (0, 0)$.

4.1 Partial observability

In this subsection we consider a setting in which preferences are mutually observed in some exogenously specified fraction $1 - \rho$ of all interactions, but are completely unobserved in the remaining fraction ρ of interactions, where $\rho \in (0, 1)$. When preferences are not mutually observed, a Bayesian equilibrium $(y^i(\tau, \Theta_t), y^j(\mathcal{T}_t, \theta))$ is played at time t, in which each player maximizes his or her preferences given the distribution of actions taken by the opponent, which in turn depends on the current distribution of the opponent's preferences.

Proposition 2: *Suppose that in the example considered in Section 3, the players observe each other's preferences with probability* $1 - \rho$. *Moreover, suppose that* $\frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3} \in T$. *Then under any regular payoff-monotonic selection dynamics, the distribution of types in the population converges in distribution to a unit mass at*

$$\tau^* = \theta^* = \frac{2(1-\rho)b^2\alpha}{8+4b-2b^2-\rho b^3}. \tag{4.2}$$

Proposition 2 shows that the emerging type is monotonic in the probability $(1 - \rho)$ of observability. Moreover, the disposition becomes asymptotically extinct, that is, $\tau^* = \theta^* = 0$, only in the extreme cases where either $\rho = 1$, so preferences are never observed, or $b = 0$, so there is no strategic interaction between the players.[13] Thus we also obtain a more precise result than Theorem 1 in terms of the generic emergence of dispositions, since for any $b \in (-1, 1) \setminus \{0\}$ and any $\rho \in [0, 1)$, players have a disposition in the limit distribution.

4.2 Noisy observability

We now consider the possibility that preferences are always mutually observed, but that the observation of preferences is subject to some randomly distributed noise. Specifically,

[13] In different but analogous settings, Dekel et al. (1998), Ely and Yilankaya (2001), Ok and Vega Redondo (2001) and Güth and Peleg (2001) also show that payoff-maximization is evolutionarily stable if preferences are completely unobservable.

we assume that before choosing actions, players i and j receive noisy signals of each other's types, update their beliefs about each other's preferences, and then play a Bayesian equilibrium given these updated beliefs.

We will again cast our analysis in the context of the example from Section 3. Now we suppose that before the players choose their actions, they receive the following signals about each other's types:

$$s^i = \tau + \nu, \qquad s^j = \theta + \nu, \tag{4.3}$$

where ν is a random variable distributed on the support $[-r, r]$ according to a cumulative distribution function $\mathcal{N}$.[14]

Proposition 3. *In the example of Section 3 with noisy observability, if the players' signals are s^i and s^j, then the dispositions do not asymptotically become extinct under any regular payoff-monotonic selection dynamics.*

4.3 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 credible, costly signals about character. Under this argument, the observable symptoms of emotional arousal reflect the operation of automatic physical reactions, like adrenaline flow, muscle tension and more rapid breathing, and the fact that these are *automatic* reactions has some survival value when facing imminent threats. A mutant who might be able to consciously control these functions, and therefore enjoy the benefits of lying without being caught, would 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 were, and thus the larger the ability of the individual to hide her true emotions.

[14]The assumption that the support of ν is symmetric around 0 is not essential. However, the assumption that the support is finite is important for our results because it makes it possible for players to distinguish between zero and non-zero types.

To model this idea, we suppose that players generate signals m^i and m^j regarding their true types τ and θ. Once again, the multiplicity of Bayesian equilibria associated with such a game at each point in time preempts a general analysis. Hence we again consider an extension of the example from Section 3. 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 $c(\tau - m^i)^2$ and $c(\theta - m^j)^2$, where $c > 0$. We suppose that the signals m^i and m^j evolve in parallel with the players' types τ and θ 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 and θ, 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 costly signaling version of the example from Section 3. Suppose that* $\frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c)-b^3} \in T$ *and* $\frac{2(1+4c)b^2\alpha}{2(4+2b-b^2)(1+4c)-b^3} \in M$. *Then the joint distribution of types and signals will converge under any regular payoff-monotonic selection dynamics to a unit mass at*

$$\begin{aligned} \tau^* &= \theta^* = \frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c)-b^3}, \\ m^* &= \frac{2\,(1+4c)\,b^2\alpha}{2(4+2b-b^2)(1+4c)-b^3}. \end{aligned}$$

Note that unless $b = 0$, so that there is strategic interaction between the players, $\tau^* = \theta^* > 0$. This implies that generically in terms of the parameters of the model, the players will have dispositions, as in the previous case of noisy observability in Section 4.1. In contrast, however, the resulting type to which the population converges is the same as in the case of noisy observability only when $c \to \infty$, that is, only when deception is infinitely costly. Otherwise, the resulting type 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 $0 < c < \infty$, $m^* > \tau^* = \theta^*$, implying that the population will converge over time to a type whose appearance will exaggerate the true disposition. That is, agents will appear to be more overconfident than they really are.

5 Conclusion

This paper is part of a large body of work that focuses on the meaning and foundations of rationality in economic models. "Rational" behavior is typically interpreted to mean

that decision makers have well-defined, stable 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 rationality assumptions. One approach to explaining such "irrational" behavior is to attribute it to various bounds on the rationality of agents, 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 rationality assumptions. 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. While work along these lines has been successful in providing foundations for various types of deviations from true payoff maximization, it is often criticized on two important grounds (see e.g., Samuelson, 2001). First, specific results typically consider preferences and dispositions that are carefully tailored to the particular game of interest, which raises the question of how robust such specific examples are and whether they extend to more general types of preferences and dispositions. Second, most of the existing work modeling the evolution of preferences assumes that preferences are perfectly observed, while it is unclear whether this assumption is reasonable or whether the results obtained still hold if this assumption is relaxed.

Our work addresses both of these questions. Under the assumption that preferences are observable, we show that in almost every game and for almost every type of distortion of a player's actual payoffs, some extent of this distortion is beneficial to the player because of the resulting effect on opponents' play. Hence, any standard evolutionary process in which selection dynamics are monotone in payoffs will not eliminate such distortions; in particular, under any such selection dynamics, the population will *not* converge to payoff maximizing behavior. This implies in turn that the evolutionary viability of dispositions is *generic*, and independent of the particular parametric models employed in most of the literature. We also show that the viability of dispositions may be robust to unobservability of preferences. Although the lack of local uniqueness of Bayesian equilibria in models with unobserved preferences precludes an extension of our results to general settings with imperfect observability, when the Bayesian equilibrium is unique (as the examples in Section 4 illustrate) dispositions remain evolutionarily viable in the sense that the population still does *not* converge to payoff maximizing behavior. Moreover, in settings where preferences are perfectly observed in a fraction of interactions and completely unobserved in others and in settings in which the players convey (costly) information about their preferences, it is even possible to completely characterize the limiting distribution, and to show that for generic combinations of parameters, the population will converge over time to a monomorphic type whose objective function does not coincide with actual payoffs.

The generic value of dispositions in strategic 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, the institutions themselves may influence the long-run preferences of participating agents; that is, preferences may be in part an endogenous feature of the particular institutional framework. Preliminary analysis in this vein includes Bar-Gill and Fershtman (2001), Güth and Ockenfels (2001), Fershtman and Heifetz (2002), and Heifetz et al. (2002). More generally, the approach has the potential to illuminate other aspects of "intrinsic motivation" (see e.g., Kreps, 1997) not fully captured by other theories, in issues ranging from corporate culture to the relations between culture and exogenous economic factors. These questions seem to present promising avenues for future research.

6 Appendix

In order to prove Theorems 1 and 2 we proceed with a sequence of lemmata. We make repeated use of the following standard definition and theorem, which we include here for completeness.[15]

Definition 7 *(regular value) Let X and S be boundaryless, C^r manifolds, and $G : X \times S \to \mathbf{R}^K$ be a C^r function, where $r \geq 1$. An element $y \in \mathbf{R}^K$ is a regular value of G if for all (x, s) such that $G(x, s) = y$, the derivative $D_{x,s}G(x, s)$ is surjective.*

In particular, notice that if there are no points (x, s) such that $G(x, s) = y$, then y is trivially a regular value of G.

