Evolution of Preferences

by

Eddie Dekel, Jeffrey C. Ely and Okan Yilankaya

JUNE 2004

Discussion Paper No.: 04-12

DEPARTMENT OF ECONOMICS
THE UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, CANADA V6T 1Z1

http://www.econ.ubc.ca

Evolution of Preferences

Eddie Dekel, Jeffrey C. Ely and Okan Yilankaya[1]

June 2004
(First Version: July 1998)

[1]Dekel: Department of Economics, Northwestern University and Eitan Berglas School of Economics, Tel Aviv University, dekel@northwestern.edu; Ely: Department of Economics, Northwestern University, ely@bu.edu; Yilankaya: Department of Economics, University of British Columbia, okan@interchange.ubc.ca. We thank seminar participants at the Econometric Society World Congress (Seattle, 2000), Conference on Economic Design (Istanbul, 2000), Koc, Michigan, Northwestern, Quenn's, Rochester, Simon Fraser, Southampton, and UBC. Dekel and Yilankaya thank NSF and SSHRC, respectively, for research support. An earlier version of this paper appeared in Yilankaya (1999).

Abstract

We model, using evolutionary game theory, the implications of endogenous determination of preferences over the outcomes of any given two-player normal form game, G. We consider a large population randomly and repeatedly matched to play G. Each individual has a preference relation over the outcomes of G which may be different than the "true" payoff function in G, and makes optimal choices given her preferences. The evolution of preferences is driven by the payoffs in G that each player obtains. We define stable outcomes (of G) as arising from the stable points of the evolutionary process described above. In our most general model players know the distribution of preferences in the population and observe their opponents' preferences with a probability $p \in [0, 1]$. They then play a (Bayesian) Nash equilibrium of the resulting game of incomplete information. In the case in which players can perfectly observe their opponents' preferences, i.e., $p = 1$, (where the game is actually one of complete information) an outcome is stable only if it is efficient. Also, an efficient outcome which arises from a strict Nash equilibrium is stable. We also characterize, for 2×2 games, both the stable outcomes and the stable distributions of preferences in the population. When preferences are unobservable, i.e., $p = 0$, we show that stability in our model of evolution of preferences coincides with the notion of neutrally stable strategy (NSS). Finally, we consider robustness of these results. The necessity and sufficiency results are robust to slight changes in p, except for the sufficiency of NSS when $p = 0$: There are in fact (Pareto-inferior) risk-dominant strict equilibria that are not stable for any $p > 0$.

1 Introduction

Economists, traditionally, have taken preferences of individuals as given and have refrained from trying to explain them. In recent years interest has arisen in trying to understand how preferences are formed, how they change through time, and how all of these processes influence economic activity. Evolutionary analysis seems to be particularly useful in tackling these questions. The main idea, borrowed from evolutionary biology, is that "successful rules" are going to proliferate, replacing unsuccessful ones.[1]

In this paper, we analyze, using evolutionary game theory, the behavioral implications of endogenous determination of preferences over the outcomes of any given two-player normal form game. The basic line of reasoning is as follows: Preferences lead to behavior, behavior determines "success", and success regulates the evolution of preferences.[2] We start with a symmetric two-player normal form game G. We imagine a large population randomly and repeatedly matched to play G. We interpret, as is standard in evolutionary game theory, the payoffs as representing "fitness", implying that evolution is driven by these payoffs. We allow each individual to have a preference relation on outcomes of G which may be different than the payoff functions. In other words, we allow "subjective" preferences to diverge from "objective" payoffs.[3] We do not, a priori, restrict the preferences that people may have; any preference relation may be represented in the population. Therefore it is possible to endogenize preferences and study their evolution. We assume that each individual makes optimal choices *given her preferences*, so that when two people are matched they play an equilibrium of the "game" given their preferences and the action set of G.[4] This is what we mean when we

[1] Some of the early proponents of this idea are Becker (1976), Frank (1987), Hirshleifer (1977), and Rubin and Paul (1979). More recent works include Guth (1995), Guth and Yaari (1992), Hansson and Stuart (1990), Robson (1996), Rogers (1994), and Waldman (1994). Also see Dekel and Scotchmer (1999), Ely and Yilankaya (2001), Fershtman and Weiss (1996), Guth and Bester (1998), Ok and Vega-Redondo (2001), Robson (2001), Sethi and Somanathan (2001), and Samuelson (2001) for a brief introduction on evolution of preferences.

[2] The "indirect evolutionary approach" was pioneered by Guth and Yaari (1992) and Guth (1995).

[3] People do not always act according to their pure self-interest (as narrowly defined), altruism, fairness, envy, etc. may effect their behavior. See, for example, the survey of Rabin (1998).

[4] To define a game one also needs to specify what each person knows about her oppo-

say preferences lead to behavior.

"Success" is measured by fitness, i.e., the payoffs of the game G. Therefore, the behavior of an individual, i.e., her part of the equilibrium play, will determine her success. Finally, the evolution of preferences will be driven by their success: The proportion of people with preferences that have yielded higher fitness will increase at the expense of those who have preferences that have yielded lower fitness. This can be the result of inheritance of preferences by children from their parents. The inheritance need not be genetic; it can be by children's emulation of their parents. Also, children can inherit preferences from their "cultural parents". (See Cavalli-Sforza and Feldman (1981) and Boyd and Richerson (1985) for the theory of cultural evolution.)

We look at the stable points of the evolutionary process described above to analyze the preferences (on outcomes of G) that we expect to see in the long run.[5] However, our main emphasis is not on the evolution of preferences *per se*, but on the implications of it for the outcome of the game we want to analyze.

In this paper we analyze the implications of *evolution of preferences*. The evolutionary game theory literature began by studying the *evolution of behavior*: behavior determines success which in turn regulates the evolution of behavior. These models of strategic interaction, where each player is committed to a particular strategy (at least for a significant amount of time), are heavily influenced from biological models of genetically determined behavior. As such, we believe, they underestimate the cognitive abilities of human beings.[6] In our model people behave rationally given their preferences. Hence, our approach can be thought as a way to embed a more sophisticated model of behavior within the natural selection paradigm. This approach of studying the evolution of preferences has been used to study evolution of various particular forms of preferences, and more recently also general preferences as we do here.[7]

For the evolutionary dynamics (or static stability concepts, like "evolutionary stable strategy", that are inspired by them) to be plausible, it is necessary that there is some *inertia*, i.e., people's choices are "locked-in",

nent's preferences. We elaborate on this issue below.

[5] In this paper, like in many models in the literature, we use a static stability concept that is inspired by evolutionary dynamics and supposed to capture its essential implications.

[6] See, for example, Fudenberg and Levine (1998) for a similar criticism.

[7] See Footnote 1.

at least for a significant amount of time. Evolution of preferences has an advantage in this respect too. As opposed to strategies in a game, one cannot change her preferences over that game easily, if at all. "Preference" is arguably a more primitive concept than "behavior".

We assume that when two players are matched they play an "equilibrium". What the "game" is, and hence which equilibrium concept is appropriate to use, depends on what players observe about their opponents' preferences. In addition to the cases where each player fully observes her opponent's preferences and does not observe anything at all, we also consider an intermediate case in which a player observes her opponent's preference with probability $p \in (0, 1)$.

In the perfectly observable preferences case, we assume that when two players are matched a Nash equilibrium of the game (given by the action set of G and their preferences) is played. After defining stability (of outcomes of G), which is very much in the spirit of the concept of *neutrally stable strategy*, we investigate the properties of stable outcomes. First, all the incumbents receive the same fitness in each of their matches with other incumbents, and, second, the average fitness they obtain must equal to the fitness of a symmetric strategy profile in G. We call a strategy *efficient* if its fitness when played against itself is at least as large as the fitness of any other symmetric strategy profile. Our first result implies that the efficient strategy gives the highest feasible fitness that any stable outcome can generate. We show that it also gives the lowest feasible fitness of any stable outcome. Thus, efficiency is a necessary condition for stability: An outcome is stable only if the average fitness of each type in the population is equal to that of the efficient strategy. This result is in contrast to the "stable only if Nash" folk theorem of the evolution of behavior paradigm.[8] We then show that if a strategy is efficient and constitutes a strict Nash equilibrium when it is played against itself, then the outcome of that equilibrium is stable. By combining the last two results, we can see that the unique stable outcome in much studied coordination games will be that of the "good equilibrium".

We next restrict our attention to 2×2 games. In this case we are able to characterize the stable outcomes. Efficiency of a pure strategy is *sufficient*, as well as necessary, for the corresponding outcome to be stable. In particular, in the Prisoners' Dilemma game, "cooperation" is the unique stable outcome

[8] This folk theorem about only Nash equilibrium outcomes being stable is a constant theme, for example, in Mailath (1998), Samuelson (1997), and Weibull (1995).

(as long as the strategy "cooperate" is efficient, of course). Games with a mixed efficient strategy, on the other hand, do not have stable outcomes with the exception of a nongeneric class of Hawk-Dove games. We also characterize the set of stable distributions of preferences for 2×2 games.

We then study the case in which players do not observe their opponents' preferences, but know the distribution of preferences in the population. We assume that, given any distribution of preferences, the aggregate play in the population corresponds to a Bayesian-Nash equilibrium of the game where a player's type is her preference. That is, we will assume that each individual, upon being selected to play, will have correct beliefs about the distribution over her opponents' play and will choose a (possibly mixed) action that is a best-reply, according to her own preferences, to this belief. We show that an outcome induced by a symmetric strategy profile (which are the only possible ones that can result from the symmetric interaction we analyze) is stable iff that strategy is neutrally stable. Thus, the tendency for efficiency in the observable preferences case disappears if preferences are unobservable, and the "stable only if Nash" folk theorem of evolutionary game theory is revived. The intuition is simple: If nobody can observe your preferences, and hence condition their behavior on that, there is no advantage in having preferences that are different from the one given by the fitness function.

