Available online at www.sciencedirect.com
ScienceDirect
Journal of Economic Theory 133 (2007) 31–57
JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

What to maximize if you must

Aviad Heifetz[a], Chris Shannon[b,*], Yossi Spiegel[c]

[a]*The Economics and Management Department, The Open University of Israel, Israel*
[b]*Department of Economics, University of California, Berkeley, USA*
[c]*The Faculty of Management, Tel Aviv University, Israel*

Received 25 November 2003; final version received 17 May 2005
Available online 6 January 2006

Abstract

The assumption that decision makers choose actions to maximize their preferences is a central tenet in economics, often justified formally or informally by appealing to evolutionary arguments. In contrast, we show that in almost every game and for almost every family of distortions of a player's actual payoffs, some degree of this distortion is beneficial to the player, and will not be driven out by any evolutionary process involving payoff-monotonic selection dynamics. Consequently, under any such selection dynamics the population will *not* converge to payoff-maximizing behavior. We also show that payoff-maximizing behavior need not prevail when preferences are imperfectly observed.

JEL classification: C72; D01

Keywords: Evolution of preferences; Evolutionary stability; Overconfidence; Interdependent preferences

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 [2] and Friedman [20] 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 by more profitable rivals, even if none of these firms deliberately maximizes

* Corresponding author.
E-mail addresses: aviadhe@openu.ac.il (A. Heifetz), cshannon@econ.berkeley.edu (C. Shannon), spiegel@post.tau.ac.il (Y. Spiegel).

0022-0531/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2005.05.013

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 [50] 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 and for almost every family of distortions of a player's actual payoffs, some degree 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 [51], and his discussion of the commitment value of decision rules. Related ideas run through work ranging from Stackelberg's [53] classic work on timing in oligopoly to the theories of reputation in Kreps and Wilson [39], and Milgrom and Roberts [42]. For similar reasons, Frank [18,19] 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 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

[1] For a brief overview of this literature, see Samuelson [48]. Examples include Güth and Yaari [26], Huck and Oechssler [33], Fershtman and Weiss [16,17], Fershtman and Heifetz [13], Rotemberg [47], Bester and Güth [7], Possajennikov [46], Bolle [8], Bergman and Bergman [6], Koçkesen et al. [36,37], Guttman [27], Sethi and Somanathan [52], Kyle and Wang [40], Benos [5], Heifetz and Segev [28], and Heifetz et al. [29].

disposition will survive evolutionary pressures. Our results generalize this work in an important way by isolating the general principle driving these results and by showing that the evolutionary emergence of dispositions is in fact *generic*.

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 nonzero 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. Interestingly, this result also implies that, generically, players can gain a strategic advantage over opponents by hiring delegates whose preferences differ from theirs to play the game on their behalf. This implies in turn that earlier results obtained in the strategic delegation literature in the context of specific models (e.g. [23,14,15,35]) are in fact generic.

Central to our results are appropriate parameterizations of games and dispositions. Since our analysis is based on first-order conditions, we restrict attention to pure-strategy equilibria in games with continuous action sets. Because we are interested in the evolutionary viability of payoff maximization rather than the emergence of one particular type of bias, such as altruism or overconfidence, we consider a disposition to encompass a family of biases indexed by a degree that can be positive, negative, or zero. In this framework a zero degree means that the player is unbiased and interested in maximizing his actual payoff. The interpretation of a positive or negative degree will typically depend on the specification of the given family of dispositions; for example, the disposition might reflect other-regarding preferences, with a positive degree corresponding to altruism and a negative degree corresponding to spite. For a generic set of payoff functions and dispositions, however, some nonzero degree of the disposition has a positive indirect effect. This guarantees that such dispositions will not be eliminated from the population under any payoff monotonic selection dynamics. 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 are 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., [41]). 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 the population does *not* converge to payoff-maximizing behavior even if preferences are observed with noise.

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 relax the assumption that types are perfectly observed and assume instead that they are observed with noise. We show, by means of a specific example, that our main results carry over to this setting. 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, $M \leqslant N$.[2] Typical strategies are denoted by $x^i = (x_1^i, \ldots, x_M^i)$ and $x^j = (x_1^j, \ldots, x_N^j)$. 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

$$\Pi_i^i \equiv D_i \Pi^i = \left(\frac{\partial \Pi^i}{\partial x_1^i}, \ldots, \frac{\partial \Pi^i}{\partial x_M^i} \right) \quad \text{and} \quad \Pi_{ij}^i \equiv D_j \Pi_i^i = \begin{pmatrix} \dfrac{\partial^2 \Pi^i}{\partial x_1^i \partial x_1^j} & \cdots & \dfrac{\partial^2 \Pi^i}{\partial x_1^i \partial x_N^j} \\ & \ddots & \\ \dfrac{\partial^2 \Pi^i}{\partial x_M^i \partial x_1^j} & \cdots & \dfrac{\partial^2 \Pi^i}{\partial x_M^i \partial x_N^j} \end{pmatrix}.$$

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

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

$$\begin{aligned} B^i &: X^i \times X^j \times E^i \to \mathbf{R}, \\ B^j &: X^i \times X^j \times E^j \to \mathbf{R} \end{aligned}$$

are the dispositions of players i and j, and τ and θ are the players' (one-dimensional) types, which are drawn from domains E^i, $E^j \subseteq \mathbf{R}$ each containing a neighborhood of 0. 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, and 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 [7] and Possajennikov [46], write the players'

[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 2 and 3.