Remark 4: In the arguments below we will frequently need to show that zero is a regular value of various maps. To this end we will rely on two useful observations. First, we will repeatedly use the assumption that these manifolds contain an open set around each point consisting of a particular type of perturbation. More precisely, fix $(\Pi^i, \Pi^j) \in \mathcal{G}$ and recall that we assume that there exist open neighborhoods $P, Q \subseteq \mathbf{R}^{M+N+M}$ of zero such that $(\bar{\Pi}^i(\cdot, \cdot, p), \bar{\Pi}^j(\cdot, \cdot, q)) \in \mathcal{G}$ for each $(p, q) \in P \times Q$, where $\bar{\Pi}^i$ and $\bar{\Pi}^j$ are given in (2.8). Now let $h : X^i \times X^j \times \mathcal{G} \to \mathbf{R}^K$ be an arbitrary C^1 function. Then zero is a regular value of h provided $Dh(x^i, x^j, \Pi^i, \Pi^j)$ has rank K (i.e., is surjective) for each $(x^i, x^j, \Pi^i, \Pi^j) \in h^{-1}(0)$. Given our assumptions about $\mathcal{G}$, to show that $Dh(x^i, x^j, \Pi^i, \Pi^j)$ has rank K it then suffices to show that

$$D_{p,q}h(x^i, x^j, \bar{\Pi}^i(x^i, x^j, 0), \bar{\Pi}^j(x^i, x^j, 0))$$

has rank K.

Second, if the derivative

$$D_{i,j}h(x^i, x^j, \Pi^i, \Pi^j)$$

[15] For example, see Hirsch (1976).

does not have rank K for any $(x^i, x^j) \in X^i \times X^j$, then zero can be a regular value of $h(\cdot, \cdot, \Pi^i, \Pi^j)$ only if $h(x^i, x^j, \Pi^i, \Pi^j) \neq 0$ for all $(x^i, x^j) \in X^i \times X^j$.

Theorem 3 *(The transversality theorem). Let X and S be finite-dimensional, boundaryless, C^r manifolds and $G : X \times S \to \mathbf{R}^K$ be a C^r function, where $r > \max\{0, \dim X - K\}$. For each $s \in S$ let $G(\cdot, s)$ be the restriction of G to $X \times \{s\}$. If $y \in \mathbf{R}^K$ is a regular value of G, then for almost every $s \in S$, y is a regular value of $G(\cdot, s)$. In addition, if $s \mapsto G(\cdot, s)$ is continuous in the Whitney C^r topology, then $\{s \in S : s$ is a regular value of $G(\cdot, s)\}$ is open.*

The first step in our argument is to show that equilibria are locally unique in almost all games. To that end, we first define the class of regular games. We will slightly abuse terminology by referring to a pair of payoff functions (Π^i, Π^j) as a game (recall that the strategy spaces X^i, X^j remain fixed throughout).

Definition 8 *(Regular games) A game is called regular if at each of its Nash equilibria (y^i, y^j), the $(M+N) \times (M+N)$ matrix*

$$\begin{pmatrix} \Pi^i_{ii}(y^i, y^j) & \Pi^i_{ij}(y^i, y^j) \\ \Pi^j_{ji}(y^i, y^j) & \Pi^j_{jj}(y^i, y^j) \end{pmatrix}$$

has full rank. We denote by $\mathcal{R} \subseteq \mathcal{G}$ the set of regular games.

Our first lemma shows that almost all games are regular.

Lemma 1 *The set of regular games $\mathcal{R}$ is an open, full-measure subset of $\mathcal{G}$.*

Proof. Fix a game $(\Pi^i, \Pi^j) \in \mathcal{G}$. Since the strategy spaces X^i, X^j are open, Nash equilibria of the game are interior. Thus, at each Nash equilibrium (y^i, y^j) of the game, the following system of $M + N$ first order conditions holds:

$$\begin{pmatrix} \Pi^i_i(y^i, y^j) \\ \Pi^j_j(y^i, y^j) \end{pmatrix} = 0.$$

Define the map $\phi : X^i \times X^j \times \mathcal{G} \to \mathbf{R}^{M+N}$ by

$$\phi(\cdot, \cdot, \Pi^i, \Pi^j) = \begin{pmatrix} \Pi^i_i(\cdot, \cdot) \\ \Pi^j_j(\cdot, \cdot) \end{pmatrix}.$$

Consider the derivative

$$D_{p^1, q^2}\phi(y^i, y^j, \bar{\Pi}^i(\cdot, \cdot, 0), \bar{\Pi}^j(\cdot, \cdot, 0)) = \begin{pmatrix} I_M & 0 \\ 0 & I_N \end{pmatrix},$$

where I_M and I_N are the $M \times M$ and $N \times N$ identity matrices. Since the matrix has rank $M+N$ for each (y^i, y^j), it follows from Remark 4 that zero is a regular value of ϕ. Therefore, the transversality theorem implies that there is a set of full measure $R \subset \mathcal{G}$ such that zero is a regular value of $\phi(\cdot,\cdot,\Pi^i,\Pi^j)$ for each game $(\Pi^i,\Pi^j) \in R$. For each $(\Pi^i,\Pi^j) \in R$, the definition of regular value and the fact that zero is a regular value of $\phi(\cdot,\cdot,\Pi^i,\Pi^j)$ implies that the derivative

$$D_{i,j}\phi(y^i,y^j,\Pi^i,\Pi^j) = \begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \end{pmatrix}$$

has full rank $M+N$ at each Nash equilibrium (y^i,y^j) of (Π^i,Π^j). Thus, using the definition of a regular game, a game $(\Pi^i,\Pi^j) \in \mathcal{G}$ is regular if and only if 0 is a regular value of $\phi(\cdot,\cdot,\Pi^i,\Pi^j)$, that is, $R = \mathcal{R}$. Thus $\mathcal{R}$ has full measure.

Finally, since the map $(\Pi^i,\Pi^j) \mapsto \phi(\cdot,\cdot,\Pi^i,\Pi^j)$ is continuous in the Whitney C^1 topology, $\mathcal{R}$ is open by the transversality theorem. ■

The next lemma shows that in a regular game, the Nash equilibrium correspondence is locally single-valued in a neighborhood of zero. This feature allows us to study the effects of small dispositions on the true equilibrium payoffs in a well-defined manner.

Lemma 2 *Consider a regular game (Π^i,Π^j) and let (y^i,y^j) be a Nash equilibrium of the game. For any pair of dispositions $(B^i,B^j) \in \mathcal{B}$, there is a neighborhood V_0 of $\tau = 0$ and a unique C^1 function*

$$Z(\cdot) \equiv (y^i(\cdot,0), y^j(\cdot,0)) : V_0 \to X^i \times X^j,$$

such that $(y^i(0,0), y^j(0,0)) = (y^i,y^j)$ and $(y^i(\tau,0), y^j(\tau,0))$ is a Nash equilibrium of the game $(\Pi^i + B^i, \Pi^j)$ when $\tau \in V_o$. Moreover,

$$\begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \end{pmatrix} \begin{pmatrix} y^i_\tau(0,0) \\ y^j_\tau(0,0) \end{pmatrix} = \begin{pmatrix} -B^i_{i\tau}(y^i,y^j,0) \\ 0 \end{pmatrix}. \tag{A.1}$$

Proof. Suppose that $\theta = 0$ (player j has no disposition), so that $B^j(\cdot,\cdot,0) \equiv 0$. Then a Nash equilibrium $(y^i(\tau,0), y^j(\tau,0))$ of the game $(\Pi^i + B^i, \Pi^j)$ satisfies the following system of $M+N$ first order conditions

$$\begin{pmatrix} \Pi^i_i(y^i,y^j) + B^i_i(y^i,y^j,\tau) \\ \Pi^j_j(y^i,y^j) \end{pmatrix} = 0. \tag{A.2}$$

Since $B^i(\cdot,\cdot,0) \equiv 0$, $B^i_i(y^i,y^j,0) \equiv 0$, hence at $\tau = 0$ this system becomes

$$\begin{pmatrix} \Pi^i_i(y^i,y^j) \\ \Pi^j_j(y^i,y^j) \end{pmatrix} = 0.$$

Since the game (Π^i, Π^j) is regular, zero is a regular value of the map

$$\begin{pmatrix} \Pi^i_i(\cdot,\cdot) \\ \Pi^j_j(\cdot,\cdot) \end{pmatrix} : \mathbf{R}^{M+N} \to \mathbf{R}^{M+N}.$$

The implicit function theorem then implies that the Nash equilibrium map $Z(\cdot) \equiv (y^i(\cdot,0), y^j(\cdot,0))$ is locally defined and C^1 in a neighborhood V_0 of $\tau = 0$. Finally, since $B^i(\cdot,\cdot,0) \equiv 0$, $B^i_{ii}(y^i,y^j,0) = B^i_{ij}(y^i,y^j,0) \equiv 0$. Then (A.1) follows by differentiating (A.2) with respect to τ and evaluating at $\tau = 0$. ■

Now let $\mathcal{U} = \mathcal{G} \times \mathcal{B}$ be the manifold of perceived payoff functions, so

$$\mathcal{U} = \left\{ \left(U^i, U^j\right) = (\Pi^i + B^i, \Pi^j + B^j) : X^i \times X^j \times \mathbf{R} \to \mathbf{R}^2 | (\Pi^i, \Pi^j) \in \mathcal{G},\ (B^i, B^j) \in \mathcal{B} \right\}. \tag{A.3}$$

Since, $B^i(x^i, x^j, 0) \equiv B^j(x^i, x^j, 0) \equiv 0$, the projection $\mathrm{Pr}_{\mathcal{G}} : \mathcal{U} \to \mathcal{G}$ maps (U^i, U^j) to the corresponding game

$$\mathrm{Pr}_{\mathcal{G}}(U^i, U^j) \equiv \left(U^i(\cdot,\cdot,0), U^j(\cdot,\cdot,0)\right),$$

while the projection $\mathrm{Pr}_{\mathcal{B}} : \mathcal{U} \to \mathcal{B}$ maps (U^i, U^j) to the corresponding dispositions

$$\mathrm{Pr}_{\mathcal{B}}(U^i, U^j) \equiv \left(U^i - U^i(\cdot,\cdot,0), U^j - U^j(\cdot,\cdot,0)\right).$$

By Lemma 1, the set $\mathcal{U}_R \equiv \mathcal{R} \times \mathcal{B}$ is an open, full-measure subset of $\mathcal{U}$.

