Review of Economic Studies (2007) **74**, 685–704 0034-6527/07/00240685$02.00

Evolution of Preferences[1]

EDDIE DEKEL
Northwestern University and Tel Aviv University
JEFFREY C. ELY
Northwestern University
and
OKAN YILANKAYA
University of British Columbia

First version received June 2004; *final version accepted July* 2006 (*Eds.*)

We endogenize preferences using the "indirect evolutionary approach". Individuals are randomly matched to play a two-person game. Individual (subjective) preferences determine their behaviour and may differ from the actual (objective) pay-offs that determine fitness. Matched individuals may observe the opponents' preferences perfectly, not at all, or with some in-between probability. When preferences are observable, a stable outcome must be efficient. When they are not observable, a stable outcome must be a Nash equilibrium and all strict equilibria are stable. We show that, for pure-strategy outcomes, these conclusions are robust to allowing almost perfect, and almost no, observability, with the notable exception that inefficient strict equilibria may fail to be stable with any arbitrarily small degree of observability (despite being stable with no observability).

1. INTRODUCTION

We study endogenous preferences using the "indirect evolutionary approach" according to which preferences induce behaviour, behaviour determines "success", and success regulates the evolution of preferences.[2] In the dynamic story that underlies our reduced-form analysis, a population of individuals is randomly matched to play a two-person game. Individual subjective preferences may differ from the objective pay-offs (*i.e.* fitness), and in general the population has heterogeneous preferences. Matched individuals play a Bayesian–Nash equilibrium determined by the individuals' preferences and their information about opponents' preferences. This behaviour then determines the aggregate outcome of the game, which in turn determines the relative fitness of the preferences in the population. Finally, the composition of the population evolves as those preferences that have yielded higher fitness will increase at the expense of those that have yielded lower fitness.

In common with much of the evolutionary literature, we propose a static solution concept to tractably capture the stable points of such a dynamic process. We use this concept primarily to investigate the stability of the resulting aggregate outcomes for any given (objective) game, but the model also sheds some light on the shape of stable preferences. We refer to the pair consisting of a distribution of preferences in the population and an aggregate outcome as a *configuration*.

[1] An earlier version of this paper appeared in Yilankaya (1999).

[2] Some of the early proponents of this idea are Becker (1976), Hirshleifer (1977), Rubin and Paul (1979), and Frank (1987). The formal model of the "indirect evolutionary approach" that we follow was pioneered by Güth and Yaari (1992) and Güth (1995).

To clarify the model we elaborate briefly on two features: the informational issues underlying how behaviour is determined and what constitutes a stable configuration. As noted, we assume equilibrium play in a match, which depends on the information a player has about her opponents. Our stability concept accommodates a range of alternatives regarding the players' information in a match. We begin by studying two extreme cases: first, where each player observes perfectly her opponents' preferences and second, where she only knows the distribution of preferences in the population. Next we consider intermediate cases, where each player observes the opponent's preferences with some intermediate probability (but not whether she herself was observed) primarily as a robustness check. In each case we assume that play corresponds to a Bayesian–Nash equilibrium of the game given the distribution of preferences and the players' information.

For a typical distribution of preferences there will be multiple modes of behaviour that form a Bayesian–Nash equilibrium within the population. Our stability criterion identifies when a distribution of preferences and a particular equilibrium will together form a stable configuration. The stability of a configuration hinges on how it responds to invasion by new preferences. Generally speaking, we wish to say that a configuration is unstable if some small invasion can move the configuration far away either because the invading preference outperforms the incumbents, thereby altering the distribution of preferences or because the entrants' presence necessarily causes a large change in aggregate behaviour.

The two methodological contributions of our study are that we consider various degrees of observability, and we allow for all possible preferences in the population. Regarding the second aspect, early studies of preference evolution starting with Güth and Yaari (1992) concentrated on observable preferences and demonstrated the possibility that non-fitness-maximizing preferences and non-Nash outcomes could be evolutionarily stable. A common theme was that certain non-fitness-maximizing preferences can have a commitment effect when they are observable. However, this literature was limited in that attention was restricted to a subset of possible preferences in some special games.