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

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 [17], 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. If for (τ, θ) the game has a pure strategy Nash equilibrium, let $\left(y^i(\tau, \theta), y^j(\tau, \theta)\right)$ denote such an equilibrium.[4] We assume for this discussion that the selection $\left(y^i(\tau, \theta), y^j(\tau, \theta)\right)$ from the Nash equilibrium correspondence is continuously differentiable at $(\tau, \theta) = (0, 0)$.[5] 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 this definition 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* types of player i: a unilaterally beneficial disposition might be beneficial for some types of player i but harmful for others.[6]

[4] Since the strategy spaces X^i and X^j are open, the equilibrium is interior. For a discussion of the issues of existence and interiority of pure strategy equilibria, see Remarks 1 and 3.

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

[6] 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.

To study how dispositions evolve, suppose that there are two large populations of individuals, one for each player, and with a continuum of individuals of each type. At each point $t \geqslant 0$ in time, this pair of populations is characterized by the pair of distributions $(\mathcal{T}_t, \Theta_t) \in \Delta(E^i) \times \Delta(E^j)$ of (τ, θ), where $\Delta(E^i)$ and $\Delta(E^j)$ denote the set of Borel probability distributions over E^i and E^j. We assume that $\mathcal{T}_0$ has full support over E^i and Θ_0 has full support over E^j. At each instance in time, an individual in one population is randomly matched with an individual of the other population to play the game Γ. The average fitness levels of the individuals of types τ and θ at time t are given by

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

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

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

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

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

Eqs. (2.5)–(2.7) reflect the idea that the proportion of more successful types in the population increases from one instance or period to another at the expense of less successful types. This may be due to the fact that more successful individuals have more descendants, who then inherit their parents' preferences either genetically or by education. An alternative explanation is that the decision rules of more successful individuals are imitated more often.

The same mathematical formulation is also compatible with the assumption that successful types translate into stronger *influence* rather than numerical proliferation. Under this interpretation, not all individuals are matched to play in each instance of time, and more successful individuals take part in a larger share of the economic interactions, and so are matched to play with a higher probability.

To guarantee that the 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,

$$\sup_{\tau\in E^i}\left|g^i(\tau,\Theta_t)\right| < M^i,$$

$$\sup_{\tau\in\mathbf{R}}\left|g^i(\tau,\Theta_t)-g^i(\tau,\widetilde{\Theta}_t)\right| < K^i\left\|\Theta_t-\widetilde{\Theta}_t\right\|,$$

$$\forall\Theta_t,\widetilde{\Theta}_t\in Y,$$

$$\sup_{\theta\in E^j}\left|g^j(\mathcal{T}_t,\theta)\right| < M^j,$$

$$\sup_{\theta\in\mathbf{R}}\left|g^j(\mathcal{T}_t,\theta)-g^j(\widetilde{\mathcal{T}}_t,\theta)\right| < K^j\left\|\mathcal{T}_t-\widetilde{\mathcal{T}}_t\right\|,$$

$$\forall\mathcal{T}_t,\widetilde{\mathcal{T}}_t\in Y,$$

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

Oechssler and Riedel [44, 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 differential equation (2.5) has a unique solution. [7]

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 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. [8]

[7] In addition, the boundedness of g^i and g^j guarantees that any set having positive probability under the initial distributions $\mathcal{T}_0$ or Θ_0 will have positive probability under $\mathcal{T}_t$ or Θ_t for all t. In particular, since we assumed that $\mathcal{T}_0$ and Θ_0 have full support on the domains E^i and E^j, so do $\mathcal{T}_t$ and Θ_t for all t.

[8] 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 [22]. 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.

In what follows, we will often make use of a particular class of payoff functions corresponding to games in which each pure strategy equilibrium is locally unique. We will slightly abuse terminology by referring to a pair of payoff functions (Π^i, Π^j) as a game (the strategy spaces X^i, X^j remain fixed throughout).

Definition 4 (*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 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^i_m, x^j_n and $x^i_m x^j_m$ independently and obtain a new pair of payoff functions in $\mathcal{G}$. To formalize this idea, let

$$p = \left(p^1, p^2, p^3\right) = \left(\left(p^1_1, \ldots, p^1_M\right), \left(p^2_1, \ldots, p^2_N\right), \left(p^3_1, \ldots, p^3_M\right)\right) \in \mathbf{R}^{M+N+M},$$
$$q = \left(q^1, q^2, q^3\right) = \left(\left(q^1_1, \ldots, q^1_M\right), \left(q^2_1, \ldots, q^2_N\right), \left(q^3_1, \ldots, q^3_M\right)\right) \in \mathbf{R}^{M+N+M}.$$

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

$$\bar{\Pi}^i(x^i, x^j, p) \equiv \Pi^i\left(x^i, x^j\right) + \sum_{m=1}^{M} p^1_m x^i_m + \sum_{n=1}^{N} p^2_n x^j_n + \sum_{m=1}^{M} p^3_m x^i_m x^j_m,$$
$$\bar{\Pi}^j(x^i, x^j, q) \equiv \Pi^j\left(x^i, x^j\right) + \sum_{m=1}^{M} q^1_m x^i_m + \sum_{n=1}^{N} q^2_n x^j_n + \sum_{m=1}^{M} q^3_m x^i_m x^j_m. \tag{2.8}$$