Lemma 3 *There is an open, full-measure subset $\mathcal{U}_B \subseteq \mathcal{U}_R$ of perceived payoff functions (U^i, U^j) for which $B^i_{i\tau}(y^i, y^j, 0) \neq 0$ at each Nash equilibrium (y^i, y^j) of (Π^i, Π^j).*

Proof. Let $\xi : X^i \times X^j \times \mathcal{U}_R \to \mathbf{R}^{M+N+M}$ be given by

$$\xi(\cdot,\cdot,\Pi^i,\Pi^j,B^i,B^j) = \begin{pmatrix} \Pi^i_i(\cdot,\cdot) \\ \Pi^j_j(\cdot,\cdot) \\ B^i_{i\tau}(\cdot,\cdot,0) \end{pmatrix}.$$

Since (Π^i, Π^j) is a regular game, by definition the $(M+N) \times (M+N)$ matrix

$$\begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \end{pmatrix}$$

has rank $M+N$ at each Nash equilibrium (y^i, y^j) of (Π^i, Π^j). Therefore, the derivative

$$D_{i,j,v}\xi(y^i,y^j,\Pi^i,\Pi^j,\overline{B}^i(\cdot,\cdot,\cdot,0),\overline{B}^j(\cdot,\cdot,\cdot,0)) = \begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) & 0 \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) & 0 \\ B^i_{i\tau i}(y^i,y^j,0) & B^i_{i\tau j}(y^i,y^j,0) & I_M \end{pmatrix}$$

has rank $M+N+M$ at each Nash equilibrium (y^i, y^j) of (Π^i, Π^j). Consequently, by Remark 4, zero is a regular value of ξ. Therefore, the transversality theorem implies that there is a full-measure subset $\mathcal{U}_B \subseteq \mathcal{U}_R$ such that zero is a regular value of the map $\xi(\cdot, \cdot, \Pi^i, \Pi^j, B^i, B^j)$ for all $(\Pi^i + B^i, \Pi^j + B^j) \in \mathcal{U}_B$. Since the map $(\Pi^i, \Pi^j, B^i, B^j) \mapsto \xi(\cdot, \cdot, \Pi^i, \Pi^j, B^i, B^j)$ is continuous in the Whitney C^1 topology, $\mathcal{U}_B$ is open by the transversality theorem as well.

Let $(\Pi^i + B^i, \Pi^j + B^j) \in \mathcal{U}_B$. Since the derivative

$$D_{i,j}\xi(x^i, x^j, \Pi^i, \Pi^j, B^i, B^j) = \begin{pmatrix} \Pi^i_{ii}(x^i, x^j) & \Pi^i_{ij}(x^i, x^j) \\ \Pi^j_{ji}(x^i, x^j) & \Pi^j_{jj}(x^i, x^j) \\ B^i_{i\tau i}(x^i, x^j, 0) & B^i_{i\tau j}(x^i, x^j, 0) \end{pmatrix}$$

has only $M+N$ columns, it cannot have rank $M+N+M$ for any $(x^i, x^j) \in X^i \times X^j$. By Remark 4, zero can be a regular value of $\xi(\cdot, \cdot, \Pi^i, \Pi^j, B^i, B^j)$ only if $\xi(x^i, x^j, \Pi^i, \Pi^j, B^i, B^j) \neq 0$ for all $(x^i, x^j) \in X^i \times X^j$. Therefore, at a (interior) Nash equilibrium (y^i, y^j) of the game (Π^i, Π^j), where

$$\begin{pmatrix} \Pi^i_i(y^i, y^j) \\ \Pi^j_j(y^i, y^j) \end{pmatrix} = 0,$$

we must have $B^i_{i\tau}(y^i, y^j, 0) \neq 0$. ■

Let $\tilde{\Pi}^j_{ji}(x^i, x^j, q)$ be the $M \times M$ matrix consisting of the first M rows of $\bar{\Pi}^j_{ji}(x^i, x^j, q)$. If $\tilde{\Pi}^j_{ji}(x^i, x^j, 0)$ has rank $M-k$, it takes k consecutive first-order perturbations (of its diagonal entries, for example) to produce a matrix of full rank. This idea is formalized in the following lemma.

Lemma 4 *For each $k = 0, \ldots, M$ there is an open, full-measure subset $\mathcal{U}_k \subseteq \mathcal{U}_B$ such that for every $(\Pi^i, \Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_k)$,*

$$\frac{\partial^{M-k}}{\partial q^3_1 \partial q^3_2 \ldots \partial q^3_{M-k}} \det \tilde{\Pi}^j_{ji}(y^i, y^j, 0) \neq 0$$

at each Nash equilibrium (y^i, y^j) of (Π^i, Π^j).

Proof. We proceed by induction on k. For the base case $k = 0$, we claim that for any Π^i and any (y^i, y^j, q)

$$\frac{\partial^M}{\partial q^3_1 \partial q^3_2 \ldots \partial q^3_M} \det \tilde{\Pi}^j_{ji}(y^i, y^j, q) = 1. \tag{A.4}$$

This follows because the determinant of $\tilde{\Pi}^j_{ji}(\cdot,\cdot)$ is a sum of products, of M factors each, and the derivative with respect to $(q^3_1,\ldots,q^3_M)$ of each of these products is zero with the exception of the diagonal product $\prod_{m=1}^M \frac{\partial^2\Pi^j}{\partial x^j_m \partial x^i_m}$. For this term, note that

$$\frac{\partial^2\Pi^j(y^i,y^j,q)}{\partial x^j_m \partial x^i_m} = q^3_m,$$

for each (y^i,y^j,q), so

$$\prod_{m=1}^M \frac{\partial^2\Pi^j(y^i,y^j,q)}{\partial x^j_m \partial x^i_m} = \prod_{m=1}^M q^3_m$$

which implies that for any (y^i,y^j,q),

$$\frac{\partial^M}{\partial q^3_1 \partial q^3_2 \ldots \partial q^3_M}\left(\prod_{m=1}^M \frac{\partial^2\Pi^j(y^i,y^j,q)}{\partial x^j_m \partial x^i_m}\right) = 1.$$

Now suppose that the claim holds for $k=\ell-1$. Then we claim there is an open, full-measure subset $\mathcal{U}_\ell \subseteq \mathcal{U}_{\ell-1}$ such that for games (Π^i,Π^j) that correspond to perceived payoff functions in $\mathcal{U}_\ell$, zero is a regular value of the map

$$\psi(\cdot,\cdot,\Pi^i,\Pi^j) \equiv \begin{pmatrix} \Pi^i_i(\cdot,\cdot) \\ \Pi^j_j(\cdot,\cdot) \\ \frac{\partial^{M-\ell}}{\partial q^3_1 \partial q^3_2 \ldots \partial q^3_M} \det \tilde{\Pi}^j_{ji}(\cdot,\cdot,0) \end{pmatrix} : X^i \times X^j \times \mathcal{G} \to \mathbf{R}^{M+N+1}. \tag{A.5}$$

To see this, note that the derivative

$$D_{p^1,q^1,q^3_{M-(\ell-1)}}\psi(y^i,y^j,\bar{\Pi}^i(\cdot,\cdot,0),\bar{\Pi}^j(\cdot,\cdot,0)) = \begin{pmatrix} I_M & 0 & 0 \\ & & 0 \\ & & \vdots \\ 0 & I_N & y^i_{M-(\ell-1)} \\ & & \vdots \\ & & 0 \\ 0 & 0 & \frac{\partial^{M-(\ell-1)}}{\partial q^3_1 \partial q^3_2 \ldots \partial q^3_M} \det \tilde{\Pi}^j_{ji}(y^i,y^j,0) \end{pmatrix} \tag{A.6}$$

has rank $M+N+1$ at each Nash equilibrium (y^i,y^j) of the game $(\Pi^i,\Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_{\ell-1})$. Consequently, by Remark 4, zero is a regular value of ψ. Therefore, the transversality theorem implies that there exists a set of full measure $\mathcal{U}_\ell \subset \mathcal{U}_{\ell-1}$ such that zero is a regular value of $\psi(\cdot,\cdot,\Pi^i,\Pi^j)$ for each $(\Pi^i,\Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_\ell)$. Since the map $(\Pi^i,\Pi^j) \mapsto \psi(\cdot,\cdot,\Pi^i,\Pi^j)$ is continuous in the Whitney C^1 topology, $\mathcal{U}^*$ is an open subset of $\mathcal{U}_M$ by the transversality theorem.