Finally, we consider the case in which each player observes her opponent's type with probability $p \in (0, 1)$. We are particularly interested whether the stability results are "continuous" in p, so that the results for the observable (respectively, unobservable) preferences case hold for high (respectively, low) enough p. First, we show that if a strategy is efficient and constitutes a strict Nash equilibrium when it is played against itself, then the outcome of that equilibrium is stable for any p. We then show that "stable only if Nash" result of the unobservable preferences and "stable only if efficient" result of the observable preferences are robust to slight changes in p: If a symmetric pure-strategy profile is not a Nash equilibrium, then its outcome is not stable for p low enough. Likewise, the outcome of a symmetric pure-strategy profile is not stable for high enough p, if that strategy is not efficient. The sufficiency result of the unobservable preferences case, however, breaks down. We provide a coordination game example, where the outcome of a risk-dominant (but payoff-dominated) strict Nash equilibrium is not stable for *any* $p > 0$.

The rest of the paper is organized as follows: We introduce the model in Section 2. We analyze the cases of observable, unobservable and imperfectly

observable preferences in Sections 3, 4 and 5, respectively. All the proofs are in the Appendix.

2 The Model

We start with a symmetric two-player normal form game G with a finite action set $A = \{a_1, a_2, ..., a_n\}$, and a payoff function $\pi : A \times A \to \mathbf{R}$. We interpret, as is standard in the evolutionary game theory literature, the payoffs as having an objective meaning, e.g., as representing "fitness". The survival of a player depends on her success in the game, which is evaluated according to the fitness function π. Let Δ represent the set of mixed strategies in G; the payoff function π extends naturally to $\Delta \times \Delta$. If $a_i \in A$, then we identify a_i with the element of Δ which assigns probability one to a_i, and we adopt this convention for all probability distributions throughout the paper. Let $\mathcal{O}$ represent the set of *outcomes* in G, i.e., the set of probability distributions on $A \times A$.

We imagine a large population randomly and repeatedly matched to play the game G. In standard evolutionary models each player is assumed to play a particular strategy in G. We, instead, allow each player to have (von Neumann-Morgenstern) preferences over outcomes in G which may be different than π. In other words, we allow "subjective" preferences to diverge from "objective" fitness. Let $\Theta \equiv [0,1]^{n^2}$ be the set of all possible (modulus affine transformations) preference relations on $A \times A$. We will often refer to the elements of Θ as "type"s. The environment will be characterized by a probability distribution on Θ, representing the distribution of preferences in the population. We will restrict attention to distributions that have finite supports. Let $\mathcal{P}(\Theta)$ be the set of all possible finite support probability distributions on Θ. Finally, let $C(\mu)$ denote the support of $\mu \in \mathcal{P}(\Theta)$.

We assume that each player is behaving "rationally" given her preferences and the "information structure", i.e., what the players observe about their opponents' types. In particular, we assume that an equilibrium (Nash or Bayesian-Nash) of the "game", which is readily defined given the information structure, is played when two players are matched. Given a particular equilibrium, we can determine the expected fitness to each type θ in the population, on which the evolution of the type distribution depends. While it is not a part of our formal model, we view the equilibrium play as arising from a process of learning which operates much faster than the evolutionary

process we model. Our model focuses on evolution of the type distribution μ. We suppose that whenever a new distribution ν arises as a consequence of evolutionary forces, the learning process always reaches equilibrium play (given ν) before subsequent evolution of types proceeds.

We consider three possible information structures, which we study in turn in the next three sections. In the first, when two players are matched they observe each other's preferences. In the second, we consider the case where they do not observe anything. Finally, in the third, we assume that when two players are matched, each player independently observes her opponent's preferences with probability $p \in (0, 1)$.

The stability criterion we use is static, and like its counterparts in the literature, is intended to capture the effects of the evolutionary dynamics discussed above. We will supply a different definition of stability for each of the three cases we consider. It should be clear from the discussions of these definitions that they follow from the same considerations, and can easily be combined in a single definition, but only by introducing even heavier notation throughout.

3 Observable Preferences

We assume that when θ and θ' are matched a Nash equilibrium of the game (given by the action set A for both players and payoff functions θ and θ') is played. Let $B(\mu)$ denote the set of all *equilibrium configurations* when the distribution of preferences in the population is given by μ, i.e., each $b \in B(\mu)$ specifies a Nash Equilibrium for all possible matches between all the types represented in the population. Formally, $B(\mu)$ is the set of all functions $b : C(\mu) \times C(\mu) \to \Delta \times \Delta$, such that $b(\theta, \theta') = (b_1(\theta, \theta'), b_2(\theta, \theta'))$ is a Nash equilibrium of the game given by θ and θ', where $b_1(\theta, \theta')$ and $b_2(\theta, \theta')$ are (possibly mixed) actions taken by, respectively, θ and θ'. Since we do not allow players to condition their actions on their positions in the game, we have $b_1(\theta, \theta') = b_2(\theta', \theta)$, so that, in the particular equilibrium configuration chosen, the action that θ takes when matched against θ' is the same regardless of her being a row or a column player. Notice that each b induces an outcome in G, which we denote by $o(b)$.[9]

Given a preference distribution in the population μ and an equilibrium

[9] We drop the arguments in $b(.)$ to simplify the exposition.

configuration b, we can calculate the expected fitness that $\theta \in \Theta$ gets:

$$\Pi_\theta(\mu \mid b) = \sum_{\theta' \in C(\mu)} \pi(b_1(\theta, \theta'), b_2(\theta, \theta'))\mu(\theta'),$$

where $\mu(\theta')$ is the population share of θ'.

A preference distribution is stable with respect to outcome x via b if x is the outcome induced by the equilibrium configuration b and every type in the population obtains the same expected fitness.

Definition 1 *A distribution μ is **stable with respect to outcome** x **via** b if $x = o(b)$ for some $b \in B(\mu)$ and $\Pi_\theta(\mu \mid b) = \Pi_{\theta'}(\mu \mid b)$ for all $\theta, \theta' \in C(\mu)$.*

We are interested in what the evolution of preferences implies about the play in the game the population is playing, G. The stability definition we use is very much in the spirit of the concept of *neutrally stable strategy* (NSS).[10] It is static and intends to capture the implications of evolutionary selection. When a mutant enters in small proportion, the incumbents should not have lower average fitness than the mutant, in other words the population should be immune to invasions by mutants.[11] Of course, this idea cannot, readily and easily, be translated into a simple definition in our setting. In the standard evolutionary game theory literature, each type (and hence the mutant) is committed to play a particular strategy. However, here each type is a preference relation; so, a mutant may take a different action when matched against different incumbents. Moreover, the mutant's action against any given incumbent may be indeterminate as well. The only restriction on the action of a mutant θ when matched against an incumbent θ', is that it has to be θ's strategy in *some* equilibrium of the game given by θ and θ'. Our stability definition is a rather strong one: we require immunity of the population against the invasion of any mutant and we consider *all* of the equilibria between the mutant and any incumbent. The idea is that when θ and θ' are matched there may be multiple equilibria and we do not know which of these will be played, so we check the stability against each of them. One can think of a

[10] A strategy σ is an NSS if for every strategy σ' there exists $\varepsilon' \in (0,1)$ such that, $\pi(\sigma, \varepsilon\sigma' + (1-\varepsilon)\sigma) \geq \pi(\sigma', \varepsilon\sigma' + (1-\varepsilon)\sigma)$ for all $\varepsilon \in (0, \varepsilon')$.

[11] However, this requirement does not imply that mutants will be driven out. We, nevertheless, use an NSS-type definition (instead of, say, ESS-type), since many preference relations are behaviorally equivalent. Moreover, we conjecture that qualitatively similar results will hold with a set-valued stability concept.

random process (maybe random only to an outside observer) that determines which equilibrium will be played. As long as each of those equilibria has a positive probability of being played, we have to check against all of them.[12] However, we have one restriction on the equilibria played in the post-entry population. We require that when two incumbents are matched against each other, they continue to play the same equilibrium after the mutant enters. The entry of a mutant should not have any effect on the play within the incumbent population, especially when the very question is the stability of that play. This is definitely the case when the equilibrium play is the result of a learning process: If this process settled on a particular equilibrium in the match between two incumbents, then why would the entrance of a mutant cause a shift to another equilibrium?

Definition 2 *An outcome x is **stable** (with μ and b) if there exists a distribution μ and an equilibrium configuration $b \in B(\mu)$ such that μ is stable with respect to x via b and the following condition holds:*

$\forall\theta\ \exists\varepsilon' > 0$ such that $\forall\varepsilon \in (0, \varepsilon')$, $\forall\theta' \in C(\mu)$, and $\forall\bar{b} \in B((1-\varepsilon)\mu+\varepsilon\theta \mid b)$, we have:

$$\Pi_{\theta'}((1-\varepsilon)\mu + \varepsilon\theta \mid \bar{b}) \geq \Pi_{\theta}((1-\varepsilon)\mu + \varepsilon\theta \mid \bar{b}),$$

where $B((1-\varepsilon)\mu + \varepsilon\theta \mid b) = \{\tilde{b} \in B((1-\varepsilon)\mu + \varepsilon\theta) : \tilde{b}(\theta_1, \theta_2) = b(\theta_1, \theta_2)$ for all $\theta_1, \theta_2 \in C(\mu)\}$.

We will refer to the distribution μ in the definition as the stable distribution.

We first show that if an outcome x is stable (with distribution μ), then all the incumbents receive the same fitness in each of their matches with other incumbents. Moreover, the average fitness that they obtain (which, of course, is equal to the fitness they obtain from each match), can be generated by a symmetric strategy profile in G. So, if an outcome is stable, then the average fitness that each type in the population gets must be equal to the payoff of a symmetric strategy profile in G.[13]

[12] Notice that one of the types in consideration (entrant) is new to the population, so we cannot claim that there is a "perceived way" of playing the game. This "convention" argument supports our assumption below that incumbents will continue to play the same equilibrium when they are matched against other incumbents.

[13] Note that this is vacuously true if we restrict our attention to monomorphic populations, i.e., those in which only one preference type is represented.

Proposition 1 *If an outcome x is stable (with μ and b), then there exists $\sigma \in \Delta$ such that for all $\theta, \theta' \in C(\mu)$,*

$$\Pi_\theta(\mu \mid b) = \pi(b_1(\theta, \theta'), b_2(\theta, \theta')) = \pi(\sigma, \sigma).$$

We call a strategy σ^* *efficient* if the payoff of the strategy profile (σ^*, σ^*) is at least as large as the payoff of any other symmetric strategy profile. This efficiency concept is meaningful in light of Proposition 1. All stable outcomes must give every type in the population the same payoff as some symmetric strategy profile. So, the stable outcomes are naturally ranked in terms of the payoffs that they generate, and $\pi(\sigma^*, \sigma^*)$ is the highest feasible payoff that any stable outcome can generate.