Using this notation, we assume that the manifold $\mathcal{G}$ is such that for every pair of payoff functions $\left(\Pi^i, \Pi^j\right) \in \mathcal{G}$ 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 $\left(B^i, B^j\right)$, define

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

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

While this framework and the resulting theorem allow for general combinations of sets of payoff functions $\mathcal{G}$ and sets of dispositions $\mathcal{B}$, notice that we could restrict attention to manifolds $\mathcal{G}$ and $\mathcal{B}$ such that for each $(\Pi^i, \Pi^j) \in \mathcal{G}$ and for each $(B^i, B^j) \in \mathcal{B}$, the resulting game Γ has

pure strategy Nash equilibria for all type profiles (τ, θ) in some neighborhood of $(0, 0)$ (see also Remark 1). [9]

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

Definition 5 (*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* $\left(\Pi^i, \Pi^j\right) \in \mathcal{G}$ *and dispositions* $(B^i, B^j) \in \mathcal{B}$:

(i) *The disposition* B^i *is unilaterally beneficial for player* i *and the disposition* B^j *is unilaterally beneficial for player* j.
(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 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

[9] Because the set of regular games having pure strategy equilibria is open, such combinations of sets of payoff functions and sets of dispositions exist.

in number at the expense of the type zero player. Thus in the limit the dispositions will not become extinct. [10]

Several remarks about Theorem 1 are now in order.

Remark 1. Theorem 1 is stated for general finite-dimensional manifolds of games and dispositions, which may include games that do not have pure strategy equilibria. Notice that in this case properties (i) and (ii) hold vacuously. As we discussed above, the theorem instead could be stated for collections of games and dispositions for which selections of pure strategy equilibria exist in a neighborhood of (0,0). We state the result as above for ease of use in extending the result to the general class of games, where the issues involved in restricting attention to games with pure strategy equilibria are slightly more complicated. We discuss this in more detail below.

Remark 2. 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 3. 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. [11] No 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. [10] 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 4. 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., [41]). In such cases, an

[10] For symmetric games, Güth and Peleg [25] 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.

[11] Dubey [11] and Anderson and Zame [3] employ a similar approach to demonstrate the generic Pareto-inefficiency of "non-vertex" Nash equilibria.

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. [12] 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.

Remark 5. Theorem 1 has an interesting implication for the strategic delegation literature. This literature has demonstrated that players can gain strategic advantage over rivals by hiring a delegate whose preferences differ from theirs to play the game on their behalf (e.g. [23,14,15,35]). Viewing the perceived payoff function of player i as representing the preferences of a delegate hired by player i to play the game on player i's behalf, part (i) of Theorem 1 implies that earlier results obtained in the strategic delegation literature are in fact generic. That is, in *almost every* strategic interaction hiring a delegate whose preferences differ from the player's own preferences is beneficial to the player because of its resulting effect on opponents' play.

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 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 [9] and Hunt et al. [34] have developed measure-theoretic analogues of Lebesgue measure 0 and full Lebesgue measure for infinite-dimensional spaces, called shyness and prevalence.

Definition 6 (*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 $y \in Y$. [13] 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.

If $A \subset Y$ is open, then a set $E \subset A$ is *relatively shy* in A if E is shy, and a set $F \subset A$ is *relatively prevalent* in A if $A \setminus Y$ is relatively shy in A.

Christensen [9] and Hunt et al. [34] show that shyness and prevalence have the properties we ought to require of measure-theoretic notions of "smallness" and "largeness." In particular, the

[12] In specific cases, however, there may be more natural candidates for such selections; see for example the analysis in Section 3 and in Heifetz and Segev [28].

[13] 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.

countable union of shy sets is shy, no relatively open subset is shy, prevalent sets are dense, and a subset of $\mathbf{R}^n$ is shy in $\mathbf{R}^n$ if and only if it has Lebesgue measure 0. It is straightforward to show that the corresponding properties hold for relatively shy and relatively prevalent subsets of an open set as well. Hunt et al. [34] also provide simple sufficient conditions for their notions of shyness and prevalence (here we adopt the somewhat more descriptive terminology from [3]). [14]

Definition 7 (*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 $(\Pi^i, \Pi^j, B^i, B^j) \in \mathcal{P}$:*

(i) *The disposition B^i is unilaterally beneficial for player i and the disposition B^j is unilaterally beneficial for player j.*
(ii) *The dispositions (B^i, B^j) do not asymptotically become extinct under any regular payoff-monotonic dynamics.*

In particular, let $\tilde{\mathcal{R}}_p \subset \tilde{\mathcal{G}}$ be the set of regular games with pure strategy equilibria. Then $\tilde{\mathcal{R}}_p \times \tilde{\mathcal{B}}$ contains an open, relatively prevalent subset satisfying (i) *and* (ii).