Let $(\Pi^i + B^i, \Pi^j + B^j) \in \mathcal{U}^*$. Since the derivative

$$D_{i,j}\zeta(x^i, x^j, \Pi^i, \Pi^j, B^i, B^j) = \begin{pmatrix} \Pi^i_{ii}(y^i, y^j) & \Pi^i_{ij}(y^i, y^j) \\ \Pi^j_{ji}(y^i, y^j) & \Pi^j_{jj}(y^i, y^j) \\ D_i\left(\Pi^i_j(y^i, y^j)y^j_\tau(0,0)\right) & D_j\left(\Pi^i_j(y^i, y^j)y^j_\tau(0,0)\right) \end{pmatrix}$$

has only $M + N$ columns, it cannot have rank $M + N + 1$. By Remark 4, zero can be a regular value of $\zeta(\cdot, \cdot, \Pi^i, \Pi^j, B^i, B^j)$ only if $\zeta(x^i, x^j, \Pi^i, \Pi^j, B^i, B^j) \neq 0$ for all $(x^i, x^j) \in X^i \times X^j$. Thus if $(\Pi^i, \Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}^*)$ and (y^i, y^j) is a (interior) Nash equilibrium of the game (Π^i, Π^j), so that $\Pi^i_i(y^i, y^j) = \Pi^j_j(y^i, y^j) = 0$, then we must have $\Pi^i_j(y^i, y^j)y^j_\tau(0,0) \neq 0$, as required. ■

Lemma 5 *Let $(U^i, U^j) \in \mathcal{U}_M$, $(\Pi^i, \Pi^j) = \mathrm{Pr}_{\mathcal{G}}(U^i, U^j)$ and $(B^i, B^j) = \mathrm{Pr}_{\mathcal{B}}(U^i, U^j)$. For every Nash equilibrium (y^i, y^j) of (Π^i, Π^j), $y^j_\tau(0,0) \neq 0$.*

Proof. Let $(U^i, U^j) \in \mathcal{U}_M$, $(\Pi^i, \Pi^j) = \mathrm{Pr}_{\mathcal{G}}(U^i, U^j)$ and $(B^i, B^j) = \mathrm{Pr}_{\mathcal{B}}(U^i, U^j)$. Let (y^i, y^j) be a Nash equilibrium of (Π^i, Π^j). Now recall from Lemma 4 that for each $k = 0, \ldots, M$ there is an open, full-measure subset $\mathcal{U}_k \subseteq \mathcal{U}_B$ such that for every $(\Pi^i, \Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_k)$,

$$\frac{\partial^{M-k}}{\partial q_1^3 \partial q_2^3 \ldots \partial q_{M-k}^3} \det \tilde{\Pi}^j_{ji}(y^i, y^j, 0) \neq 0.$$

When $k = M$, this implies that

$$\det \tilde{\Pi}^j_{ji}(y^i, y^j) \neq 0.$$

Hence, $\Pi^j_{ji}(y^i, y^j)$ has rank M.

Now note from (A.1) that

$$\Pi^j_{ji}(y^i, y^j)y^i_\tau(0,0) + \Pi^j_{jj}(y^i, y^j)y^j_\tau(0,0) = 0, \tag{A.7a}$$

and

$$\Pi^i_{ii}(y^i, y^j)y^i_\tau(0,0) + \Pi^i_{ij}(y^i, y^j)y^j_\tau(0,0) = -B^i_{i\tau}(y^i, y^j, 0), \tag{A.7b}$$

and suppose by way of contradiction that $y^j_\tau(0,0) = 0$. Since $\Pi^j_{ji}(y^i, y^j)$ has rank M, it is injective. Then since $y^j_\tau(0,0) = 0$, (A.7a) implies that $y^i_\tau(0,0) = 0$. Recalling from Lemma 3 that $-B^i_{i\tau}(y^i, y^j, 0) \neq 0$, this means that (A.7b) cannot hold, a contradiction. ■

Lemma 6 *There is an open, full-measure subset $\mathcal{U}^* \subseteq \mathcal{U}_M$ such that if $(\Pi^i, \Pi^j) = \mathrm{Pr}_{\mathcal{G}}(U^i, U^j)$ and $(B^i, B^j) = \mathrm{Pr}_{\mathcal{B}}(U^i, U^j)$ for some $(U^i, U^j) \in \mathcal{U}^*$, then for every Nash equilibrium (y^i, y^j) of the game (Π^i, Π^j),*

$$\Pi^i_j(y^i, y^j)y^j_\tau(0,0) \neq 0.$$

Proof. Denote by J_n the $(M+N)\times(M+N)$ matrix obtained from

$$\begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \end{pmatrix}$$

after replacing the n-th column by

$$\begin{pmatrix} -B^i_{i\tau}(y^i,y^j,0) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Then by (A.1) and Cramer's rule

$$y^j_\tau(0,0) = \left(\ldots, \frac{\det J_n}{\det \begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \end{pmatrix}}, \ldots \right).$$

In particular, note that $y^j_\tau(0,0)$ is independent of p^2.

Let $\zeta : X^i \times X^j \times \mathcal{G} \times \mathcal{B} \to \mathbf{R}^{M+N+1}$ be given by

$$\zeta(\cdot,\cdot,\Pi^i,\Pi^j,B^i,B^j) = \begin{pmatrix} \Pi^i_i(\cdot,\cdot) \\ \Pi^j_j(\cdot,\cdot) \\ \Pi^i_j(\cdot,\cdot)y^j_\tau(0,0) \end{pmatrix}. \tag{A.8}$$

Since $y^j_\tau(0,0)$ is independent of p^2 and since $D_{p^2}\bar{\Pi}^i_j(\cdot,\cdot,p) = 1$,

$$D_{p^2}\left(\bar{\Pi}^i_j(y^i,y^j,p)y^j_\tau(0,0)\right) = y^j_\tau(0,0).$$

Since by Lemma 5, $y^j_\tau(0,0) \neq 0$, it follows that if (y^i,y^j) is a Nash equilibrium of (Π^i,Π^j), then the derivative

$$D_{p^1,q^2,p^2}\zeta(y^i,y^j,\bar{\Pi}^i,\bar{\Pi}^j,B^i,B^j) = \begin{pmatrix} I_M & 0 & 0 \\ 0 & I_N & 0 \\ 0 & 0 & y^j_\tau(0,0) \end{pmatrix}$$

has rank $M+N+1$. Consequently, by Remark 4, zero is a regular value of ζ. Therefore, by the transversality theorem, there is a full-measure subset $\mathcal{U}^* \subset \mathcal{U}_M$ such that zero is a regular value of $\zeta(\cdot,\cdot,\Pi^i,\Pi^j,B^i,B^j)$ for all $(\Pi^i+B^i,\Pi^j+B^j) \in \mathcal{U}^*$. Since the map $(\Pi^i,\Pi^j,B^i,B^j) \mapsto \zeta(\cdot,\cdot,\Pi^i,\Pi^j,B^i,B^j)$ is continuous in the Whitney C^1 topology, $\mathcal{U}^*$ is an open subset of $\mathcal{U}_M$ by the transversality theorem.