Definition 3 *$\sigma^* \in \Delta$ is **efficient** if $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma)$ for all $\sigma \in \Delta$.*

When σ^* is efficient, we will refer $\pi(\sigma^*, \sigma^*)$ as the efficient payoff. We now show that if (a^*, a^*) is a strict Nash equilibrium of G, where $a^* \in A$ is efficient, then it is stable as well. Consider a population consisting of types for which a^* is a strictly dominant strategy and any entrant type. If the entrants play anything but a^* against the incumbents, they will be driven out, since (a^*, a^*) is a strict Nash equilibrium. If they play a^* on the other hand, their expected fitness can never exceed that of the incumbents, since a^* is efficient.

Proposition 2 *If $a^* \in A$ is efficient and (a^*, a^*) is a strict Nash equilibrium of G, then (a^*, a^*) is stable.*

In Proposition 1 we showed that, for an outcome to be stable, each type in the population must receive the same fitness in each of its encounters with other types in the stable distribution, which implies that the average fitness that each type in the population gets must be equal to the payoff of a symmetric strategy profile in G. It follows from the definition of efficiency that if an outcome is stable, then the average fitness of the population cannot be larger than the efficient payoff. Our next result proves that it cannot be smaller either: An outcome is stable only if the average fitness of each type in the population is equal to the efficient payoff. The idea is simple, and best demonstrated for monomorphic populations, where its "secret handshake" flavor is clear.[14] Suppose the incumbents' fitness is less than

[14] See Robson (1990).

the efficient payoff. We can always find an entrant which would do better than the incumbents in the post-entry population. Consider, for example, a coordination game. The outcomes of the "bad equilibria" are not stable, because an entrant whose preferences coincide with the fitness function can invade by playing, as part of the post-entry equilibrium configuration, the bad action against the incumbents and the good one against itself. Now, consider a Prisoners' Dilemma game. The defection outcome is not stable. Any population where defection is played can be invaded by an entrant who has "coordination" type preferences, i.e., types for which defection (respectively, cooperation) is the unique best response to defection (respectively, cooperation). There is a post-entry equilibrium configuration in which the entrant and the incumbents both defect when they face each other, and the entrant cooperates when matched with itself. Our next result shows that these arguments can be generalized: Efficiency is a necessary condition for stability.

Proposition 3 *If an outcome x is stable (with μ and b), then for all $\theta \in C(\mu)$, $\Pi_\theta(\mu \mid b) = \pi(\sigma^*, \sigma^*)$, where σ^* is efficient.*

Combining Propositions 2 and 3 yields a unique prediction for a class of games which include the much studied coordination games:

Corollary 1 *Let G be such that $\pi(a_1, a_1) \geq \pi(\sigma, \sigma)$ for all $\sigma \in \Delta$ and $\pi(a_1, a_1) > \pi(a_i, a_1)$ for $i \neq 1$. The (unique) stable outcome is (a_1, a_1).*

We now restrict our attention to 2×2 games.

3.1 2×2 Games

In this subsection we exclusively study a class of games that attracted considerable attention in the literature: 2×2 games. First, we characterize the stable outcomes. In Proposition 3 we showed that efficiency was necessary for stability. It turns out that, in 2×2 games, efficiency of a pure strategy is sufficient for the corresponding outcome to be stable. Moreover, games with a mixed efficient strategy do not have stable outcomes with the exception of a nongeneric class of Hawk-Dove games. Hence, for generic games, the existence of a pure efficient strategy is both necessary and sufficient for the existence of a stable outcome. Finally, we characterize the stable distributions of preferences. The Prisoners' Dilemma case is particularly interesting.

All of the types that can be in any stable distribution belong to a certain equivalence class which has a "secret handshake" flavor: Against any opponent, they cooperate with positive probability in equilibrium only if their opponent is cooperating with probability one.

In order to simplify the exposition in establishing these results, we now introduce more notation and review a basic fact about 2×2 games.

FACT Consider any 2×2 (normal) game form with the strategy set $\{A, B\}$. In terms of equilibrium behavior, all possible payoff functions that a player may have, belong to one and only one of the following equivalence classes: $\mathcal{AA}$, $\mathcal{AB}_\alpha$, $\mathcal{BA}_\alpha$, $\mathcal{BB}$, and θ^o, where $\alpha \in [0, 1]$, and

$$\mathcal{AA} = \begin{array}{c|cc} & A & B \\ \hline A & 1 & 1 \\ B & 0 & 0 \end{array}, \quad \mathcal{AB}_\alpha = \begin{array}{c|cc} & A & B \\ \hline A & 1-\alpha & 0 \\ B & 0 & \alpha \end{array}, \quad \mathcal{BA}_\alpha = \begin{array}{c|cc} & A & B \\ \hline A & -1+\alpha & 0 \\ B & 0 & -\alpha \end{array},$$

$$\mathcal{BB} = \begin{array}{c|cc} & A & B \\ \hline A & 0 & 0 \\ B & 1 & 1 \end{array}, \quad \theta^o = \begin{array}{c|cc} & A & B \\ \hline A & 0 & 0 \\ B & 0 & 0 \end{array}.$$

Notice that X is a best-response to A and Y is a best-response to B for the payoff functions in $XY \in \{\mathcal{AA}, \mathcal{AB}_\alpha, \mathcal{BA}_\alpha, \mathcal{BB}\}$. All players with payoff functions within a given class will have the same set of equilibria in any game (game is defined by players and their payoff functions, in addition to the game form). A payoff function for which A strictly dominates B belongs to $\mathcal{AA}$, and any player with this kind of payoffs will play A in any equilibrium of any game. A player with a payoff function that belongs to $\mathcal{AB}_\alpha$, in any equilibrium of any game, will play A (respectively, B) if her opponent is playing A (respectively, B) and mix between A and B if her opponent plays A with probability α. For example, in the game that an $\mathcal{AB}_\alpha$ is matched with an $\mathcal{AA}$, the unique equilibrium is (A, A); when an $\mathcal{AB}_\alpha$ is matched with an $\mathcal{AB}_\beta$, where α,$\beta \in (0, 1)$, there are three equilibria: (A, A), (B, B) and a mixed strategy equilibrium in which $\mathcal{AB}_\alpha$ (respectively, $\mathcal{AB}_\beta$) plays A with probability β (respectively, α).

We next present a result which, when combined with Proposition 3, characterizes stable outcomes in 2×2 games. Let G be

	A	B
A	a, a	b, c
B	c, b	d, d

where $a \geq d$, without loss of generality. Suppose that A is efficient. We show that (A, A) is stable, and hence all the types in any stable distribution obtain a as fitness. If A is not efficient, i.e., if the efficient strategy is mixed, then there is no stable outcome unless G is a Hawk-Dove game ($c \geq a, b \geq d$) with $b = c$.

Proposition 4 *a) If A is efficient, then (A, A) is stable.*

b) If A is not efficient, then the outcome induced by (σ^*, σ^*) is stable iff $b = c \geq a$, where σ^* is efficient.

We now consider stable distributions, i.e., the preferences that are selected. We showed that in Prisoners' Dilemma and Hawk-Dove games, in which A is efficient, a monomorphic population consisting of $\mathcal{AB}_1$'s is a stable distribution. It turns out that it is the only one.[15] To illustrate this, consider a Prisoners' Dilemma game and restrict attention to monomorphic populations. It is clear why $\mathcal{AA}$ (cooperate (A) dominates defect (B)) cannot be stable. $\mathcal{BB}$ enters and in the unique equilibrium it defects while the incumbent is cooperating. But, why cannot the type with coordination game payoffs ($\mathcal{AB}_\alpha$) that cooperates be stable, since it defects when the opponent is defecting and cooperates when the opponent is cooperating? The only problem arises from the mixed strategy equilibrium against the entrant $\mathcal{AB}_\beta$, where $\beta > \alpha$. In that equilibrium $\mathcal{AB}_\alpha$ plays cooperation with probability β and $\mathcal{AB}_\beta$ plays cooperation with probability α, i.e., $\mathcal{AB}_\alpha$ is cooperating too much relative to $\mathcal{AB}_\beta$. The only type that is immune to this problem is $\mathcal{AB}_1$ (the limit in this heuristic selection process): in any equilibrium against any type, it cooperates with positive probability *only if* its opponent is cooperating with probability one. Hence, a monomorphic population consisting of $\mathcal{AB}_1$'s is the unique stable distribution. Notice that $\mathcal{AB}_1$ is not generic in the set of payoff functions. Moreover, the equilibrium chosen when it is matched against itself ((A, A)) is not perfect, since A is weakly dominated. However, notice that, as the discussion above illustrates, both the types and the equilibrium configurations chosen are endogenous to the model, they are the outcomes of the "selection process."

We also show below that in coordination games any type for which (A, A) is an equilibrium when matched against itself can be in a stable distribution.

[15] Note that $\mathcal{AB}_1$ is an equivalance class. All types who are indifferent between A and B (respectively, strictly prefer B) when the opponent plays A (respectively, B) belong to this class.

Also, in Hawk-Dove games where the efficient strategy is mixed, a monomorphic population consisting of $\mathcal{AB}_\alpha$ is the unique stable distribution, where α is the weight that the efficient strategy puts on A. This is interesting, since the fitness function of these games is given by $\mathcal{BA}_\alpha$. Why, then, cannot $\mathcal{BA}_\alpha$ be in any stable distribution? Because, a type for which "Hawk" is a dominant strategy ($\mathcal{AA}$) enters, and since $\mathcal{BA}_\alpha$ plays "Dove" in the unique equilibrium when $\mathcal{BA}_\alpha$ and $\mathcal{AA}$ are matched, $\mathcal{AA}$ obtains a higher average fitness than the incumbents.