In Section 3, we study the case of perfect observability, allowing for all possible preferences. A key aspect of the model with observable preferences is that individuals can condition their behaviour on the specific match, effectively correlating their behaviour with the opponents'. This enables entrants to coordinate on efficient play, thereby destabilizing Nash equilibria of the objective game, so that preferences distinct from the objective pay-offs can be stable and induce play that is not a Nash equilibrium of the objective game. Indeed, we show in Proposition 2 that efficiency is a necessary condition for stability.[3] Proposition 1 provides a companion sufficiency result: efficient strict equilibria are stable. By the first proposition, those previous results on preference evolution that imposed restrictions on the possible preferences in the population can be valid without such restrictions only if they selected efficient equilibria. Moreover, it identifies efficiency as the driving force behind the selection of behaviour (rather than altruism, spite, or other features of preferences).

Our Section 4 studies the case where players do not observe their opponents' preferences and know only the distribution of preferences in the population. In this case, an entrant's play is necessarily independent of a particular opponents' play and hence, as stated in Proposition 5, any non-Nash equilibrium outcome can be destabilized by an entering population with preferences that induce the (objective) best reply. It is also straightforward to show that any strict Nash

[3] The logic is reminiscent of the "secret handshake" result of Robson (1990) and related studies of evolutionary stability in games with communication: a population of entrants with preferences that enable cooperation among themselves and maintaining the previous equilibrium with the existing population destabilizes any inefficient outcome.

equilibrium outcome is stable.[4] Thus being Nash equilibrium is necessary and strict equilibrium, sufficient, for an outcome to be stable when there is no observability.

In Section 5, we develop our second contribution. As suggested by Samuelson (2001), it is important to investigate the robustness of these polar cases. Indeed, our model can accommodate varying assumptions on observability, and we use this to investigate the robustness of the preceding results, for the case of pure-strategy outcomes.[5] Our Proposition 7 shows that in this sense our first necessary result is robust: efficiency is a necessary condition for pure-strategy outcomes to be stable when observability is almost perfect.[6] The necessity result with no observability is similarly robust: even when there is a small degree of observability, a pure-strategy outcome is stable only if it is a Nash equilibrium (Proposition 8).[7]

Regarding the sufficiency conditions, efficient strict equilibrium outcomes remain stable with any degree of observability. The most interesting conclusion, however, is that the sufficiency result of the unobservable preferences case is not robust. We provide a coordination game example, where the outcome of a *strict* Nash equilibrium is not stable for *any* strictly positive probability of observing preferences. The unstable strict Nash equilibrium is pay-off dominated, suggesting that the efficiency force for observable preferences has implications for any degree of observability; only when preferences are completely unobservable does this force disappear.

The last modelling issue to discuss is that of existence. As is typical in studies that adopt a static solution concept to capture a dynamic process, existence will not be guaranteed in general. In our case, existence problems also arise as a result of a tension between efficiency, which is necessary for stability in the almost observable case and Nash equilibrium, which is necessary in the almost unobservable case. If in a game without an efficient Nash equilibrium the regions where these conditions are necessary overlap, then there does not exist a stable outcome. In those cases where existence fails, analysis of a fully dynamic model would be a useful direction for future research.[8] Our results identifying stable configurations when they exist should therefore be viewed as a first step in a more general theory of the joint evolution of preferences and behaviour.

2. THE MODEL

2.1. *The environment*

We study a symmetric two-player normal-form game G with a finite action set $A = \{a_1, a_2, \ldots, a_n\}$, and a pay-off function $\pi : A \times A \to \mathbf{R}$. We interpret, as is standard in the evolutionary game theory literature, the pay-offs as representing "success" or "fitness". Let Δ represent the set of mixed actions in G; the pay-off 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 1 to a_i, and we adopt this convention for

[4] On the other hand, some Nash outcomes will be unstable, and our concept is therefore a refinement of Nash equilibrium. These results are consistent with those in Ely and Yilankaya (2001) and Ok and Vega-Redondo (2001) who also studied general preference evolution with no observability.