As with Theorem 1, here too we could give other versions of this result restricted to games with pure strategy Nash equilibria. This becomes somewhat more delicate, however, due to the fact that the subset of $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$ for which each game has pure strategy equilibria is not necessarily open, nor necessarily convex. The difficulty lies in extending the notion of prevalence to a relative one. Anderson and Zame [3] provide one such extension, but, crucially, they require the domain to be convex. To restrict to games with pure strategy equilibria, we have taken the simplest approach by considering the subset $\tilde{\mathcal{R}}_p \subset \tilde{\mathcal{G}}$ of regular games with pure strategy equilibria, which is open. Then it follows immediately that $\tilde{\mathcal{R}}_p \times \tilde{\mathcal{B}}$ contains an open, relatively prevalent subset satisfying (i) and (ii) above. Alternatively, given any convex subset $\mathcal{C} \subset \tilde{\mathcal{G}}_p$, one can show that there exists a relatively prevalent subset of $\mathcal{C} \times \tilde{\mathcal{B}}$ satisfying (i) and (ii). Justifying a restriction to a convex set of games with pure strategy equilibria seems difficult, however.

[14] Anderson and Zame [3] have extended the work of Hunt et al. [34] and Christensen [9] by defining prevalence and shyness relative to a convex subset that may be a shy 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 et al. [34].

3. Noisy 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, in important classes of games play can converge to a Nash equilibrium even if players have very limited knowledge or adapt their behavior myopically, for instance by following some version of fictitious play (see e.g. [21]).

In this section, we pursue further the possibility that preferences may not be perfectly observed. Specifically, we assume that players observe each other's preferences with some randomly distributed noise. The natural solution concept for this setting is Bayesian equilibrium. Unfortunately, as we discussed in the introduction, Bayesian equilibria are typically not locally unique; consequently, it is impossible to generalize Theorems 1 and 2 to this setting. Nonetheless, using a specific example that gives rise to a unique Bayesian equilibrium for any given distribution $(\mathcal{T}, \Theta)$ of types, we show that in the absence of this technical obstacle, the evolutionary viability of dispositions is maintained. 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 $(\tau, \theta) = (0, 0)$.

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

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

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

$$B^i(x^i, x^j, \tau) = \tau x^i, \quad B^j(x^i, x^j, \theta) = \theta x^j, \quad \tau, \theta \in \mathbf{R}. \tag{3.2}$$

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

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

From (3.3) it is clear that the 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.

This example can be used to illustrate our more general results. Here, if perceived payoff functions are completely observable, then any regular payoff monotonic dynamics results in a distribution of types that converges to a unit mass at a type that is positive as long as $b \neq 0$, that is, as long as the game is one with nontrivial strategic interaction. We prove this, along with some more general results, in Heifetz et al. [31].

To extend these results to a setting with partial observability, we assume that the observation of opponents' perceived payoffs is subject to some randomly distributed noise. Specifically, we assume that before choosing actions players i and j receive the following signals about each other's types:

$$s^i = \tau + v, \quad s^j = \theta + v, \tag{3.4}$$

where v is a random variable distributed on the support $[-r, r]$ according to a cumulative distribution function $\mathcal{N}$ with a positive density. The assumption that the support of v is symmetric

around 0 is not essential; however, the assumption that the support is bounded is important as it makes it possible for players to distinguish between zero and non-zero types. [15]

Given the signals, s^i and s^j, the players update their beliefs about each other's preferences, and then play a Bayesian equilibrium given these updated beliefs. In this setting we now prove the following result:

Proposition 1. *Suppose that the players have the perceived payoff functions specified in* (3.3) *and they receive the signals s^i and s^j specified in* (3.4), *and moreover, the initial distributions of both τ and θ have full support. Then the dispositions do not asymptotically become extinct under any regular payoff-monotonic selection dynamics.*

In the working paper version [30] we also establish some positive convergence results for this game with noisy observability. When the supports of the initial bias distributions $\mathcal{T}_0, \Theta_0$ are confined to some large enough compact interval, then under any regular payoff-monotonic dynamics the distributions $\mathcal{T}_t, \Theta_t$ converge weakly to a unit mass at $\tau^* = \theta^* = \frac{b^2\alpha}{4+2b-b^2}$. In particular, as in the case of full observability, this value is nonzero as long as $b \neq 0$, thus as long as there is nontrivial strategic interaction. Similar results hold if the preferences U^i, U^j are unobserved in some fraction ρ of the interactions (in which case the corresponding Bayesian equilibrium is played). Finally, similar results obtain in a version of this model incorporating costly signaling of types. Here player j observes a signal m^i of player i's type τ, where player i incurs fitness cost $c\left(m^i - \tau\right)^2$ which is convex in the distance between the signal m^i and the true type τ; and analogously for player j. Now the distributions of type–signal pairs $\left(\tau, m^i\right)$ and $\left(\theta, m^j\right)$ evolve according to some regular payoff-monotonic dynamics. Then these distributions converge to a unit mass at values that are nonzero as long as $b \neq 0$. For details and more discussion of all of these results, see Heifetz et al. [31].

4. Conclusion

The literature on the evolution of preferences, while successful in providing foundations for various types of dispositions and biases, is often criticized on two important grounds (see, e.g., [48]). 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 positive or negative 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

[15] In different but related models, Acemoglu and Yildiz [1], Dekel et al. [10], Ely and Yilankaya [12], Ok and Vega-Redondo [45] and Güth and Peleg [25] show that payoff-maximization is evolutionarily stable if preferences are completely unobservable. In our setting, this would correspond to the limit case in which the noise is distributed with an improper uniform prior on the entire real line.