Let $(\Pi^i+B^i,\Pi^j+B^j) \in \mathcal{U}^*$. Since the derivative

$$D_{i,j}\zeta(x^i,x^j,\Pi^i,\Pi^j,B^i,B^j) = \begin{pmatrix} \Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j) \\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j) \\ D_i\left(\Pi^i_j(y^i,y^j)y^j_\tau(0,0)\right) & D_j\left(\Pi^i_j(y^i,y^j)y^j_\tau(0,0)\right) \end{pmatrix}$$

has only $M+N$ columns, it cannot have rank $M+N+1$. By Remark 4, zero can be a regular value of $\zeta(\cdot,\cdot,\Pi^i,\Pi^j,B^i,B^j)$ only if $\zeta(x^i,x^j,\Pi^i,\Pi^j,B^i,B^j)\neq 0$ for all $(x^i,x^j)\in X^i\times X^j$. Thus if $(\Pi^i,\Pi^j)\in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}^*)$ and (y^i,y^j) is a (interior) Nash equilibrium of the game (Π^i,Π^j), so that $\Pi^i_i(y^i,y^j)=\Pi^j_j(y^i,y^j)=0$, then we must have $\Pi^i_j(y^i,y^j)y^j_\tau(0,0)\neq 0$, as required. ■

Lemma 7 *For perceived payoffs $(U^i,U^j)\in\mathcal{U}^*$, $f^i_\tau(0,0)\neq 0$.*

Proof. At $(\tau,\theta)=(0,0)$ we have

$$f^i_\tau(0,0)=\Pi^i_i(y^i,y^j)y^i_\tau(0,0)+\Pi^i_j(y^i,y^j)y^j_\tau(0,0).$$

As (y^i,y^j) is a Nash equilibrium, $\Pi^i_i(y^i,y^j)=0$. By Lemma 6, $\Pi^i_j(y^i,y^j)y^j_\tau(0,0)\neq 0$. Hence $f^i_\tau(0,0)\neq 0$. ■

Next, consider the "fitness" game in which players i and j choose their types, τ and θ, to maximize their fitness, $f^i(\tau,\theta)$ and $f^j(\tau,\theta)$. Note that Lemma 7 shows that for perceived payoffs $(U^i,U^j)\in\mathcal{U}^*$, the profile $(\tau,\theta)=(0,0)$ is not a Nash equilibrium of this fitness game, since $f^i_\tau(0,0)$ means that player i's best response to $\theta=0$ is nonzero. Moreover, this will be enough to allow us to conclude that the dispositions do not become asymptotically extinct under any regular payoff-monotonic selection dynamics, as the next lemma shows.

Lemma 8 *If the dispositions (B^i,B^j) asymptotically become extinct in the game (Π^i,Π^j), then the types $(\tau,\theta)=(0,0)$ are a Nash equilibrium of the fitness game.*

Proof. Let δ_0 denote the unit mass at $(0,0)$. Suppose, by way of contradiction, that $(\tau,\theta)=(0,0)$ is not a Nash equilibrium of the fitness game. Then without loss of generality, for some $\tau\neq 0$ we have $f^i(\tau,0)>f^i(0,0)$. Since f^i is continuous, there exists a neighborhood A of δ_0 and neighborhoods V_0 of 0 and V_τ of τ such that if $\Theta\in A$, $\hat\tau\in V_0$ and $\tilde\tau\in V_\tau$, then $\int f^i(\tilde\tau,\theta)d\Theta_t>\int f^i(\hat\tau,\theta)d\Theta_t$. Now since (B^i,B^j) becomes asymptotically extinct, there exists t' sufficiently large so that for every $t\geq t'$, $\Theta_t\in A$, and hence for every $t\geq t'$, $\int f^i(\tilde\tau,\theta)d\Theta_t>\int f^i(\hat\tau,\theta)d\Theta_t$ for any $\tilde\tau\in V_\tau$ and $\hat\tau\in V_0$. Then, using (2.6), the growth rates satisfy $g^i(\tilde\tau,\Theta_t)>g^i(\hat\tau,\Theta_t)$ for every $t\geq t'$, $\tilde\tau\in V_\tau$ and $\hat\tau\in V_0$ as well. By (2.5), this implies that for $t\geq t'$ we have $\frac{d}{dt}\mathcal{T}_t(V_{\tilde\tau})>\frac{d}{dt}\mathcal{T}_t(V_0)$. This means that $\mathcal{T}_t$ does not converge weakly to δ_0, a contradiction. ■

Proof of Theorem 1. Part (i) of the theorem follows immediately from Lemma 7. As for part (ii), Lemma 8 implies that if $(\tau,\theta)=(0,0)$ is not a Nash equilibrium of the fitness game then the dispositions (B^i,B^j) do not become asymptotically extinct. Since Lemma 7 implies that for perceived payoffs in $\mathcal{U}^*$, $(\tau,\theta)=(0,0)$ is not a Nash equilibrium of the

fitness game, it follows that for $(U^i, U^j) \in \mathcal{U}^*$, the dispositions $(B^i, B^j) = \mathrm{Pr}_{\mathcal{B}}(U^i, U^j)$ do not become asymptotically extinct in the game $(\Pi^i, \Pi^j) = \mathrm{Pr}_{\mathcal{G}}(U^i, U^j)$. ■

Proof of Theorem 2. Let

$$\mathcal{P} = \left\{ \left(\Pi^i, \Pi^j, B^i, B^j\right) \in \tilde{\mathcal{G}} \times \tilde{\mathcal{B}} : f^i_\tau(0,0) \neq 0 \right\}.$$

Thus every $(\Pi^i, \Pi^j, B^i, B^j) \in \mathcal{P}$ satisfies part (i) of the theorem, and by Lemma 8 it also satisfies part (ii). It remains to show that $\mathcal{P}$ is finitely prevalent in $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$. To this end, we first claim that $\mathcal{P}$ is open. To see this, note that by (2.10),

$$f^i_\tau(0,0) = \Pi^i_i\left(y^i(0,0), y^j(0,0)\right) y^i_\tau(0,0) + \Pi^i_j\left(y^i(0,0), y^j(0,0)\right) y^j_\tau(0,0)$$

and by (A.1) and (A.2), $y^i(0,0)$, $y^j(0,0)$, $y^i_\tau(0,0)$ and $y^j_\tau(0,0)$ are continuous in (Π^i, Π^j, B^i, B^j) on $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$. Thus $\Pi^i_i\left(y^i(0,0), y^j(0,0)\right)$ and $\Pi^i_j\left(y^i(0,0), y^j(0,0)\right)$ are continuous on $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$ as well. This implies that $f^i_\tau(0,0)$ is continuous in (Π^i, Π^j, B^i, B^j) on $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$, which suffices to show that $\mathcal{P}$ is an open subset of $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$.

Now let

$$\begin{aligned}
\mathcal{V} \;=\; \Bigg\{ & (\widehat{\Pi}^i, \widehat{\Pi}^j) \in \tilde{\mathcal{G}} \mid \widehat{\Pi}^i(x^i, x^j) = \sum_{m=1}^{M} p^1_m x^i_m + \sum_{n=1}^{N} p^2_n x^j_n + \sum_{m=1}^{M} p^3_m x^i_m x^j_m \\
& \text{for some } p \in \mathbf{R}^{M+N+M}, \\
& \widehat{\Pi}^j(x^i, x^j) = \sum_{m=1}^{M} q^1_m x^i_m + \sum_{n=1}^{N} q^2_n x^j_n + \sum_{m=1}^{M} q^3_m x^i_m x^j_m \\
& \text{for some } q \in \mathbf{R}^{M+N+M} \Bigg\},
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{W} \;=\; \Bigg\{ & (\widehat{B}^i, \widehat{B}^j) \in \tilde{\mathcal{B}} \mid \widehat{B}^i(x^i, x^j, \tau) = \tau \sum_{m=1}^{M} v_m x^i_m \text{ for some } v \in \mathbf{R}^M, \\
& \widehat{B}^j(x^i, x^j, \theta) = \theta \sum_{n=1}^{N} w_n x^j_n \text{ for some } w \in \mathbf{R}^N \Bigg\}.
\end{aligned}$$

Now by Theorem 1, for every $(\Pi^i, \Pi^j, B^i, B^j) \in \tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$, $[(\mathcal{V} \times \mathcal{W}) + (\Pi^i, \Pi^j, B^i, B^j)] \cap \mathcal{P}$ has full measure in $\mathcal{V} \times \mathcal{W}$. Equivalently, $(\mathcal{P} - (\Pi^i, \Pi^j, B^i, B^j)) \cap (\mathcal{V} \times \mathcal{W})$ has full measure in $\mathcal{V} \times \mathcal{W}$. Thus $\mathcal{P}$ is finitely prevalent. Since finitely prevalent sets are prevalent, the proof is complete. ■

To prove Proposition 1 we need the following theorem, which is of independent interest. It generalizes Theorem 1 in Samuelson and Zhang (1992) to the case of games

with infinitely many strategies. In keeping with the example, we state the theorem for symmetric two-player games with a compact strategy space; the result carries over however to asymmetric games with virtually the same proof.