[5] While Ok and Vega-Redondo (2001) do not directly allow for different assumptions on observability, some aspects of those differences can be seen through variations in their matching technology. They use this to argue that preference evolution has no effect on outcomes when preferences are not observed and show by example that there might be such an effect with observability. Their remark 4 (see also pp. 244–245) provides a more detailed discussion of these issues and our papers.

[6] This is in the same spirit as Sethi and Somanathan (2001), who showed in subsequent work that reciprocal (non-fitness) preferences evolve with perfect and almost-perfect observability in a class of games.

[7] Güth (1995) also considers the case of partial observability. His model differs in many ways so a detailed comparison would not be insightful; one important difference is that he models almost perfect observability by having the preferences be common knowledge with probability p close to 1, whereas in our model they are only common p belief.

[8] A recent paper, von Widekind (2004), restores existence via a different extension. He extends our model with observability to allow for non-expected utility preferences (we assume expected utility). This guarantees existence in all 2 × 2 games and extends our efficiency result: stability is equivalent to efficiency in these games.

all probability distributions throughout the paper. We are interested in what *outcomes* in G are stable, where an outcome is a probability distribution on $A \times A$.

We imagine a large population randomly and repeatedly matched to play G, or more accurately, to play a game that has the same action set as G. In standard evolutionary models, each player is assumed to play a particular action in G. Instead, we 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) utility functions on $A \times A$. We will often refer to $\theta \in \Theta$ as a "preference type" or "type". We write $\theta(\sigma, \sigma')$ for the expected utility of type θ when she plays σ and her opponent plays σ'. The environment will be described by a probability distribution on Θ, representing the distribution of preferences in the population. We will restrict attention to distributions that have finite supports, reflecting the assumption that the population is large but finite. Let $\mathcal{P}(\Theta)$ be the set of all possible finite support probability distributions on Θ. Finally, let $C(\mu)$ denote the support of $\mu \in \mathcal{P}(\Theta)$.

To complete the description of the strategic interaction between players, we need to specify what they know about each other's preferences when they are matched. We focus on the complete information scenario, where players observe their opponents' preferences, and the unobservable preferences case. We also allow for intermediate information structures where players observe their opponents' preferences with probability $p \in (0,1)$.[9]

2.2. *The solution concept*

We present a reduced-form stability concept intended to capture the essential features of the following three components of the evolutionary process: *mutation*, which introduces preferences into the population, *optimization*, by which agents adapt their behaviour given the preferences represented in the population, and *natural selection*, by which the preference composition is updated as successful preferences replicate. We model mutation by considering exogenous changes in the distribution of preferences resulting from the entry of new types in small proportions. Given a preference distribution μ, it is assumed that the population learns to play a Bayesian–Nash equilibrium of the incomplete information game defined by μ and the information structure.[10] Finally, natural selection is modelled by a static stability concept. The concept identifies populations of preferences that cannot be invaded by mutants who—in the resulting equilibrium—have larger fitness pay-offs than the incumbents.

2.2.1. Equilibrium play in a match. Suppose that the distribution of preferences in the population is given by μ. The interaction can be analysed via the following two-player Bayesian game, $\Gamma_p(\mu)$. The preferences of the two players are drawn independently from μ, and each player with independent probability p observes the preferences of the other. With the complementary probability $1-p$, the player observes the uninformative signal $\emptyset$.

A strategy for preference θ is a rule $b_\theta : C(\mu) \cup \emptyset \to \Delta$ specifying a mixed action conditional on each possible observation. We assume that aggregate play in the population corresponds to a

[9] Thus the "type" of a player in the usual sense in these games of incomplete information designates both their preference type, θ, and their private information (whether they observed an opponent and what they observed). In the extreme cases where $p \in \{0,1\}$ only the preference type matters for the game, and hence we refer to this as the players' type.

[10] While it is not a part of our formal model, the justification for our concept is based on the view that equilibrium play arises from a process (*e.g.* of learning), which operates much faster than the evolutionary process we model. Whenever the distribution of preferences changes, we assume that the learning process always reaches equilibrium play before subsequent evolution proceeds.

symmetric Bayesian–Nash equilibrium of this game. That is, we assume that each individual, upon being selected to play has correct beliefs about the distribution of her opponents' play and chooses a mixed action that is a best reply to this belief according to her own preferences.