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 noisy observability of preferences. Although the lack of local uniqueness of Bayesian equilibria in models in which preferences are observed with noise precludes a general extension of our results, when the Bayesian equilibrium is unique, dispositions remain evolutionarily viable in such settings in the sense that the population still does *not* converge to payoff maximizing behavior.

Acknowledgements

We are grateful for valuable comments from Joerg Oechssler, Bob Anderson, Bill Zame, Eddie Dekel, Youngse Kim, Menachem Yaari, three referees and the associate editor, and participants of the 11th European Workshop in General Equilibrium Theory.

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. [16]

Definition 8 (*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 $r \geqslant 1$. 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 $D_{x,s}G(x, s)$ has rank K.

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 6. 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

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

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 $(x^i, x^j) \in X^i \times X^j$.

[16] For example, see Hirsch [32].

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. This follows from the genericity of regular games, established in Lemma 1, and the local uniqueness of equilibria in regular games, established in Lemma 2.

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 6 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 $\phi(\cdot, \cdot, \Pi^i, \Pi^j)$ imply 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 $\left(U^i, U^j\right)$ 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 $\left(U^i, U^j\right)$ 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 6, 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 6, 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 $(\Pi^i, \Pi^j) \in \Pr_{\mathcal{G}}(\mathcal{U}_k)$,*

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

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)

$$\frac{\partial^{M}}{\partial q_1^3 \partial q_2^3 \ldots \partial q_{M}^3} \det \tilde{\Pi}_{ji}^{j}(y^i, y^j, q) = 1. \tag{A.4}$$

This follows because the determinant of $\tilde{\Pi}_{ji}^{j}(\cdot, \cdot)$ is a sum of products, of M factors each, and the derivative with respect to $(q_1^3, \ldots, q_M^3)$ of each of these products is zero with the exception of the diagonal product $\prod_{m=1}^{M} \frac{\partial^2 \Pi^j}{\partial x_m^j \partial x_m^i}$. For this term, note that

$$\frac{\partial^2 \Pi^j(y^i, y^j, q)}{\partial x_m^j \partial x_m^i} = q_m^3,$$

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

$$\prod_{m=1}^{M} \frac{\partial^2 \Pi^j(y^i, y^j, q)}{\partial x_m^j \partial x_m^i} = \prod_{m=1}^{M} q_m^3,$$

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

$$\frac{\partial^{M}}{\partial q_1^3 \partial q_2^3 \ldots \partial q_{M}^3} \left(\prod_{m=1}^{M} \frac{\partial^2 \Pi^j(y^i, y^j, q)}{\partial x_m^j \partial x_m^i} \right) = 1.$$

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

$$\psi(\cdot, \cdot, \Pi^i, \Pi^j) \equiv \begin{pmatrix} \Pi_i^i(\cdot, \cdot) \\ \Pi_j^j(\cdot, \cdot) \\ \dfrac{\partial^{M-\ell}}{\partial q_1^3 \partial q_2^3 \ldots \partial q_M^3} \det \tilde{\Pi}_{ji}^{j}(\cdot, \cdot, 0) \end{pmatrix} : X^i \times X^j \times \mathcal{G} \to \mathbf{R}^{M+N+1}. \tag{A.5}$$

To see this, note that the derivative

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

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 6, 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, $\mathcal{U}_\ell$ is an open subset of $\mathcal{U}_M$ by the transversality theorem. □

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* $\left(y^i, y^j\right)$ *of* $\left(\Pi^i, \Pi^j\right)$, $y^j_\tau(0,0) \neq 0$.

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 $(\Pi^i, \Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_k)$,

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

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* $\left(\Pi^i, \Pi^j\right)$,

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

Proof. Fix $(\Pi^i, \Pi^j) \in \tilde{\mathcal{G}}$ and $(B^i, B^j) \in \tilde{\mathcal{B}}$. For each $(x^i, x^j) \in X^i \times X^j$ and for each n, denote by $J_n(x^i, x^j)$ the $(M+N) \times (M+N)$ matrix obtained from

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

after replacing the nth column by

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

Let $z : X^i \times X^j \times \tilde{\mathcal{G}} \times \tilde{\mathcal{B}} \to \mathbf{R}^N$ be given by

$$\begin{aligned} &z(x^i, x^j, \Pi^i, \Pi^j, B^i, B^j) \\ &\quad = \left(\det J_1(x^i, x^j), \ldots, \det J_n(x^i, x^j), \ldots, \det J_N(x^i, x^j)\right). \end{aligned}$$

In particular, note that z is independent of p^2. Now let $\zeta : X^i \times X^j \times \tilde{\mathcal{G}} \times \tilde{\mathcal{B}} \to \mathbf{R}^{M+N+1}$ be given by