Proposition 5 *For any generic 2×2 game G, μ is a stable distribution iff its support is a subset of $M(G)$, where $M(G)$ is defined below.*

a) For games in which A is efficient, i.e., $a > \pi(\sigma, \sigma)\ \forall \sigma \neq a$:

i) If $a > c$ and $a > b$, then $M(G) = \{\mathcal{AA}, \mathcal{AB}_\alpha, \mathcal{BA}_1, \theta^o\}$, $\alpha \in [0, 1]$.

ii) If $a > c$ and $b > a$, then $M(G) = \{\mathcal{AA}, \mathcal{AB}_\alpha\}$, where $\alpha \leq \frac{a-d}{b-d}$.

iii) If $c > a$, then $M(G) = \{\mathcal{AB}_1\}$.

b) If $\sigma^* \neq A$ is efficient and A is not and $b = c > a$, then $M(G) = \{\mathcal{AB}_{\alpha^*}\}$, where $\alpha^* = \sigma^*(A)$.

4 Unobservable Preferences

In this section we consider the case in which players do not observe their opponents' types. Suppose that the distribution of preferences is given by μ. The situation can be analyzed as a two-player Bayesian game, $\Gamma(\mu)$, where, for both players, the set of possible actions is A, the set of possible types is $C(\mu)$, the payoff function of type θ is given by θ, and for each type the probability distribution over the other player's types is given by the common prior distribution μ. Notice that we have symmetry, since the players cannot condition their behavior on their position in G. We will assume that, given the population distribution μ, the aggregate play in the population corresponds to a Bayesian-Nash Equilibrium of this game. That is, we will assume that each individual, upon being selected to play, will have correct beliefs about the distribution over her opponents' play and will choose a (possibly mixed) action that is a best-reply, according to her own preferences, to this belief. Let $B(\mu)$ denote the set of all Bayesian-Nash Equilibria when the distribution of preferences in the population is given by μ. Each $b \in B(\mu)$ determines

an action for each type θ in $C(\mu)$, $b_\theta \in \Delta$, such that

$$b_\theta \in \arg\max_{\sigma \in \Delta} \sum_{\theta'} \theta(\sigma, b_{\theta'})\mu(\theta').$$

The average fitness of type θ, given the population distribution μ and the equilibrium $b \in B(\mu)$, is, then

$$\Pi_\theta(b \mid \mu) = \sum_{\theta'} \pi(b_\theta, b_{\theta'})\mu(\theta').$$

Like in the previous section, a preference distribution μ is stable with respect to outcome x via b if x is the outcome induced by the equilibrium b of $\Gamma(\mu)$ and every type in the population obtains the same expected fitness. The stability definition is in the same spirit as well:

Definition 4 *An outcome x is **stable** (with μ and b) if there exists a distribution μ and an equilibrium $b \in B(\mu)$ such that μ is stable with respect to x via b and $\forall\theta$ $\exists\varepsilon' > 0$ such that $\forall\varepsilon \in (0, \varepsilon')$, $\forall\theta' \in C(\mu)$, we have:*

i) $B((1-\varepsilon)\mu + \varepsilon\theta \mid b) \neq \emptyset$, *and*

ii) $\Pi_{\theta'}((1-\varepsilon)\mu+\varepsilon\theta \mid \bar{b}) \geq \Pi_\theta((1-\varepsilon)\mu+\varepsilon\theta \mid \bar{b})$, *for all* $\bar{b} \in B((1-\varepsilon)\mu+\varepsilon\theta \mid b)$, *where*

$B((1-\varepsilon)\mu+\varepsilon\theta \mid b) = \{\tilde{b} \in B((1-\varepsilon)\mu+\varepsilon\theta) : \tilde{b}_{\theta'} = b_{\theta'}$ *for all* $\theta' \in C(\mu)\}$.

When a mutant enters, we require each incumbent to do at least as well as the mutant in any equilibrium in the post-entry population in which the incumbents' actions are left unchanged, which we call *focal* equilibria. The arguments for restricting the set of equilibria this way and checking for *all* equilibria in this restricted set are the same as the ones made in the observable types case.[16] The only difference between the two definitions in the observable and unobservable types cases, then, is the requirement, in the latter case, of non-emptiness of the set of focal equilibria in the post-entry population. This is not a real difference, since non-emptiness is trivially satisfied in the observable-types case. We require the existence of a focal equilibrium

[16] One may argue that the arguments presented can only justify restricting the post-entry equilibrium to those where incumbents' actions are "close enough" to their actions in the before-entry equilibrium instead of being exactly equal. Our stability definition can easily be changed to deal with this concern without qualitatively affecting the result below.

in the post-entry population, because we want the resulting outcome to be "close" to the outcome that we are claiming to be stable. If an entrant causes the outcome induced by the equilibrium play of the agents to move significantly, even for arbitrary small population share of the entrant, then that outcome cannot be considered stable. Consider the following Prisoner's Dilemma game:

	A	B
A	2	0
B	3	1

Now, consider a monomorphic population consisting of $\mathcal{AB}_1$'s (indifferent between A and B if the opponent plays A, strictly prefers B if the opponent plays B) playing A in the chosen equilibrium. No entrant can invade, since in any equilibrium in any post-entry population (not only within the set of focal equilibria), the incumbents play A with positive probability only if the entrant plays A with probability one. In a sense, the incumbents are stable. However, (A, A), the cooperation outcome is not stable. Consider the entrant for which B strictly dominates A; in the unique equilibrium in the post-entry population, even for arbitrarily small shares for the entrant, everyone plays B. So, the introduction of even a very small share of the entrant causes a large shift in the outcome, from "all cooperate" to "all defect". Hence, even though no entrant can earn a strictly higher fitness than the incumbents in any equilibrium, we cannot claim that (A, A) is stable.

Given the symmetric nature of the interaction in the population, any strategy profile in the Bayesian Game, $\Gamma(\mu)$, as well as any equilibrium $b \in B(\mu)$, induces a symmetric outcome in G:

FACT Any outcome induced by any strategy profile in $\Gamma(\mu)$ is identical to the outcome induced by (σ, σ), for some $\sigma \in \Delta$.

We now show that, when preferences are unobservable, stability in our model of evolution of preferences has the same implications with the concept of neutrally stable strategy (NSS) in the game G.[17] An outcome induced by a symmetric strategy profile, which are the only possible ones that can result from the interaction we analyze, is stable *iff* that strategy is an NSS. Instead of emphasizing the equivalence with NSS, it may be more informative to look at slightly different necessary and sufficient conditions separately.

[17] Note that if σ is an NSS, then (σ, σ) is a Nash equilibrium, and conversely, if (σ, σ) is a strict Nash equilibrium, then σ is an NSS.

Firstly, "stable only if Nash" folk theorem is revived: Nash behavior is a necessary condition for stability. Consider a monomorphic population that is not playing a Nash equilibrium of the game G. This implies that an entrant who plays a (pure) action that is a strictly better response (in terms of fitness) to incumbents' play would be more successful than the incumbents. Conversely, a strict Nash equilibrium outcome will be stable, say, with incumbents who have preferences for which the Nash action is strictly dominant. In this case any entrant who does not play the Nash action will be driven out.

The result and the related discussion illustrate how close the relation between models of evolution of preferences and the standard evolutionary game theory are when preferences are completely unobservable. The basic intuition is straightforward: If nobody can observe your preferences, and hence condition their behavior on that, there is no advantage in having preferences that are different from the one given by the fitness function.

Proposition 6 *The outcome induced by (σ, σ) is stable iff σ is an NSS.*

5 Imperfectly Observable Preferences

In this section we consider an intermediate case where preferences are imperfectly observable. In particular, we assume that each player observes the preferences of the opponent she is matched against with probability $p \in (0, 1)$ (with the complementary probability, $1 - p$, she does not observe anything) independent of what her opponent observes. Suppose that the distribution of preferences is given by μ. The situation can be analyzed as a two-player Bayesian game, $\Gamma_p(\mu)$. Again we have symmetry, since the players cannot condition their behavior on their position in G. We will assume that, given the population distribution μ, the aggregate play in the population corresponds to a Bayesian-Nash Equilibrium of this game. That is, we will assume that each individual, upon being selected to play, will have correct beliefs about the distribution over her opponents' play and will choose an (possibly mixed) action that is a best-reply, according to her own preferences, to this belief. Let $B_p(\mu)$ denote the set of all Bayesian-Nash Equilibria when the distribution of preferences in the population is given by μ. Each equilibrium determines, for each type in the population, a set of actions that she would take when she observes any other type in the population, and an action that she would take when she does not observe anything. Formally,

each $b \in B_p(\mu)$ determines an action for each type θ in $C(\mu)$ in the case she observes $\theta^{'} \in C(\mu)$, $b_\theta(\theta^{'}) \in \Delta$, and an action in the case she does not observe anything, b_θ, such that

$$b_\theta(\theta^{'}) \in \arg\max_{\sigma \in \Delta} p\theta(\sigma, b_{\theta'}(\theta)) + (1-p)\theta(\sigma, b_{\theta'}),$$

and

$$b_\theta \in \arg\max_{\sigma \in \Delta} \sum_{\theta'} [p\theta(\sigma, b_{\theta'}(\theta)) + (1-p)\theta(\sigma, b_{\theta'})]\mu(\theta^{'}).$$

The average fitness of type θ, given the population distribution μ and the equilibrium $b \in B_p(\mu)$, is, then

$$\Pi_\theta(\mu \mid b) = \sum_{\theta'} \{p[p\pi(b_\theta(\theta^{'}), b_{\theta'}(\theta)) + (1-p)\pi(b_\theta(\theta^{'}), b_{\theta'})] + (1-p)[p\pi(b_\theta, b_{\theta'}(\theta)) + (1-p)\pi(b_\theta, b_{\theta'})]\}\mu(\theta^{'}).$$

As in the previous sections, a preference distribution μ is stable with respect to outcome x via b if x is the outcome induced by the equilibrium b of $\Gamma_p(\mu)$ and every type in the population obtains the same expected fitness. The stability definition is in the same spirit as well:

Definition 5 *An outcome x is **stable** (with μ and b) for $p \in (0,1)$, if there exists a distribution μ and an equilibrium $b \in B_p(\mu)$ such that μ is stable with respect to x via b and $\forall \theta\ \exists \varepsilon^{'} > 0$ such that $\forall \varepsilon \in (0, \varepsilon^{'})$, $\forall \theta^{'} \in C(\mu)$, we have:*

i) $B_p((1-\varepsilon)\mu + \varepsilon\theta \mid b) \neq \emptyset$, and

ii) $\Pi_{\theta'}((1-\varepsilon)\mu + \varepsilon\theta \mid \overline{b}) \geq \Pi_\theta((1-\varepsilon)\mu + \varepsilon\theta \mid \overline{b})$, for all $\overline{b} \in B_p((1-\varepsilon)\mu + \varepsilon\theta \mid b)$, where

$B_p((1-\varepsilon)\mu + \varepsilon\theta \mid b) = \{\tilde{b} \in B_p((1-\varepsilon)\mu + \varepsilon\theta) : \tilde{b}_{\theta'} = b_{\theta'}$ and $\tilde{b}_{\theta'}(\theta^{''}) = b_{\theta'}(\theta^{''})$ for all $\theta^{'}, \theta^{''} \in C(\mu)\}$.