To state this result we need some additional notation. We consider a symmetric game with common strategy space T 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 $n \geq 1$,

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

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 $n \geq 1$,

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

Theorem 4 *Let T be a compact space of strategies, $f : T \times T \to \mathbf{R}$ be the continuous payoff function of a symmetric two-player game, and $g : T \times T \to \mathbf{R}$ a regular, payoff monotonic growth-rate function. Let G_t be the population dynamics defined by the differential equation*

$$\frac{d}{dt} G_t(A) = \int_A g(t, G_t) dG_t, \quad A \subseteq T \text{ Borel measurable} \tag{A.9}$$

given initial distribution G_0. 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

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

Let

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

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

$$s = \frac{1}{2} \sup_{y \in D_{k-1}} [f(x,y) - f(d,y)], \tag{A.10}$$

and

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

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

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,

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

and

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

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 $d' \in \overline{W}_d$,

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

and

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

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 $t \geq \bar{t}$, $G_t(C) \leq \frac{\epsilon}{8M}$ and $G_t(T \setminus C) > 1 - \epsilon$. Then for any $x' \in \overline{V}_x, d' \in \overline{W}_d$ and $t \geq \bar{t}$,

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

By the continuity of f, (A.11) holds also when G_t is replaced by any probability measure $\mu \in A \equiv \overline{\{G_t\}_{t \geq \bar{t}}}$, 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

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

Then (A.12) 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 $t \geq \bar{t}$

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

Hence, by (A.9), for $t \geq \bar{t}$,

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

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

Proof of Proposition 1. Consider the fitness game in which players i and j simultaneously choose their types, τ and θ, to maximize their respective fitness, f^i and f^j. The best-response functions in this game are

$$BR^i(\theta) = \frac{b^2\left((2-b)\,\alpha - b\theta\right)}{4(2-b^2)}, \qquad BR^j(\tau) = \frac{b^2\left((2-b)\,\alpha - b\tau\right)}{4(2-b^2)}.$$

Then the strategy sets are one-dimensional compact intervals, the functions f^i and f^j are smooth and strictly concave in the players' own strategies, and the slopes of the best-response functions are less than 1 in absolute value. It follows from Moulin (1984, Theorem 4) that the game can be solved by iterated elimination of strongly dominated strategies. The unique outcome that survives this process is

$$\tau^* = \theta^* = \frac{b^2\alpha}{4 + 2b - b^2} > 0.$$

By Theorem 4 in the Appendix, all other types which are serially dominated (i.e., do not survive the iterated elimination process) become asymptotically extinct under any regular payoff-monotonic selection dynamics. Consequently, the selection dynamics converges to a unit mass at $\frac{b^2\alpha}{4+2b-b^2}$. ■

Proof of Proposition 2. When preferences are mutually observable, the equilibrium actions are specified in (3.4). 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:

$$U^i(x^i, \overline{x}; \tau) = (\alpha + \tau - b\overline{x} - x^i)x^i. \tag{A.15}$$

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

$$BR^i(\overline{x}; \tau) = \frac{\alpha + \tau - b\overline{x}}{2}, \qquad BR^j(\overline{x}; \theta) = \frac{\alpha + \theta - b\overline{x}}{2}. \tag{A.16}$$

On the equilibrium path, the beliefs of the two players about $\overline{x}$ must be correct. Taking expectations on both sides of equation (A.16), 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 $BR^j(\overline{x}; \theta)$ yields equilibrium actions

$$\widehat{y}^i = \frac{\alpha + \tau - b\frac{\alpha+\omega}{2+b}}{2}, \qquad \widehat{y}^j = \frac{\alpha + \theta - b\frac{\alpha+\omega}{2+b}}{2}.$$

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

$$\left(\alpha - b\frac{\alpha + \theta - b\frac{\alpha+\omega}{2+b}}{2} - \frac{\alpha + \tau - b\frac{\alpha+\omega}{2+b}}{2}\right)\left(\frac{\alpha + \tau - b\frac{\alpha+\omega}{2+b}}{2}\right). \tag{A.17}$$

With probability $1 - \rho$, preferences are observed and individual i's payoff is as in (4.5), whereas with probability ρ preferences are not observed and i's payoff is given by (A.17). Hence the expected fitness of player i when the player's type is τ, the type of player j is θ, and the current average type of player j in the population is ω is given by

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

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 and θ, to maximize their fitness. The best-response function of player i in this game is

$$\begin{aligned} BR^i(\theta;\rho,\omega) &= \frac{4\alpha b^2(2-b)(1-\rho)}{2(16-8b^2+\rho b^4)} + \frac{b\rho(2-b)(4b+8-b^3-2b^2)}{2(16-8b^2+\rho b^4)}\omega \\ &\quad - \frac{b(\rho b^4+4b^2-12\rho b^2+16\rho)}{2(16-8b^2+\rho b^4)}\theta. \end{aligned} \tag{A.18}$$

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

In what follows we prove that the population converges over time to a stable monomorphic type. To this end, let $\tau^* = \theta^*$ be defined implicitly by the equation

$$\tau^* = BR^i(\tau^*;\rho,\tau^*). \tag{A.19}$$

Solving this equation yields

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

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

The idea behind the rest of the proof is as follows. The fitness game played at each point in time depends on the current average type ω. Hence, we cannot prove the convergence result as in Proposition 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 under any payoff-monotonic selection dynamics, the average type ω will eventually converge to an interval $[\tau_\ell - \delta, \tau_h + \delta]$, where δ is some small positive number.[16] The fact that $\tau^* \in [\tau_\ell - \delta, \tau_h + \delta]$ enables us to show that further types are serially dominated and thus that types outside some 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, $\mathcal{T}_t$. The evolution of Θ_t is analogous. Let $T = [\underline{\tau}, \bar{\tau}]$ and

$$\overleftarrow{\tau} = \inf\left\{\tau' > \tau^* : \forall \tau > \tau' \; \exists V_\tau \ni \tau, V_\tau \text{ open, s.t. } \lim_{t\to\infty} \mathcal{T}_t(V_\tau) = 0\right\} \tag{A.20a}$$

$$\underrightarrow{\tau} = \sup\left\{\tau' < \tau^* : \forall \tau < \tau' \; \exists V_\tau \ni \tau, V_\tau \text{ open, s.t. } \lim_{t\to\infty} \mathcal{T}_t(V_\tau) = 0\right\} \tag{A.20b}$$

Here we use the convention that $\overleftarrow{\tau} = \bar{\tau}$ if the infimum in (A.20a) ranges over an empty set, and similarly that $\underrightarrow{\tau} = \underline{\tau}$ if the supremum in (A.20b) ranges over an empty set.

[16] 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 $[\tau_\ell - \delta, \tau_h + \delta]$.

Let $\varepsilon > 0$, and set $A = [\underline{\tau}, \underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}] \cup [\overleftarrow{\tau} + \frac{\varepsilon}{2}, \overline{\tau}]$. Then A is a compact subset of T. For each $\tau \in A$, let V_τ be a neighborhood of τ as given in (A.20a,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 $\lim_{t\to\infty} \mathcal{T}_t(V_{\tau_k}) = 0$ for each $k = 1, \ldots, n$, there exists a time t_ε such that for $t > t_\varepsilon$, $\mathcal{T}_t(V_{\tau_k}) < \frac{\varepsilon}{2nM}$ for each $k = 1, \ldots, n$, where $M = \max\{\varepsilon, \overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right), \left(\underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}\right) - \underline{\tau}\}$. Hence, for $t > t_\varepsilon$ we conclude

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

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

$$\begin{aligned} \omega &< \frac{\varepsilon}{2M}\overline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) = \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2M}\left(\overline{\tau} - \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right)\right) \quad \text{(A.21)} \\ &\leq \left(\overleftarrow{\tau} + \frac{\varepsilon}{2}\right) + \frac{\varepsilon}{2} = \overleftarrow{\tau} + \varepsilon, \end{aligned}$$

and

$$\begin{aligned} \omega &> \frac{\varepsilon}{2M}\underline{\tau} + \left(1 - \frac{\varepsilon}{2M}\right)\left(\underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}\right) = \left(\underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2M}\left(\left(\underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}\right) - \underline{\tau}\right) \quad \text{(A.22)} \\ &\geq \left(\underset{\rightarrow}{\tau} - \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2} = \underset{\rightarrow}{\tau} - \varepsilon. \end{aligned}$$

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

Next, note from equation (A.18) that the slope of the best-response function of type τ in the fitness game in the (τ, θ) space is given by

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

Now consider the case where $b < 0$. In this case, this slope is negative and less than 1 in absolute value. Hence, fixing the value of ω, there exists a unique symmetric Nash equilibrium in the fitness game. Moreover, since $b < 0$, equation (A.18) shows that $BR^i(\cdot; \rho, \omega)$ is decreasing in ω. Hence, the "highest" symmetric Nash equilibrium in the fitness game is attained when $\omega = \max\{\underset{\rightarrow}{\tau} - \varepsilon, \underline{\tau}\}$ and the "lowest" equilibrium is attained when $\omega = \min\{\overleftarrow{\tau} + \varepsilon, \overline{\tau}\}$. Let the highest and lowest symmetric Nash equilibria be $(\overleftarrow{\tau_\varepsilon}, \overleftarrow{\tau_\varepsilon})$ and $(\underset{\rightarrow}{\tau_\varepsilon}, \underset{\rightarrow}{\tau_\varepsilon})$, respectively. That is, $\overleftarrow{\tau_\varepsilon}$ and $\underset{\rightarrow}{\tau_\varepsilon}$ are the solutions to the equations $\overleftarrow{\tau_\varepsilon} = BR^i(\overleftarrow{\tau_\varepsilon}; \rho, (\underset{\rightarrow}{\tau} - \varepsilon) \vee \underline{\tau})$ and $\underset{\rightarrow}{\tau_\varepsilon} = BR^i(\underset{\rightarrow}{\tau_\varepsilon}; \rho, (\overleftarrow{\tau} + \varepsilon) \wedge \overline{\tau})$.