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

For the remainder of the argument, we restrict ζ to the set $\mathcal{U}_M$ defined in Lemma 4. Fix $(\Pi^i, \Pi^j) \in \mathrm{Pr}_{\mathcal{G}}(\mathcal{U}_M)$ and $(B^i, B^j) \in \mathrm{Pr}_{\mathcal{B}}(\mathcal{U}_M)$; by definition (Π^i, Π^j) is a regular game. Now let (y^i, y^j) be a Nash equilibrium of (Π^i, Π^j). By (A.1) and Cramer's rule,

$$\begin{aligned} y^j_\tau(0,0) &= \left(\ldots, \frac{\det J_n(y^i, y^j)}{\det \begin{pmatrix} \Pi^i_{ii}(y^i, y^j) & \Pi^i_{ij}(y^i, y^j) \\ \Pi^j_{ji}(y^i, y^j) & \Pi^j_{jj}(y^i, y^j) \end{pmatrix}}, \ldots \right) \\ &= \frac{1}{\det \begin{pmatrix} \Pi^i_{ii}(y^i, y^j) & \Pi^i_{ij}(y^i, y^j) \\ \Pi^j_{ji}(y^i, y^j) & \Pi^j_{jj}(y^i, y^j) \end{pmatrix}} z(y^i, y^j, \Pi^i, \Pi^j, B^i, B^j). \end{aligned}$$

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

$$D_{p^2}\left(\bar{\Pi}^i_j(y^i,y^j,p)z(y^i,y^j)\right)=z(y^i,y^j,\Pi^i,\Pi^j,B^i,B^j)$$
$$=\det\begin{pmatrix}\Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j)\\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j)\end{pmatrix}y^j_\tau(0,0).$$

By Lemma 5, $y^j_\tau(0,0)\neq 0$, and because (Π^i,Π^j) is a regular game,

$$\det\begin{pmatrix}\Pi^i_{ii}(y^i,y^j) & \Pi^i_{ij}(y^i,y^j)\\ \Pi^j_{ji}(y^i,y^j) & \Pi^j_{jj}(y^i,y^j)\end{pmatrix}\neq 0.$$

Thus if (y^i,y^j) is a Nash equilibrium of (Π^i,Π^j), then the derivative

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

has rank $M+N+1$. Consequently, by Remark 5, 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 $\left(\Pi^i+B^i,\Pi^j+B^j\right)\in\mathcal{U}^*$. 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 $\left(\Pi^i+B^i,\Pi^j+B^j\right)\in\mathcal{U}^*$. Since the derivative

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

has only $M+N$ columns, it cannot have rank $M+N+1$. By Remark 5, 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 $\left(x^i,x^j\right)\in X^i\times X^j$. 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

$$\Pi^i_j(y^i,y^j)z(y^i,y^j,\Pi^i,\Pi^j,B^i,B^j)\neq 0.$$

Using this together with the fact that (Π^i,Π^j) is a regular game yields $\Pi^i_j(y^i,y^j)y^j_\tau(0,0)\neq 0$ 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

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

where (y^i, y^j) is a Nash equilibrium of $\left(\Pi^i, \Pi^j\right)$. Hence $\Pi^i_i(y^i, y^j) = 0$. 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 $f^i_\tau(0,0) \neq 0$ 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* $\left(B^i, B^j\right)$ *become asymptotically extinct in the game* Γ, *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 the unit mass at 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 $\left(B^i, B^j\right)$ becomes asymptotically extinct, there exists t' sufficiently large so that for every $t \geqslant t'$, $\Theta_t \in A$, and hence for every $t \geqslant t'$, $\int f^i(\tilde{\tau},\theta)\,d\Theta_t > \int f^i(\hat{\tau},\theta)\,d\Theta_t$ for any $\tilde{\tau} \in V_\tau$ and $\hat{\tau} \in V_0$. Because the dynamics are regular, $\mathcal{T}_t$ and Θ_t have full support for each t (see footnote 7). Then, using (2.6), the growth rates satisfy $g^i(\tilde{\tau},\Theta_t) > g^i(\hat{\tau},\Theta_t)$ for every $t \geqslant t'$, $\tilde{\tau} \in V_\tau$ and $\hat{\tau} \in V_0$ as well. By (2.5), this implies that for $t \geqslant t'$ 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 a unit mass at $\tau = 0$, a contradiction. □

Proof of Theorem 1. Lemma 7 proves the existence of an open, full-measure set of perceived payoffs $\mathcal{U}^*$ such that B^i is unilaterally beneficial to player i. An analogous proof establishes the existence of an open, full-measure set of perceived payoffs $\mathcal{U}^{**}$ such that B^j is unilaterally beneficial to player j. Part (i) of the theorem follows by observing that the intersection of $\mathcal{U}^*$ and $\mathcal{U}^{**}$ is also an open and full-measure set of perceived payoffs. As for part (ii), Lemma 7 implies that for perceived payoffs in $\mathcal{U}^*$, $(\tau,\theta) = (0,0)$ is not a Nash equilibrium of the fitness game, and by Lemma 8 it follows that for $\left(U^i, U^j\right) \in \mathcal{U}^*$, the dispositions $(B^i, B^j) = \mathrm{Pr}_{\mathcal{B}}\left(U^i, U^j\right)$ do not become asymptotically extinct in the game $(\Pi^i, \Pi^j) = \mathrm{Pr}_{\mathcal{G}}\left(U^i, U^j\right)$. □

Proof of Theorem 2. With ζ as defined in the proof of Lemma 6, let

$$\mathcal{P} = \left\{\left(\Pi^i, \Pi^j, B^i, B^j\right) \in \tilde{\mathcal{G}} \times \tilde{\mathcal{B}} : \left(\Pi^i, \Pi^j\right) \text{ is regular and 0 is a regular value of } \zeta\right\}.$$

By the arguments in Lemmas 6 and 7, every $\left(\Pi^i, \Pi^j, B^i, B^j\right) \in \mathcal{P}$ 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. This follows from the fact that the set of regular games $\tilde{\mathcal{R}}$ is open in $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$, that the set of functions in $C^1(X^i \times X^j, \mathbf{R}^{M+N+1})$ transverse to $\{0\}$ is open in the Whitney C^1 topology, and from the fact that ζ is continuous on $\tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$ in the Whitney C^1 topology.