The set of focal equilibria considered is a natural extension of the ones considered in previous sections. Each incumbent is taking the same action when she observes any other incumbent and when she does not observe anything in all focal equilibria. There is no restriction on the entrant's actions or on the action taken by incumbents when they observe the entrant's preferences. It is easy to see that when $p = 1$ and $p = 0$ the definition of stability reduces to the corresponding definitions in the observable and unobservable preferences cases, respectively.

We first show that (strict) equilibrium combined with efficiency implies stability for *any* probability of observability.

Proposition 7 *If $a^* \in A$ is efficient and (a^*, a^*) is a strict Nash equilibrium of G, then (a^*, a^*) is stable for all $p \in (0, 1)$.*

In studying imperfectly observable preferences, we are particularly interested in whether the results about stability in the observable (respectively, unobservable) preferences case continue to hold for high (respectively, low) values of probability of observability. In other words, are the stability results of previous sections "continuous" in p?

In Proposition 6, we showed that if σ is an NSS, then the outcome induced by (σ, σ) is stable when preferences are unobservable. The following example demonstrates that even a weaker version of this is not true in the case of imperfectly observable preferences, even for arbitrarily small values of p: (B, B) is a strict Nash equilibrium, but it is not stable for any $p > 0$. Notice, as well, that (B, B) is also the risk-dominant equilibrium.[18]

Example 1 *Consider the following game:*

	A	B
A	6	0
B	5	2

We will show that even though (B, B) is a strict Nash equilibrium, it is not stable for any $p > 0$. Suppose that (B, B) is stable with distribution μ. Notice that every incumbent must have preferences for which B is a best-response to itself. Consider the coordination type, $\mathcal{AB}_\alpha$, where $0 < \alpha \leq p$, as the entrant. There is a post-entry focal equilibrium in which entrants play A if they observe each other, and play B otherwise, and the incumbents continue to play B regardless of what they observe. To see this, note that incumbents' opponents are playing B for sure, so it is a best-response for them to continue to play B. When an entrant observes an incumbent, she knows for sure that her opponent is playing B. When she does not observe anything, she is very likely to be facing an incumbent (who plays B) as an opponent. In either case it is optimal for the entrant to play B. When an entrant observes another entrant who plays A with probability p A must be her best-response, and this happens as long as the entrant's type assigns a

[18] In symmetric 2×2 games, a strategy is *risk-dominant* (Harsanyi and Selten, 1988) if it is a better response to a mixed strategy where both strategies are played with probability 1/2.

high enough payoff to (A, A), i.e., $\alpha \leq p$ in our normalization. For this post-entry equilibrium we specified, incumbents' and entrants' fitnesses are given by

$$\Pi_\theta(. \mid .) = 2, \ \forall \theta \in C(\mu),$$

and

$$\Pi_{\mathcal{AB}_\alpha}(. \mid .) = (1-\varepsilon)2 + \varepsilon[6p^2 + 5p(1-p) + 2(1-p)^2].$$

Since $\Pi_{\mathcal{AB}_\alpha}(. \mid .) > 2$ for $p > 0$, and hence (B, B) is not stable for any $p > 0$.

We next show that a version of the "stable only if efficient" result of the observable preferences case holds for high enough probability of observability: The outcome of a symmetric pure-strategy profile is not stable for p high enough, if that strategy is not efficient.

Proposition 8 *If there exists a $\sigma \in \Delta$ such that $\pi(a, a) < \pi(\sigma, \sigma)$, then there exists a $\overline{p} \in (0, 1)$ such that (a, a) is not stable for $p \in (\overline{p}, 1)$.*

Combining this result with the previous example, we provide some support for efficiency in coordination games:

Corollary 2 *Consider (strict) coordination games. The outcome of the risk-dominant equilibrium is not stable for large enough p, unless the equilibrium is also payoff-dominant. There exist games in which the outcome of the risk-dominant equilibrium is not stable for any $p > 0$. In contrast, the outcome of the payoff-dominant equilibrium is stable for all $p > 0$.*

The selection between the risk-dominant and the payoff-dominant equilibrium in coordination games have been studied extensively in evolutionary models.[19] The risk-dominant equilibrium is selected in the models of Ellison (1993), Kandori, Mailath and Rob (1993), and Young (1993); the payoff-dominant equilibrium is favored in Ely (2002) and Robson and Vega Redondo (1996); in Binmore and Samuelson (1997) either can be selected. Evolutionary analysis of cheap-talk games provide extensive support for efficiency. (See, for example, Bhaskar (1998), Kim and Sobel (1995), and Matsui (1991).)

To conclude, we show that a version of the "stable only if Nash" result of the unobservable preferences case holds for low enough probability of observability: If a pure-strategy profile is not a Nash equilibrium, then it is not stable for p low enough.

[19] Among the nonevolutionary models, Harsanyi and Selten (1988) selects the payoff-dominant, whereas Carlsson and van Damme (1993) selects the risk-dominant equilibrium.

Proposition 9 *If (a,a) is not a Nash equilibrium of G, then there exists a $\overline{p} \in (0,1)$ such that (a,a) is not stable for $p \in (0,\overline{p})$.*

6 Appendix

Proof of Proposition 1. First we will show that if x is stable (with μ and b), then all the types in μ obtain the same fitness, in b, when they are matched with any given type. Then we show that any given type gets the same fitness when it is matched with any other type, hence obtaining the average fitness in each and every match, proving the first equality. This, in particular, means that any type will receive the average fitness when it is matched against itself. Since each type has to play a symmetric equilibrium when matched against itself, the average fitness must be equal to the payoff of a symmetric strategy profile (σ,σ) in G.

Claim i) :

$$\pi(b_1(\theta^{'},\theta),b_2(\theta^{'},\theta)) = \pi(b_1(\theta^{''},\theta),b_2(\theta^{''},\theta))\forall\theta,\theta^{'},\theta^{''} \in C(\mu).$$

Let θ^o be the type who is indifferent between all the actions against any action of the opponent. Suppose that x is stable (with μ and b). Let $m(\theta) \in \arg\max_{\theta^{'}}\pi(b_1(\theta^{'},\theta),b_2(\theta^{'},\theta))$, i.e., $m(\theta)$ is the incumbent which gets the highest (equilibrium) fitness. Let θ^o be the entrant, and consider the equilibrium configuration $\overline{b}$, where $(\overline{b}_1(\theta^o,\theta),\overline{b}_2(\theta^o,\theta)) = (b_1(m(\theta),\theta),b_2(m(\theta),\theta))$ for all $\theta \in C(\mu)$, and $\overline{b}_1(\theta^o,\theta^o) = \overline{b}_2(\theta^o,\theta^o) = \arg\max_{\sigma\in\Delta}\pi(\sigma,\sigma)$. This can be done, because any equilibrium in the match between θ and $\theta^{'}$ is also an equilibrium between θ and θ^o, since θ^o's best-response set includes any other type's best-response set. Now, suppose that the claim above is not true, then there exists a $\theta \in C(\mu)$ such that

$$\Pi_{\theta^o}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) > \Pi_{\theta}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b})$$

for all $\varepsilon > 0$, showing that x is not stable (with μ and b), a contradiction.

Claim ii):

$$\pi(b_1(\theta,\theta^{'}),b_2(\theta,\theta^{'})) = \Pi_{\theta}(\mu \mid b)\forall\theta,\theta^{'} \in C(\mu).$$

Suppose that x is stable (with μ and b), and the statement above is not true. There must exist θ, $\theta^* \in C(\mu)$ such that

$$\pi(b_1(\theta,\theta^*),b_2(\theta,\theta^*)) < \Pi_{\theta}(\mu \mid b),$$

since the average fitness of θ is $\Pi_\theta(\mu \mid b)$. Now, the claim proven above implies that

$$\pi(b_1(\theta^*,\theta^*), b_2(\theta^*,\theta^*)) = \pi(b_1(\theta,\theta^*), b_2(\theta,\theta^*)) < \Pi_\theta(\mu \mid b).$$

Let θ^o be the entrant, and consider the equilibrium configuration $\overline{b}$, where $(\overline{b}_1(\theta^o,\theta'), \overline{b}_2(\theta^o,\theta')) = (b_1(\theta^*,\theta'), b_2(\theta^*,\theta'))$ for all $\theta' \in C(\mu)$, and $(\overline{b}_1(\theta^o,\theta^o), \overline{b}_2(\theta^o,\theta^o)) = (b_1(\widetilde{\theta},\widetilde{\theta}), b_2(\widetilde{\theta},\widetilde{\theta}))$, where $\widetilde{\theta}$ is chosen such that

$$\pi(b_1(\theta,\widetilde{\theta}), b_2(\theta,\widetilde{\theta})) > \Pi_\theta(\mu \mid b).$$

(Such a $\widetilde{\theta}$ must exist, since the average fitness of θ is $\Pi_\theta(\mu \mid b)$.) The average fitnesses of θ^* and θ^o, when θ^o's population share is ε, respectively, are:

$$\Pi_{\theta^*}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) = (1-\varepsilon)\Pi_\theta(\mu \mid b) + \varepsilon\pi(b_1(\theta^*,\theta^*), b_2(\theta^*,\theta^*)),$$

$$\Pi_{\theta^o}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) = (1-\varepsilon)\Pi_\theta(\mu \mid b) + \varepsilon\pi(\overline{b}_1(\theta^o,\theta^o), \overline{b}_2(\theta^o,\theta^o)),$$

since θ^o is imitating θ^*'s behavior against all incumbents, including itself. Now, we have,

$$\begin{aligned}\pi(\overline{b}_1(\theta^o,\theta^o), \overline{b}_2(\theta^o,\theta^o)) &= \pi(b_1(\widetilde{\theta},\widetilde{\theta}), b_2(\widetilde{\theta},\widetilde{\theta})) \\ &= \pi(b_1(\theta,\widetilde{\theta}), b_2(\theta,\widetilde{\theta})) > \Pi_\theta(\mu \mid b) > \pi(b_1(\theta^*,\theta^*), b_2(\theta^*,\theta^*))\end{aligned}$$

where the second inequality follows from *Claim i*) above. So,

$$\Pi_{\theta^o}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) > \Pi_{\theta^*}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b})$$

for all $\varepsilon > 0$, and hence x is not stable (with μ and b), a contradiction. ■