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

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

As a result, $f^i(\widetilde{\tau},\theta;\rho,\omega) < f^i(\tau,\theta;\rho,\omega)$ implies $f^i(\widetilde{\tau},\theta;\rho,\widetilde{\omega}) < f^i(\tau,\theta;\rho,\widetilde{\omega})$ and similarly $f^i(\widetilde{\tau},\theta;\rho,\widetilde{\omega}) > f^i(\tau,\theta;\rho,\widetilde{\omega})$ implies $f^i(\widetilde{\tau},\theta;\rho,\omega) > f^i(\tau,\theta;\rho,\omega)$. 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 4, 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

$$\begin{aligned}
\overleftarrow{\tau} &\leq \inf_{\varepsilon>0} \overleftarrow{\tau}_\varepsilon \equiv \overleftarrow{\tau}_0 = BR^i(\overleftarrow{\tau}_0;\rho,\underrightarrow{\tau}), \\
\underrightarrow{\tau} &\geq \sup_{\varepsilon>0} \underrightarrow{\tau}_\varepsilon \equiv \underrightarrow{\tau}_0 = BR^i(\underrightarrow{\tau}_0;\rho,\overleftarrow{\tau}).
\end{aligned}$$

Subtracting the second inequality from the first, using equation (A.18), and rearranging terms yields:

$$0 \leq \overleftarrow{\tau} - \underrightarrow{\tau} \leq \overleftarrow{\tau}_0 - \underrightarrow{\tau}_0 = \frac{b\rho\,(2-b)\,(2+b)^2}{\rho b^4 + 4\rho b^3 + 4b^2 - 4\rho b^2 - 8b - 8b\rho - 16}\left(\overleftarrow{\tau} - \underrightarrow{\tau}\right). \quad \text{(A.23)}$$

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

We now consider the case where $b > 0$. Then, $\frac{\partial(f^i)^2(\tau,\theta;\rho,\omega)}{\partial\tau\partial\omega} = \frac{b\rho}{4} > 0$, so if $\omega < \widetilde{\omega}$ and $\tau < \widetilde{\tau}$, then

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

As a result, $f^i(\widetilde{\tau},\theta;\rho,\omega) > f^i(\tau,\theta;\rho,\omega)$ implies $f^i(\widetilde{\tau},\theta;\rho,\widetilde{\omega}) > f^i(\tau,\theta;\rho,\widetilde{\omega})$ and similarly $f^i(\widetilde{\tau},\theta;\rho,\widetilde{\omega}) < f^i(\tau,\theta;\rho,\widetilde{\omega})$ implies $f^i(\widetilde{\tau},\theta;\rho,\omega) < f^i(\tau,\theta;\rho,\omega)$. Since $b > 0$, equation (A.18) implies that $BR^i(\cdot;\rho,\omega)$ is upward sloping and increasing in ω. Hence, 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 4, 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

$$\begin{aligned}
\overleftarrow{\tau} &\leq \inf_{\varepsilon>0} \overleftarrow{\tau}_\varepsilon \equiv \overleftarrow{\tau}_0 = BR^i(\underrightarrow{\tau}_0;\rho,\overleftarrow{\tau}), \\
\underrightarrow{\tau} &\geq \sup_{\varepsilon>0} \underrightarrow{\tau}_\varepsilon \equiv \underrightarrow{\tau}_0 = BR^i(\overleftarrow{\tau}_0;\rho,\underrightarrow{\tau}).
\end{aligned}$$

Subtracting the second inequality from the first, using equation (A.18), and rearranging terms yields:

$$\begin{aligned}
\overleftarrow{\tau}_0 - \underrightarrow{\tau}_0 &= \frac{b\rho\,(2-b)\,(4b+8-b^3-2b^2)}{2\,(16-8b^2+\rho b^4)}\left(\overleftarrow{\tau} - \underrightarrow{\tau}\right) \\
&\quad + \frac{b\,(\rho b^4 + 4b^2 - 12\rho b^2 + 16\rho)}{2\,(16-8b^2+\rho b^4)}\left(\overleftarrow{\tau}_0 - \underrightarrow{\tau}_0\right).
\end{aligned}$$

This implies in turn that

$$0 \leq \overleftarrow{\tau} - \underrightarrow{\tau} \leq \overleftarrow{\tau_0} - \underrightarrow{\tau_0} = \frac{(-4+b^2)(2-b)b\rho}{\rho b^4 - 4b^3\rho + 4b^2 - 4\rho b^2 + 8b + 8b\rho - 16}\left(\overleftarrow{\tau} - \underrightarrow{\tau}\right). \quad \text{(A.24)}$$

Since $b < 1$, the coefficient of $\overleftarrow{\tau} - \underrightarrow{\tau}$ is strictly smaller than 1, implying that $\overleftarrow{\tau} = \underrightarrow{\tau} = \tau^*$ as desired. ■

Proof of Proposition 3. Before the players choose their actions, they observe the signals s^i and s^j, but not the true types τ and θ. Player i with type τ and signal s^i chooses an action x^i so as to maximize the expected perceived payoff

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

where the expectation is taken over players j who produced the signal s^j when they meet somebody with signal s^i, and $\chi^j(s^i, s^j)$ is the (current) average action of these players. The problem of player j is analogous.

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

$$x^i = \frac{\alpha + \tau - b\chi^j(s^i, s^j)}{2}, \qquad x^j = \frac{\alpha + \theta - b\chi^i(s^i, s^j)}{2}. \quad \text{(A.25)}$$

Let $\tau(s^i)$ be the (current) average type of player i who produces the signal s^i and let $\theta(s^j)$ be the (current) average type of player j who produces the signal s^j. Taking expectations on both sides of (A.25) yields

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

Solving this pair of equations yields

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

Substituting this in (A.25) reveals that the equilibrium actions of players i and j are

$$\widehat{x}^i = \frac{\alpha + \tau - b\frac{2\alpha + 2\theta(s^j) - \alpha b - b\tau(s^i)}{4-b^2}}{2}, \qquad \widehat{x}^j = \frac{\alpha + \theta - b\frac{2\alpha + 2\tau(s^i) - \alpha b - b\theta(s^j)}{4-b^2}}{2}.$$

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

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

Now, suppose that Θ_t converges to a unit mass at 0. We will show that it is impossible for $\mathcal{T}_t$ to also converge to a unit mass at 0. Since Θ_t converges to a unit mass at 0, then the posterior belief of player i regarding player j's type, $\theta(s^j)$, also converges to a unit mass at 0. Thus, the average fitness of player i with type τ who produces the signal s^i converges to

$$\begin{aligned} f^i(\tau, s^i) &= \left(\alpha - b\frac{\alpha(2-b) - b\tau(s^i)}{4-b^2} - \frac{\alpha + \tau - b\frac{\alpha(2-b)-b\tau(s^i)}{4-b^2}}{2}\right)\frac{\alpha + \tau - b\frac{\alpha(2-b)-b\tau(s^i)}{4-b^2}}{2} \\ &= \frac{b^2}{4-b^2}\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right)\tau(s^i) + \left(\frac{\alpha}{2+b} - \frac{\tau}{2}\right)\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right). \end{aligned}$$

Now suppose by way of contradiction that $\mathcal{T}_t$ also converges to a unit mass at 0. If player i produces a signal $s^i \in [-r, r]$, then player j cannot rule out the possibility that player i's type is $\tau = 0$. Therefore, $\tau(s^i)$ converges to 0 for all $s^i \in [-r, r]$. Now, consider player i whose type τ is positive but close to 0 (the argument when τ is negative and close to 0 is analogous). With probability $\mathcal{N}(r - \tau)$, the player produces a signal $s^i \in [-r + \tau, r]$. Given such a signal, player j cannot rule out the possibility that player i's type is 0, so player i's payoff in this case converges to

$$\left(\frac{\alpha}{2+b} - \frac{\tau}{2}\right)\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right).$$