When type θ is matched with type θ' and plays a mixed action σ, the expected utility for θ is

$$p\,\theta(\sigma, b_{\theta'}(\theta)) + (1-p)\,\theta(\sigma, b_{\theta'}(\emptyset)).$$

This pay-off is the average over two possibilities. With probability p, the opponent observes the preferences of type θ and thus plays $b_{\theta'}(\theta)$, and with probability $1-p$, the opponent observes $\emptyset$ and plays $b_{\theta'}(\emptyset)$.

An equilibrium b is thus characterized by two properties. First, type θ chooses an optimal action conditional on observing that the opponent's type is θ':

$$b_\theta(\theta') \in \arg\max_{\sigma \in \Delta} [p\theta(\sigma, b_{\theta'}(\theta)) + (1-p)\theta(\sigma, b_{\theta'}(\emptyset))], \tag{1}$$

for each $\theta' \in C(\mu)$. Second, type θ chooses an optimal action conditional on observing nothing informative:

$$b_\theta(\emptyset) \in \arg\max_{\sigma \in \Delta} \sum_{\theta' \in C(\mu)} [p\theta(\sigma, b_{\theta'}(\theta)) + (1-p)\theta(\sigma, b_{\theta'}(\emptyset))]\mu(\theta'), \tag{2}$$

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

For the observable preferences case, that is, $p = 1$, we ignore (2), and the conditions reduce to the requirement that in each (θ, θ') match, play forms a Nash equilibrium of the complete information game with pay-off functions θ and θ'. For the unobservable preferences case, that is, $p = 0$, we ignore (1). Let $B_p(\mu)$ denote the set of all Bayesian–Nash equilibria of the game $\Gamma_p(\mu)$.

Fitness, preference evolution, and stability. Given a population distribution μ and an equilibrium $b \in B_p(\mu)$, the average fitness of type $\theta \in C(\mu)$ is denoted $\Pi_\theta(\mu \mid b)$ and is given by equation (3).

$$\Pi_\theta(\mu \mid b) = \sum_{\theta' \in C(\mu)} [p^2 \pi(b_\theta(\theta'), b_{\theta'}(\theta)) + p(1-p)\pi(b_\theta(\theta'), b_{\theta'}(\emptyset)) \\ + p(1-p)\pi(b_\theta(\emptyset), b_{\theta'}(\theta)) + (1-p)^2 \pi(b_\theta(\emptyset), b_{\theta'}(\emptyset))]\mu(\theta'). \tag{3}$$

This fitness, which depends on the equilibrium played, is the measure of evolutionary success for types. Consequently, evolution depends both on the distribution of preferences *and* the equilibrium played given this distribution. Hence, our stability definition applies to *configurations*, (μ, b), where $b \in B_p(\mu)$. Every configuration induces a distribution over actions, called the *outcome*, denoted by $x(\mu, b)$. When there is no chance of confusion, we may drop the arguments of $x(\mu, b)$ for expositional purposes.

A configuration is stable if it satisfies two conditions. First, it must be *balanced*: all types present must receive the same fitness. If the configuration were not balanced, then some types have higher fitness than others and natural selection would alter the configuration as the former types multiply and the latter types recede.

Definition 1. A configuration (μ, b) is *balanced* if $\Pi_\theta(\mu \mid b) = \Pi_{\theta'}(\mu \mid b)$ for all θ, θ' in $C(\mu)$.

Second, a stable configuration must resist entry by mutants. There are two ways in which a mutation can destabilize the configuration. First, the mutant type could achieve a higher fitness than the incumbent types. Selection would then favour the mutant and the distribution of preferences would diverge from the original configuration. Second, the behaviour of the mutant could unravel the original equilibrium behaviour causing the distribution of actions to diverge.[11] Our stability definition identifies configurations that are immune to either type of change.