Proof of Proposition 2. Consider a monomorphic population μ, consisting of θ, for which a^* strictly dominates any other strategy. Clearly, μ is stable with respect to (a^*, a^*) via the unique equilibrium configuration b in which everyone plays a^*. Let θ' be any entrant and $\overline{b}$ be any equilibrium configuration in the post-entry population. θ will play a^* regardless of the opponent's type; suppose, in $\overline{b}$, θ' plays σ and σ' when matched against θ and itself, respectively. We have

$$\Pi_\theta((1-\varepsilon)\theta + \varepsilon\theta' \mid \overline{b}) = (1-\varepsilon)\pi(a^*,a^*) + \varepsilon\pi(a^*,\sigma),$$

and

$$\Pi_{\theta'}((1-\varepsilon)\theta + \varepsilon\theta' \mid \overline{b}) = (1-\varepsilon)\pi(\sigma,a^*) + \varepsilon\pi(\sigma',\sigma').$$

Now, if $\sigma \neq a^*$, then $\pi(a^*, a^*) > \pi(\sigma, a^*)$, since (a^*, a^*) is a strict Nash equilibrium. If $\sigma = a^*$, the hypothesis of the proposition implies that $\pi(a^*, \sigma) = \pi(a^*, a^*) \geq \pi(\sigma^{'}, \sigma^{'})$ for any $\sigma^{'} \in \Delta$. In either case we can find an $\varepsilon^{'} > 0$ such that $\Pi_\theta(.) \geq \Pi_{\theta'}(.)$ for all $\varepsilon \in (0, \varepsilon^{'})$, showing that (a^*, a^*) is stable. ■

Proof of Proposition 3. Suppose that x is stable (with μ and b). Proposition 1 and the efficiency of σ^* imply that

$$\Pi_\theta(\mu \mid b) \leq \pi(\sigma^*, \sigma^*).$$

Now, suppose that

$$\Pi_\theta(\mu \mid b) < \pi(\sigma^*, \sigma^*).$$

Let θ^o be the entrant, and choose the post-entry equilibrium configuration $\overline{b}$ such that:

i) The equilibrium when θ^o and any incumbent $\theta^{'} \in C(\mu)$ are matched is the same as the equilibrium played when $\theta^{'}$ and $\theta \in C(\mu)$ are matched, i.e., $(\overline{b}_1(\theta, \theta^o), \overline{b}_2(\theta, \theta^o)) = (b_1(\theta, \theta^{'}), b_2(\theta, \theta^{'}))$ for all $\theta^{'} \in C(\mu)$.

ii) When θ^o is matched against itself (σ^*, σ^*) is played. Then, we have

$$\Pi_\theta((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) = (1-\varepsilon)\Pi_\theta(\mu \mid b) + \varepsilon\pi(\overline{b}_1(\theta, \theta^o), \overline{b}_2(\theta, \theta^o)),$$

and

$$\Pi_{\theta^o}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) = (1-\varepsilon)\Pi_\theta(\mu \mid b) + \varepsilon\pi(\overline{b}_1(\theta^o, \theta^o), \overline{b}_2(\theta^o, \theta^o)).$$

Since $(\overline{b}_1(\theta, \theta^o), \overline{b}_2(\theta, \theta^o)) = (\overline{b}_2(\theta, \theta^o), \overline{b}_1(\theta, \theta^o)) = (b_1(\theta, \theta), b_2(\theta, \theta))$, it follows from Proposition 1 that

$$\pi(\overline{b}_1(\theta, \theta^o), \overline{b}_2(\theta, \theta^o)) = \Pi_\theta(\mu \mid b).$$

Moreover,

$$\pi(\overline{b}_1(\theta^o, \theta^o), \overline{b}_2(\theta^o, \theta^o)) = \pi(\sigma^*, \sigma^*) > \Pi_\theta(\mu \mid b).$$

Hence, for all $\varepsilon > 0$,

$$\Pi_{\theta^o}((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}) > \Pi_\theta((1-\varepsilon)\mu + \varepsilon\theta^o \mid \overline{b}),$$

showing that x is not stable, a contradiction. ■

Proof of Proposition 4. *a*) (A is efficient) We will consider two cases:

i) $a > c$: In this case, which consists of coordination games and games in which the efficient (pure) strategy (A) strictly dominates the other strategy (B), Proposition 2 implies that (A, A) is stable.

ii) $a \leq c$: In this case, which consists of Prisoners' Dilemma and Hawk-Dove games, (A, A) is stable with a monomorphic population of $\mathcal{AB}_1$. $\mathcal{AB}_1$ is the type for which both A and B are best responses to A, and B is the unique best response to B. Consider a monomorphic population consisting of $\mathcal{AB}_1$, and the equilibrium configuration in which they play (A, A). Let θ be any entrant and $\overline{b}$ be any equilibrium configuration in the post-entry population. Suppose that in $\overline{b}$, the equilibrium between $\mathcal{AB}_1$ and θ is (σ_1, σ_2), and the equilibrium when θ is matched with itself is (σ_3, σ_3). The expected fitness to the incumbent and the entrant are, respectively,

$$\Pi_{\mathcal{AB}_1}((1-\varepsilon)\mathcal{AB}_1 + \varepsilon\theta \mid \overline{b}) = (1-\varepsilon)a + \varepsilon\pi(\sigma_1, \sigma_2),$$

and

$$\Pi_{\theta}((1-\varepsilon)\mathcal{AB}_1 + \varepsilon\theta \mid \overline{b}) = (1-\varepsilon)\pi(\sigma_2, \sigma_1) + \varepsilon\pi(\sigma_3, \sigma_3).$$

In any equilibrium against any type of the opponent, $\mathcal{AB}_1$ plays A with positive probability only if the opponent plays A with probability one, i.e., $\sigma_1(A) > 0 \Rightarrow \sigma_2 = A$. Suppose $\sigma_2 = A$, in which case $\mathcal{AB}_1$ is indifferent between A and B, and consider all possible equilibria. The expected fitness to the incumbent and the entrant are, respectively,

$$\Pi_{\mathcal{AB}_1}(.) = (1-\varepsilon)a + \varepsilon[qa + (1-q)c],$$

and

$$\Pi_{\theta}(.) = (1-\varepsilon)[qa + (1-q)b] + \varepsilon\pi(\sigma_3, \sigma_3),$$

where $q \in [0, 1]$. Since $c \geq a$, efficiency of A implies that $a \geq b$. So,

$$a \geq qa + (1-q)b.$$

Also, efficiency of A implies that

$$qa + (1-q)c \geq a \geq \pi(\sigma_3, \sigma_3)$$

for any $\sigma_3 \in \Delta$. Hence, $\Pi_{\mathcal{AB}_1}(.) \geq \Pi_{\theta}(.)$, irrespective of ε, proving that (A, A) is stable.

Now, suppose that $\sigma_2 \neq A$, which implies that $\sigma_1 = B$. Considering all possible equilibrium configurations, the expected fitness to the incumbent and the entrant are, respectively,

$$\Pi_{\mathcal{AB}_1}(.) = (1-\varepsilon)a + \varepsilon[qc + (1-q)d],$$

and

$$\Pi_{\theta}(.) = (1-\varepsilon)[qb + (1-q)d] + \varepsilon\pi(\sigma_3, \sigma_3),$$

where $q \in [0,1]$. We have $a \geq qb + (1-q)d$, since $a \geq b$ and $a \geq d$. If the inequality is strict, then we are done: we can find $\varepsilon^{'} > 0$ such that $\Pi_{\mathcal{AB}_1}(.) \geq \Pi_{\theta}(.)$ for all $\varepsilon \in (0, \varepsilon^{'})$. So, suppose that $a = qb + (1-q)d$. Now, we have

$$qc + (1-q)d \geq qb + (1-q)d = a \geq \pi(\sigma_3, \sigma_3)$$

for any $\sigma_3 \in \Delta$. Hence, $\Pi_{\mathcal{AB}_1}(.) \geq \Pi_{\theta}(.)$, irrespective of ε, proving that (A, A) is stable.

b) (A is not efficient) Let $\sigma^* = \arg\max_{\sigma \in \Delta} \pi(\sigma, \sigma)$, i.e.,

$$\alpha^* = \sigma^*(A) = \frac{b + c - 2d}{2(b + c - a - d)} \in (0,1),$$

and $\sigma^*(B) = 1 - \alpha^*$. Notice that $\pi(\sigma^*, \sigma^*) = d + \frac{(b+c-2d)^2}{4(b+c-a-d)}$. Since $\pi(\sigma^*, \sigma^*) > a$ implies that $b + c > 2a$, we know that σ^* is unique. Thus, Propositions 1 and 3 imply that, if an outcome is stable, then (σ^*, σ^*) must be played in each matching within the stable distribution. So any stable distribution must be a distribution on $\{\mathcal{AB}_{\alpha^*}, \mathcal{BA}_{\alpha^*}, \theta^o\}$. We now consider four classes of 2×2 games in turn:

i) $a \geq c$ and $d \geq b$ (Coordination games): A is always an efficient strategy for this class of games.

ii) $a \geq c$ and $b \geq d$: If $c \geq b$, then A is efficient. So, let $b > c$. Suppose that the outcome induced by (σ^*, σ^*) is stable. We can show that $\mathcal{AB}_0$ can enter and obtain strictly higher expected fitness against incumbents than the incumbents obtain against themselves in the equilibrium configuration $(\overline{b})$ in which $\mathcal{AB}_0$ mixes between A and B (playing A with probability α^*) and the incumbents play B when they are matched. Let μ be the stable distribution. We have, for any $\theta \in \mu$,

$$\Pi_{\theta}((1-\varepsilon)\mu + \varepsilon\mathcal{AB}_0 \mid \overline{b}) = (1-\varepsilon)\pi(\sigma^*, \sigma^*) + \varepsilon[\alpha^* c + (1-\alpha^*)d],$$

and

$$\Pi_{\mathcal{AB}_0}((1-\varepsilon)\mu + \varepsilon\mathcal{AB}_0 \mid \bar{b}) = (1-\varepsilon)[\alpha^* b + (1-\alpha^*)d] + \varepsilon\pi(\sigma,\sigma).$$