Now let

$$\mathcal{V} = \Bigg\{ (\widehat{\Pi}^i, \widehat{\Pi}^j) \in \tilde{\mathcal{G}} \mid \widehat{\Pi}^i(x^i, x^j) = \sum_{m=1}^{M} p_m^1 x_m^i + \sum_{n=1}^{N} p_n^2 x_n^j + \sum_{m=1}^{M} p_m^3 x_m^i x_m^j$$
$$\text{for some } p \in \mathbf{R}^{M+N+M},$$
$$\widehat{\Pi}^j(x^i, x^j) = \sum_{m=1}^{M} q_m^1 x_m^i + \sum_{n=1}^{N} q_n^2 x_n^j + \sum_{m=1}^{M} q_m^3 x_m^i x_m^j$$
$$\text{for some } q \in \mathbf{R}^{M+N+M} \Bigg\}$$

and

$$\mathcal{W} = \Bigg\{ (\widehat{B}^i, \widehat{B}^j) \in \tilde{\mathcal{B}} \mid \widehat{B}^i(x^i, x^j, \tau) = \tau \sum_{m=1}^{M} v_m x_m^i \quad \text{for some } v \in \mathbf{R}^M,$$
$$\widehat{B}^j(x^i, x^j, \theta) = \theta \sum_{n=1}^{N} w_n x_n^j \quad \text{for some } w \in \mathbf{R}^N \Bigg\}.$$

Now by Theorem 1, for every $\left(\Pi^i, \Pi^j, B^i, B^j\right) \in \tilde{\mathcal{G}} \times \tilde{\mathcal{B}}$, $\left[(\mathcal{V} \times \mathcal{W}) + \left(\Pi^i, \Pi^j, B^i, B^j\right)\right] \cap \mathcal{P}$ has full measure in $\mathcal{V} \times \mathcal{W}$. Equivalently, $\left(\mathcal{P} - \left(\Pi^i, \Pi^j, B^i, B^j\right)\right) \cap (\mathcal{V} \times \mathcal{W})$ 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. □

Proof of Proposition 1. Before choosing their actions, the players 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 produce 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. Player j's problem 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

$$x^i = \frac{\alpha + \tau - b\chi^j(s^i, s^j)}{2}, \quad x^j = \frac{\alpha + \theta - b\chi^i(s^i, s^j)}{2}. \tag{A.9}$$

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.9) yields

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

Solving this pair of equations yields

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

Substituting this in (A.9) reveals that the equilibrium actions of players i and j are

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

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

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

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 $\tau > 0$,

$f^i\left(\tau, s^i\right)$ is increasing in $\tau(s^i)$. Consequently, the overall average fitness of player i with type τ will be bounded from below asymptotically by

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

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

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

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. □

References

[1] D. Acemoglu, M. Yildiz, Evolution of perceptions and play, MIT Press, Cambridge, MA, 2001, Mimeo.
[2] A. Alchian, Uncertainty, evolution and economic theory, J. Polit. Economy 58 (1950) 211–221.
[3] R.M. Anderson, W.R. Zame, Genericity with infinitely many parameters, Advances Theoretical Econ. 1 (2001) 1–62.
[5] A.V. Benos, Aggressiveness and survival of overconfident traders, J. Finan. Markets 1 (1998) 353–383.
[6] N. Bergman, Y. Bergman, Ecologies of preferences with envy as an antidote to risk-aversion in bargaining, The Hebrew University of Jerusalem, 2000, Mimeo.
[7] H. Bester, W. Güth, Is altruism evolutionary stable?, J. Econ. Behav. Organ. 34 (2) (1998) 211–221.
[8] F. Bolle, Is altruism evolutionarily stable? And envy and malevolence?—Remarks on Bester and Güth, J. Econ. Behav. Organ. 42 (1) (2000) 131–133.
[9] J.P.R. Christensen, Topology and borel structure, North Holland Mathematical Studies, vol. 10, North-Holland, Amsterdam, 1974.
[10] E. Dekel, J. Ely, O. Yilankaya, Evolution of preferences, Northwestern University, 1998, Mimeo.
[11] P. Dubey, Inefficiency of Nash equilibria, Mathematics Operations Res. 11 (1) (1986) 1–8.
[12] J. Ely, O. Yilankaya, Nash equilibrium and the evolution of preferences, J. Econ. Theory 97 (2) (2001) 255–272.
[13] C. Fershtman, A. Heifetz, Read my lips, watch for leaps: a theory of endogenous political instability, Foerder Working Paper 8-02, Tel Aviv University, 2002.
[14] C. Fershtman, K. Judd, Incentive equilibrium in oligopoly, Amer. Econ. Rev. 77 (5) (1987) 927–940.
[15] C. Fershtman, K. Judd, E. Kalai, Observable contracts: strategic delegation and cooperation, Int. Econ. Rev. 32 (3) (1991) 551–559.
[16] C. Fershtman, Y. Weiss, Why do we care about what others think about us?, in: A. Ben Ner, L. Putterman (Eds.), Economics, Values and Organization, Cambridge University Press, Cambridge, MA, 1997.
[17] C. Fershtman, Y. Weiss, Social rewards, externalities and stable preferences, J. Public Econ. 70 (1998) 53–74.
[18] R. Frank, If homo economicus could choose his own utility function, would he choose one with a conscience?, Amer. Econ. Rev. 77 (4) (1987) 593–604.
[19] R. Frank, Passions within Reason—The Strategic Role of the Emotions, W.W. Norton & Company, New York, 1988.
[20] M. Friedman, Essays in Positive Economics, University of Chicago Press, 1953.
[21] D. Fudenberg, D. Levine, The Theory of Learning in Games, MIT Press, Cambridge, MA, 1998.
[22] M. Golubitsky, V. Guillemin, Stable Mappings and Their Singularities, Springer, New York, 1973.
[23] J. Green, Commitment with third parties, Ann. Econ. Statist. 25–26 (1992) 101–121.
[25] W. Güth, B. Peleg, When will payoff maximization survive? An indirect evolutionary analysis, J. Evolutionary Econ. 11 (2001) 479–499.