With probability $1 - \mathcal{N}(r - \tau)$, the player produces a signal $s^i \in (r, r + \tau]$. In that case, player j realizes that player i's type cannot be 0 and is bounded from below by $s^i - r$. Since $\tau > 0$, $f^i(\tau, s^i)$ is increasing in $\tau(s^i)$. Consequently, the overall average fitness of player i with type τ will be bounded from below asymptotically by

$$\begin{aligned} &\mathcal{N}(r - \tau)\left(\frac{\alpha}{2+b} - \frac{\tau}{2}\right)\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right) \\ &+ \int_{r-\tau}^{\tau}\left[\frac{b^2}{4-b^2}\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right)(\tau + \nu - r) + \left(\frac{\alpha}{2+b} - \frac{\tau}{2}\right)\left(\frac{\alpha}{2+b} + \frac{\tau}{2}\right)\right] d\mathcal{N}(\nu). \end{aligned}$$

The derivative of this expression with respect to τ, evaluated at $\tau = 0$, is

$$\mathcal{N}'(r)\frac{rb^2\alpha}{(4-b^2)(2+b)} > 0.$$

Thus asymptotically some $\tau > 0$ dominates $\tau = 0$. The disposition is therefore unilaterally beneficial to player i, which implies that $\mathcal{T}_t$ cannot converge to a unit mass at $\tau = 0$ under any regular payoff-monotonic selection dynamics. ■

Proof of Proposition 4. The proposition follows from Theorem 4 once we show that (τ^*, m^*) is the only combination of strategies that survives iterated elimination of strongly dominated strategies (τ, m) in the fitness game.

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

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

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

$$x^i = \frac{\alpha + \tau - b\chi^j(m^i, m^j)}{2}, \qquad x^j = \frac{\alpha + \theta - b\chi^i(m^i, m^j)}{2}. \tag{A.26}$$

Let $\tau(m^i)$ and $\theta(m^j)$, 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 (A.26) yields:

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

Solving these two equations yields

$$\chi^i(m^i, m^j) = \frac{2\alpha + 2\tau(m^i) - \alpha b - b\theta(m^j)}{4 - b^2}, \qquad \chi^j(m^i, m^j) = \frac{2\alpha + 2\theta(m^j) - \alpha b - b\tau(m^i)}{4 - b^2}. \tag{A.27}$$

Substituting $\chi^i(m^i, m^j)$ and $\chi^j(m^i, m^j)$ in (A.26) reveals that the equilibrium actions of players i and j are given by

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

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

$$\begin{aligned} f^i\left(\left(\tau, m^i\right), m^j\right) &= \left(\alpha - b\frac{2\alpha + 2\theta(m^j) - \alpha b - b\tau(m^i)}{4 - b^2} - \frac{\alpha + \tau - b\frac{2\alpha + 2\theta(m^j) - \alpha b - b\tau(m^i)}{4 - b^2}}{2}\right) \\ &\quad \times \frac{\alpha + \tau - b\frac{2\alpha + 2\theta(m^j) - \alpha b - b\tau(m^i)}{4 - b^2}}{2} - c(m^i - \tau)^2. \end{aligned} \tag{A.28}$$

The corresponding average fitness of player j is analogous. Maximizing $f^i((\tau, m^i), m^j)$ with respect to τ and $f^j((\theta, m^j), m^i)$ with respect to θ 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

$$\tau^*(m^i) = \frac{4c}{1 + 4c}m^i, \qquad \theta^*(m^j) = \frac{4c}{1 + 4c}m^j.$$

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

$$\lim_{t\to\infty} \tau\left(m^i\right) = \tau^*(m^i) = \frac{4c}{1+4c} m^i, \qquad \lim_{t\to\infty} \theta\left(m^j\right) = \theta^*(m^j) = \frac{4c}{1+4c} m^j. \tag{A.29}$$

Taking the limit of the expressions in (A.28) as $t \to \infty$ and using (A.29), yields

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

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

$$BR^i(m^j) = \frac{2b^2\alpha(2-b)(1+4c)}{(4-b^2)^2 + 32c(2-b^2)} - \frac{8cb^3}{(4-b^2)^2 + 32c(2-b^2)} m^j. \tag{A.30}$$

The slope of BR^i is less than 1 in absolute value. Since the strategy sets in the fitness game are one-dimensional compact intervals (recall that $m^i, m^j \in M$, where M is a one-dimensional compact interval), the functions f^i and f^j are smooth and strictly concave in the players' own strategies, and the slopes of the best-response functions are less than 1 in absolute value, it follows from Moulin (1984, Theorem 4) that the fitness game can be solved by iterated elimination of strongly dominated strategies. The unique signal that survives this process is

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

which can be found by setting $BR^i(m^j) = m^j = m^*$ in equation (A.30). By Theorem 4, the distribution of signals converges to a unit mass at m^*. Using (A.29), the resulting types are

$$\tau^* = \theta^* = \frac{8cb^2\alpha}{2(4+2b-b^2)(1+4c) - b^3}$$

as specified in the proposition. ■

References

[1] Acemoglu, D. and M. Yildiz (2001), "Evolution of Perceptions and Play", mimeo, MIT.

[2] Alchian, A. (1950), "Uncertainty, Evolution and Economic Theory," *Journal of Political Economy*, **58**, pp. 211-221

[3] Anderson, R.M. and W.R. Zame (2001), "Genericity with Infinitely Many Parameters," *Advances in Theoretical Economics*, **1**, pp. 1-62.

[4] Bar-Gill, O. and C. Fershtman (2001) "The Limit of Public Policy: Endogenous Preferences," Foerder working paper 5-01, Tel Aviv University.

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

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

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

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

[9] Christensen, J.P.R. (1974), *Topology and Borel Structure*, North Holland Mathematical Studies Vol. 10, Amsterdam: North-Holland.

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

[11] Dubey P. (1986), "Inefficiency of Nash Equilibria," *Mathematics of Operations Research*, **11**(1), pp. 1-8.

[12] Ely J. and O. Yilankaya (2001), "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory* , **97**(2), pp. 255-272.

[13] Fershtman C. and A. Heifetz (2002), "Read My Leaps, Watch for Leaps: A Theory of Endogenous Political Instability," Foerder working paper 8-02, Tel Aviv University.

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

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

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

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

[18] Friedman M. (1953), *Essays in Positive Economics,* University of Chicago Press.

[19] Fudenberg D. and D. Levine (1998), *The Theory of Learning in Games*, MIT Press, Cambridge MA.

[20] Golubitsky, M. and V. Guillemin (1973), *Stable Mappings and Their Singularities*, Springer-Verlag, New York.

[21] Güth W. and A. Ockenfels (2001), "The Coevolution of Morality and Legal Institutions - An Indirect Evolutionary Approach," Mimeo.

[22] Güth W. and B. Peleg (2001), "When will Payoff Maximization Survive? An Indirect Evolutionary Analysis," *Journal of Evolutionary Economics*, **11**, pp. 479-499.

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

[24] Guttman, J.M. (2000), "On the Evolutionary Stability of Preferences for Reciprocity," *European Journal of Political Economy* **16**, pp. 31-50.

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

[26] Heifetz, A., E. Segev and E. Talley (2002), "Endogenous Preferences and Market Intervention," Mimeo.

[27] Hirsch, M. (1976), *Differential Topology*, Springer-Verlag, New York.

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

[29] Hunt, B.R., T. Sauer and J.A. Yorke (1992), "Prevalence: A Translation-Invariant 'Almost Every' on Infinite-Dimensional Spaces," *Bulletin (New Series) of the American Mathematical Society* **27**, pp. 217-238.

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

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

[32] Kreps, D.(1997), "Intrinsic Motivation and Extrinsic Incentives," American Economic Review, 87(2): 359 364.

[33] Kreps D. and R. Wilson (1982), "Reputation and Imperfect Information," *Journal of Economic Theory,* **27**, pp. 253-279.

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

[35] Leininger W., P.B. Linhart and R. Radner (1989), "Equilibria of the Sealed-Bid Mechanism for Bargaining with Incomplete Information," *Journal of Economic Theory*, **48**, pp.63-106.

[36] Milgrom P. and J. Roberts (1982), "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory,* **27**, pp. 280-312.

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

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

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

[40] Possajennikov A. (2000), "On The Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior and Organization*, **42**(1) pp. 125-129.

[41] Rotemberg, J.J. (1994) "Human Relation in the Workplace," *Journal of Political Economy*, **102**, pp. 684-717.

[42] Samuelson, L. (2001), "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, **97**, pp. 225-230.

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

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

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

[46] Sethi, R. and E. Somanathan (2001), "Preference Evolution and Reciprocity," *Journal of Economic Theory*, **97**, pp. 273-297.

[47] Stackelberg, H. von (1934): *Marktform und Gleichgewicht*, Vienna and Berlin: Springer.