It is easy to show that, for $b > c$,

$$\alpha^* b + (1-\alpha^*)d > \pi(\sigma^*,\sigma^*).$$

Hence, we do not have stability.

iii) $c \geq a$ and $d \geq b$ (Prisoners' Dilemma): We have $c \geq b$. If $c = b$, then A is efficient. So, let $c > b$. Suppose that μ is a stable distribution. Let $\mathcal{AB}_1$ be the entrant, and consider the equilibrium configuration in which when it is matched with incumbents the mixed strategy equilibrium is played ($\mathcal{AB}_1$ playing A with probability α^*, and incumbents playing A). The entrant's expected fitness from its encounters with the incumbents are

$$\alpha^* a + (1-\alpha^*)c > \pi(\sigma^*,\sigma^*),$$

which shows that μ is not a stable distribution, a contradiction.

iv) $c \geq a$ and $b \geq d$ (Hawk-Dove): If $b > c$ (respectively, $c > b$), then the exact arguments in case *ii*) (respectively, *iii*)) apply, so there is no stable outcome. If $b = c$, then the monomorphic population of $\mathcal{AB}_{\alpha^*}$ is stable. Let θ be any entrant and $\bar{b}$ be any equilibrium configuration in the post-entry population. Suppose that in $\bar{b}$, the equilibrium between $\mathcal{AB}_{\alpha^*}$ and θ is (σ_1,σ_2), and the equilibrium when θ is matched with itself is (σ_3,σ_3). The expected fitness to the incumbent and the entrant are, respectively,

$$\Pi_{\mathcal{AB}_{\alpha^*}}((1-\varepsilon)\mathcal{AB}_{\alpha^*} + \varepsilon\theta \mid \bar{b}) = (1-\varepsilon)\pi(\sigma^*,\sigma^*) + \varepsilon\pi(\sigma_1,\sigma_2),$$

and

$$\Pi_{\theta}((1-\varepsilon)\mathcal{AB}_{\alpha^*} + \varepsilon\theta \mid \bar{b}) = (1-\varepsilon)\pi(\sigma_2,\sigma_1) + \varepsilon\pi(\sigma_3,\sigma_3).$$

If $\sigma_2 = A$ (respectively, B), then $\sigma_1 = A$ (respectively, B); in either case $\pi(\sigma^*,\sigma^*) > \pi(\sigma_2,\sigma_1)$. In the mixed strategy equilibrium, the entrant must play A with probability α^*, in which case, straightforward calculations show that, $\max_{\sigma_1\in\Delta} \pi(\sigma_2,\sigma_1) = \pi(\sigma^*,\sigma^*)$. Therefore, $\mathcal{AB}_{\alpha^*}$ is stable. ■

Proof of Proposition 5. *a*) Propositions 1 and 3 imply that, if an outcome is stable, then (A,A) must be played in each matching within the stable distribution. So $M(G)$ must be a subset of $\{\mathcal{AA},\ \mathcal{AB}_\alpha,\ \mathcal{BA}_1,\ \theta^o\}$, $\alpha \in [0,1]$. Moreover if an entrant receives a against any incumbent, then,

the genericity of G implies that the incumbent must receive a as well in that equilibrium. So for stability we only need to check whether there is an entrant which can obtain fitness greater than a against any incumbent in any equilibrium.

i) In this case fitness from (A, A) (which is a) is greater than fitness from any other strategy profile. Therefore any type for which (A, A) is an equilibrium can be in a stable distribution.

ii) If a stable distribution contains $\mathcal{BA}_1$ or θ^o, $\mathcal{AA}$ can enter and obtain $b > a$ in the matchings against $\mathcal{BA}_1$ and θ^o, thereby getting a higher average fitness than $\mathcal{BA}_1$ and θ^o in any post-entry equilibrium configuration. Now, suppose that a stable distribution contains $\mathcal{AB}_\alpha$, where $\alpha > \frac{a-d}{b-d}$. $\mathcal{AB}_0$ can enter, and in the equilibrium between $\mathcal{AB}_0$ and $\mathcal{AB}_\alpha$ in which $\mathcal{AB}_0$ plays A with probability α and $\mathcal{AB}_\alpha$ plays A, obtain $\alpha b + (1-\alpha)d > a$. $\mathcal{AA}$ is stable because any entrant in any equilibrium receives a convex combination of a and c, which is less than or equal to a. $\mathcal{AB}_\alpha$, where $\alpha \leq \frac{a-d}{b-d}$, is stable since any entrant in any pure strategy equilibrium gets a or d, and the highest that it gets in a mixed strategy equilibrium is either $\alpha b + (1-\alpha)d$ or $\alpha a + (1-\alpha)c$, which are both less than or equal to a.

iii) We showed in the proof of Proposition 4 that $\mathcal{AB}_1$ is stable. Now suppose that a stable distribution contains $\mathcal{AA}$, $\mathcal{BA}_1$ or θ^o. $\mathcal{AB}_1$ can enter. There is an equilibrium between $\mathcal{AB}_1$ and $\mathcal{AA}$ (also $\mathcal{BA}_1$ and θ^o) in which $\mathcal{AB}_1$ plays B and $\mathcal{AA}$ plays A, which gives $\mathcal{AB}_1$ a fitness of $c > a$. Suppose that a stable distribution contains $\mathcal{AB}_\alpha$, where $\alpha \in [0, 1)$. There is an equilibrium in which $\mathcal{AB}_1$ mixes between A and B, and $\mathcal{AA}$ plays A, which gives $\mathcal{AB}_1$ a fitness of $\alpha a + (1-\alpha)c > a$. Hence $\mathcal{AB}_1$ is the only stable distribution.

b) Propositions 1 and 3 imply that, if an outcome is stable, then (σ^*, σ^*) must be played in each matching within the stable distribution. So $M(G)$ must be a subset of $\{\mathcal{AB}_{\alpha^*}, \mathcal{BA}_{\alpha^*}, \theta^o\}$. We showed in the proof of Proposition 4 that $\mathcal{AB}_{\alpha^*}$ is stable. Suppose that a stable distribution contains $\mathcal{BA}_{\alpha^*}$ or θ^o. $\mathcal{AA}$ enters and obtains $b > a$ in the matchings with $\mathcal{BA}_{\alpha^*}$ and θ^o. Hence $\mathcal{AB}_{\alpha^*}$ is the only stable distribution. ■

Proof of Proposition 6. *i*) (only if) The proof is by contradiction. If $\sigma \in \Delta$ is not an NSS, then there exists $\sigma' \in \Delta$ such that

$$\pi(\sigma', (1-\varepsilon)\sigma + \varepsilon\sigma') > \pi(\sigma, (1-\varepsilon)\sigma + \varepsilon\sigma'),$$

for arbitrarily small ε. Suppose that (σ, σ) is stable with some $\mu \in \mathcal{P}(\Theta)$ and $b \in B(\mu)$, but σ is not an NSS. Consider an entrant θ such that σ' is

a best response to $(1-\varepsilon)\sigma+\varepsilon\sigma^{'}$ for all $\varepsilon>0$. (Such a θ always exists, for example θ^{o}, the type which is indifferent between all the outcomes.) If the set of focal post-entry equilibria, $B((1-\varepsilon)\mu+\varepsilon\theta \mid b)$, is empty, then (σ,σ) is not stable, a contradiction. So, suppose that the set of focal post-entry equilibria is non-empty. Since the aggregate play by incumbents are given by σ in all focal equilibria, there exists $\overline{b}\in B((1-\varepsilon)\mu+\varepsilon\theta \mid b)$, for which we have

$$\Pi_{\theta}((1-\varepsilon)\mu+\varepsilon\theta \mid \overline{b})=\pi(\sigma^{'},(1-\varepsilon)\sigma+\varepsilon\sigma^{'})>\pi(\sigma,(1-\varepsilon)\sigma+\varepsilon\sigma^{'}).$$

Since $\pi(\sigma,(1-\varepsilon)\sigma+\varepsilon\sigma^{'})$ is the average expected fitness of incumbents, there exists a $\theta^{'}\in C(\mu)$ such that

$$\Pi_{\theta}((1-\varepsilon)\mu+\varepsilon\theta \mid \overline{b})>\Pi_{\theta^{'}}((1-\varepsilon)\mu+\varepsilon\theta \mid \overline{b}),$$

showing that (σ,σ) is not stable, a contradiction.

ii) (if) Suppose that σ is an NSS, and consider the monomorphic population consisting of θ^{o}'s, where everyone is playing σ. For any entrant, the set of focal post-entry equilibria is, clearly, non-empty. Consider any focal post-entry equilibrium, $\overline{b}$, in which the entrant, θ, plays $\sigma^{'}$. We have

$$\Pi_{\theta^{o}}((1-\varepsilon)\theta^{o}+\varepsilon\theta \mid \overline{b})=\pi(\sigma,(1-\varepsilon)\sigma+\varepsilon\sigma^{'})\geq\pi(\sigma^{'},(1-\varepsilon)\sigma+\varepsilon\sigma^{'})=\Pi_{\theta}((1-\varepsilon)\theta^{o}+\varepsilon\theta \mid \overline{b}),$$

where the inequality follows from the fact that σ is an NSS. Hence, (σ,σ) is stable. ■