[26] W. Güth, M. Yaari, Explaining reciprocal behavior in simple strategic games: an evolutionary approach, in: U. Witt (Ed.), Explaining Forces and Changes: Approaches to Evolutionary Economics, University of Michigan Press, 1992.
[27] J.M. Guttman, On the evolutionary stability of preferences for reciprocity, Europ. J. Polit. Econ. 16 (2000) 31–50.
[28] A. Heifetz, E. Segev, The evolutionary role of toughness in bargaining, Games Econ. Behav. 49 (2004) 117–134.
[29] A. Heifetz, A., E. Segev, E. Talley, Market design with endogenous preferences, 2004, Mimeo.
[30] A. Heifetz, C. Shannon, Y. Spiegel, What to maximize if you must, CORE Discussion Paper 2003/47, 2003.
[31] A. Heifetz, C. Shannon, Y. Spiegel, The dynamic evolution of dispositions, 2004, Mimeo.
[32] M. Hirsch, Differential Topology, Springer, New York, 1976.
[33] S. Huck, J. Oechssler, The indirect evolutionary approach to explaining fair allocations, Games Econ. Behav. 28 (1999) 13–24.
[34] B.R. Hunt, T. Sauer, J.A. Yorke, Prevalence: a translation-invariant 'almost every' on infinite-dimensional spaces, Bull. (New Series) Amer. Math. Society 27 (1992) 217–238.
[35] M. Katz, Game-playing agents: unobservable contracts as precommitments, RAND J. Econ. 22 (1991) 307–328.
[36] L. Koçkesen, E.A. Ok, R. Sethi, Evolution of interdependent preferences in aggregative games, Games Econ. Behav. 31 (2000) 303–310.
[37] L. Koçkesen, E.A. Ok, R. Sethi, The strategic advantage of negatively interdependent preferences, J. Econ. Theory 92 (2000) 274–299.
[39] D. Kreps, R. Wilson, Reputation and imperfect information, J. Econ. Theory 27 (1982) 253–279.
[40] A.S. Kyle, A. Wang, Speculation duopoly with agreement to disagree: can overconfidence survive the market test?, J. Finance LII (1997) 2073–2090.
[41] W. Leininger, P.B. Linhart, R. Radner, Equilibria of the sealed-bid mechanism for bargaining with incomplete information, J. Econ. Theory 48 (1989) 63–106.
[42] P. Milgrom, J. Roberts, Predation, reputation, and entry deterrence, J. Econ. Theory 27 (1982) 280–312.
[44] J. Oechssler, F. Riedel, Evolutionary dynamics on infinite strategy spaces, Econ. Theory 17 (2001) 141–162.
[45] E.A. Ok, F. Vega-Redondo, On the evolution of individualistic preferences: an incomplete information scenario, J. Econ. Theory 97 (2001) 231–254.
[46] A. Possajennikov, On the evolutionary stability of altruistic and spiteful preferences, J. Econ. Behav. Organ. 42 (1) (2000) 125–129.
[47] J.J. Rotemberg, Human relation in the workplace, J. Polit. Econ. 102 (1994) 684–717.
[48] L. Samuelson, Introduction to the evolution of preferences, J. Econ. Theory 97 (2001) 225–230.
[50] A. Sandroni, Do markets favor agents able to make accurate predictions, Econometrica 68 (2000) 1303–1341.
[51] T. Schelling, The Strategy of Conflict, Harvard University Press, Cambridge, MA, 1960.
[52] R. Sethi, E. Somanathan, Preference evolution and reciprocity, J. Econ. Theory 97 (2001) 273–297.
[53] H. von Stackelberg, Marktform und Gleichgewicht, Springer, Vienna, Berlin, 1934.