Proof of Proposition 7. Consider a monomorphic population consisting of θ, for which a^{*} strictly dominates any other strategy. Let $\theta^{'}$ be any entrant and $\overline{b}$ be any focal equilibrium in the post-entry population. θ will play a^{*} regardless of whether she observes the entrant's preferences or not. We have

$$\Pi_{\theta}((1-\varepsilon)\mu+\varepsilon\theta^{'} \mid \overline{b})=(1-\varepsilon)\pi(a^{*},a^{*})+\varepsilon[p\pi(a^{*},b_{\theta^{'}}(\theta))+(1-p)\pi(a^{*},b_{\theta^{'}})],$$

and

$$\begin{aligned}\Pi_{\theta^{'}}((1-\varepsilon)\mu+\varepsilon\theta^{'} \mid \overline{b})&=(1-\varepsilon)[p\pi(b_{\theta^{'}}(\theta),a^{*})+(1-p)\pi(b_{\theta^{'}},a^{*})]+\varepsilon[p^{2}\pi(b_{\theta^{'}}(\theta^{'}),b_{\theta^{'}}(\theta^{'}))\\&\quad+p(1-p)\pi(b_{\theta^{'}}(\theta^{'}),b_{\theta^{'}})+p(1-p)\pi(b_{\theta^{'}},b_{\theta^{'}}(\theta^{'}))+(1-p)^{2}\pi(b_{\theta^{'}},b_{\theta^{'}})].\end{aligned}$$

Now, notice that, unless $b_{\theta^{'}}(\theta)=b_{\theta^{'}}=a^{*}$,

$$\pi(a^{*},a^{*})>p\pi(b_{\theta^{'}}(\theta),a^{*})+(1-p)\pi(b_{\theta^{'}},a^{*}),$$

since (a^*, a^*) is a strict Nash equilibrium, and hence we can find an $\varepsilon^{'} > 0$ such that $\Pi_\theta(. \mid .) > \Pi_{\theta'}(. \mid .)\ \forall \varepsilon \in (0, \varepsilon^{'})$, making (a^*, a^*) stable. So, suppose that $b_{\theta'}(\theta) = b_{\theta'} = a^*$. In this case, we have

$$\Pi_{\theta'}(. \mid .) = \pi(a^*, a^*),$$

and

$$\Pi_{\theta'}(. \mid .) = (1-\varepsilon)\pi(a^*, a^*) + \varepsilon[p^2\pi(b_{\theta'}(\theta^{'}), b_{\theta'}(\theta^{'})) + p(1-p)\pi(b_{\theta'}(\theta^{'}), a^*) + p(1-p)\pi(a^*, b_{\theta'}(\theta^{'}) + (1-p)^2\pi(a^*, a^*)].$$

Since a^* is efficient,

$$\pi(a^*, a^*) \geq p^2\pi(b_{\theta'}(\theta^{'}), b_{\theta'}(\theta^{'})) + p(1-p)\pi(b_{\theta'}(\theta^{'}), a^*) + p(1-p)\pi(a^*, b_{\theta'}(\theta^{'}) + (1-p)^2\pi(a^*, a^*)].$$

Therefore, $\Pi_\theta(. \mid .) \geq \Pi_{\theta'}(. \mid .)$ for all $\varepsilon \geq 0$, proving that (a^*, a^*) is stable. ■

Proof of Proposition 8. Suppose that (a, a) is stable with μ. Let θ^o be the entrant, with share ε. For all ε, there exists a post-entry focal equilibrium in which incumbents play a regardless of what they observe, and θ^o plays a after observing any incumbent or observing nothing, and plays σ, where $\pi(\sigma, \sigma) > \pi(a, a)$, after observing itself. We have

$$\Pi_\theta(. \mid .) = \pi(a, a),\ \forall \theta \in C(\mu),$$

and

$$\Pi_{\theta^o}(. \mid .) = (1-\varepsilon)\pi(a, a) + \varepsilon[p^2\pi(\sigma, \sigma) + p(1-p)\pi(\sigma, a) + p(1-p)\pi(a, \sigma) + (1-p)^2\pi(a, a).$$

Since $\pi(\sigma, \sigma) > \pi(a, a)$, there exists a $\overline{p} \in (0, 1)$ such that (a, a) is not stable for $p \in (\overline{p}, 1)$. ■

Proof of Proposition 9. Suppose that (a, a) is stable with μ. If (a, a) is not a Nash Equilibrium, then there exists $a^* \in A$ such that $\pi(a^*, a) > \pi(a, a)$. Let the entrant, $\theta^{'}$, be such that a^* strictly dominates any other strategy. For all focal equilibria

$$\Pi_\theta(. \mid .) = (1-\varepsilon)\pi(a, a) + \varepsilon[p\pi(b_\theta(\theta^{'}), a^*) + (1-p)\pi(a, a^*)],\ \forall \theta \in C(\mu),$$

and

$$\Pi_{\theta'}(. \mid .) = (1-\varepsilon)[p \sum_{\theta \in C(\mu)} \pi(a^*, b_\theta(\theta^{'})) + (1-p)\pi(a^*, a)] + \varepsilon\pi(a^*, a^*).$$

Since $\pi(a^*, a) > \pi(a, a)$, there exists a $\overline{p} \in (0, 1)$ such that for all $p \in (0, \overline{p})$,

$$p \sum_{\theta \in C(\mu)} \pi(a^*, b_\theta(\theta')) + (1 - p)\pi(a^*, a) > \pi(a, a),$$

showing that (a, a) is not stable for $p \in (0, \overline{p})$. ■

References

[1] Becker, G. S., 1976, "Altruism, Egoism, and Genetic Fitness," *Journal of Economic Literature*, **14**, 817-26.

[2] Bester, H. and W. Guth, 1998, "Is Altruism Evolutionarily Stable?" *Journal of Economic Behavior and Organization*, **34**, 193-209.

[3] Bhaskar, V., 1998, "Noisy Communication and the Evolution of Cooperation," *Journal of Economic Theory*, **82**, 110-31.

[4] Binmore, K. and L. Samuelson, 1997, "Muddling through: Noisy Equilibrium Selection," *Journal of Economic Theory*, **74**, 235-65.

[5] Boyd, R. and P. J. Richerson, 1985, *Culture and the Evolutionary Process*, Chicago and London: University of Chicago Press.

[6] Carlsson, H. and E. van Damme, 1993, "Global Games and Equilibrium Selection," *Econometrica*, **61**, 989-1018.

[7] Cavalli-Sforza, L. L. and M. W. Feldman, 1981, *Cultural Transmission and Evolution: A Quantitative Approach*, Princeton: Princeton University Press.

[8] Dekel, E. and S. Scotchmer, 1999, "On the Evolution of Attitudes toward Risk in Winner-Take-All Games," *Journal of Economic Theory*, **87**, 125-43.

[9] Ellison, G., 1993, "Learning, Local Interaction, and Coordination," *Econometrica*, **61**, 1047-72.

[10] Ely, J. C., 2002, "Local Conventions," *Advances in Theoretical Economics*, **2**(1), Article 1. http://www.bepress.com/bejte/advances/vol2/iss1/art1

[11] Ely J. C. and O. Yilankaya, 2001, "Nash Equilibrium and Evolution of Preferences," *Journal of Economic Theory*, **97**, 255-72.

[12] Fershtman, C. and Y. Weiss, 1996, "Social Rewards, Externalities and Stable Preferences," mimeo.

[13] Frank, R. H., 1987, "If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience," *American Economic Review*, **77**, 593-604.

[14] Fudenberg, D. and D. K. Levine, 1998, *The Theory of Learning in Games*, Cambridge and London: MIT Press.

[15] Guth, W., 1995, "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, **24**, 323-44.

[16] Guth, W. and M. Yaari, 1992, "Explaining Reciprocal Behavior in a Simple Strategic Game" in U. Witt, *Explaining Process and Change-Approaches to Evolutionary Economics*, pp. 23-24, Ann Arbor: The University of Michigan Press.

[17] Hansson, I. and C. Stuart, 1990, "Malthusian Selection of Preferences," *American Economic Review*, **80**, 529-44.

[18] Harsanyi J. C. and R. Selten, 1988, *A General Theory of Equilibrium Selection in Games*, Cambridge and London: MIT Press.

[19] Hirshleifer, J., 1977, "Economics from a Biological Viewpoint," *The Journal of Law and Economics*, **20**, 1-52.

[20] Kandori, M., G. J. Mailath and R. Rob, 1993, "Learning, Mutation, and Long-Run Equilibria in Games," *Econometrica*, **61**, 29-56.

[21] Kim, Y. G. and J. Sobel, 1995, "An Evolutionary Approach to Pre-Play Communication," *Econometrica*, **63**, 1181-94.

[22] Mailath, G. J., 1998, "Do People Play Nash Equilibrium? Lessons from Evolutionary Game Theory," *Journal of Economic Literature*, **36**, 1347-74.

[23] Matsui, A., 1991, "Cheap Talk and Coordination in Society," *Journal of Economic Theory*, **54**, 245-58.

[24] Ok, E. A. and F. Vega-Redondo, 2001, "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, **97**, 231-54.

[25] Rabin, M., 1998, "Psychology and Economics," *Journal of Economic Literature*, **36**, 11-46.

[26] Robson, A. J., 1990, "Efficiency in Evolutionary Games: Darwin, Nash, and the Secret Handshake," *Journal of Theoretical Biology*, **144**, 379-96.

[27] Robson, A. J., 1996, "The Evolution of Attitudes to Risk," *Games and Economic Behavior*, **14**, 190-207.

[28] Robson, A. J., 2001, "The Biological Basis of Economic Behavior," *Journal of Economic Literature*, **39**, 11-33.

[29] Robson, A. J. and F. Vega-Redondo, 1996, "Efficient Equilibrium Selection in Evolutionary Games with Random Matching," *Journal of Economic Theory*, **70**, 65-92.

[30] Rogers, A. R., 1994, "Evolution of Time Preferences by Natural Selection," *American Economic Review*, **84**, 460-81.

[31] Rubin, P. H. and C. W. Paul II, 1979, "An Evolutionary Model of Taste for Risk," *Economic Inquiry*, **42**, 585-96.

[32] Samuelson, L., 1997, *Evolutionary Games and Equilibrium Selection*, Cambridge and London: MIT Press.

[33] Samuelson, L., 2001, "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, **97**, 225-30.

[34] Sethi, R. and E. Somanathan, 2001, "Preference Evolution and Reciprocity," *Journal of Economic Theory*, **97**, 273-97.

[35] Waldman, M., 1994, "Systematic Errors and the Theory of Natural Selection," *American Economic Review*, **84**, 482-97.

[36] Weibull, J. W., 1995, *Evolutionary Game Theory*, Cambridge and London: MIT Press.

[37] Yilankaya, O., 1999, *On the Evolution of Preferences*, PhD Dissertation, Northwesten University.

[38] Young, P., 1993, "The Evolution of Conventions," *Econometrica*, **61**, 57-84.