To precisely formulate a definition along these lines we must make some assumptions about which post-entry equilibria are relevant. If all equilibria are considered, then stability is too hard to satisfy: whenever the original population admits multiple equilibria, any entry could destabilize it by triggering a switch to another equilibrium. Instead, we define the subset of equilibria that are *focal* relative to the original configuration, and assume that *any* focal equilibrium can arise after an entry. To motivate our definition, consider the observable case, and imagine that a mutation has taken place resulting in the entry of a new preference. Prior to the entry, incumbent types have played against one another long enough to learn an equilibrium $b \in B_1(\mu)$. We assume that entry by a small group of new types will not undo this. On the other hand, the incumbents had no previous experience being matched against the new type, so any outcome in such matches (consistent with equilibrium) is a plausible result of the ensuing adaptation process.

To develop the formal definition, we shall introduce some notation. If the original distribution of preferences is μ and the mutant preference is $\tilde{\theta}$, we define $N_\varepsilon(\mu,\tilde{\theta})$ to be the set of all preference distributions resulting from entry by no more than ε mutants. Formally,

$$N_\varepsilon(\mu,\tilde{\theta}) = \{\mu' : \mu' = (1-\varepsilon')\mu + \varepsilon'\tilde{\theta},\, \varepsilon' < \varepsilon\}.$$

Beginning with a configuration (μ, b), and following an entry by $\tilde{\theta}$ leading to $\tilde{\mu} \in N_\varepsilon(\mu,\tilde{\theta})$, an equilibrium $\tilde{b} \in B_p(\tilde{\mu})$ is *focal* if incumbents' behaviour is unchanged, that is, $\tilde{b}_\theta(\theta') = b_\theta(\theta')$ (whenever $p > 0$) and $\tilde{b}_\theta(\emptyset) = b_\theta(\emptyset)$ (whenever $p < 1$) for all $\theta, \theta' \in C(\mu)$. Notice that a focal equilibrium does not restrict the behaviour of entrants, nor does it restrict the behaviour of incumbents when they observe that they have been matched with entrants. Let $B_p(\tilde{\mu} \mid b)$ denote the set of all focal equilibria relative to b if the distribution is $\tilde{\mu}$.

We assume that any focal equilibrium can potentially arise following a mutation, and thus our definition of stability requires that in all of them, entrants earn no higher fitness than any incumbent. However, not all post-entry populations will have focal equilibria. In that case, the adaptation process would lead to some other equilibrium in $B_p(\tilde{\mu})$. If all these equilibria involve play that is "far" from the original configuration, then clearly the configuration is not stable. Our definition requires that there exist some nearby post-entry equilibria and that in all nearby equilibria the entrants do not outperform incumbents.[12]

To present our formal definition of stability, we first define the set of "nearby" equilibria.

Definition 2. Given a configuration (μ, b), a parameter δ, and a post-entry population $\tilde{\mu} \in N_\varepsilon(\mu,\tilde{\theta})$, let $B_p^\delta(\tilde{\mu} \mid b) = \{\tilde{b} \in B_p(\tilde{\mu}) : |x(\tilde{b},\tilde{\mu}) - x(b,\mu)| < \delta\}$.

[11] The second can happen without the first: consider a Prisoners' Dilemma (PD) game, and $p = 0$. Suppose the incumbents have preferences, which make *defect* a weakly, but not strictly, dominant action, and suppose they are cooperating. When a mutant type enters and plays defect, the incumbents will also switch to defect. Although the mutants earn no higher pay-off than the incumbents, the outcome switches from all cooperate to all defect. The cooperative outcome is therefore not stable.

[12] If there are no focal equilibria, but there are "nearby" equilibria, then there are several alternative ways to proceed. First, rather than require only the existence of *some* nearby equilibrium, we could insist that all post-entry equilibria are close to the original. Second, rather than requiring that the incumbents outperform the entrants in all nearby equilibria, we could require only that this be true in at least one nearby equilibrium. We chose our version largely for consistency with the spirit behind the rest of the definition, but we emphasize that all of our results would hold under any of these versions.

Definition 3. A configuration (μ, b) is stable if it is balanced and if for every $\delta > 0$ there exists $\varepsilon > 0$ such that for every $\tilde{\theta} \in \Theta$ and $\tilde{\mu} \in N_\varepsilon(\mu, \tilde{\theta})$,

1. $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$ for all $\tilde{b} \in B_p(\tilde{\mu} \mid b)$ and $\theta \in C(\mu)$.
2. If $B_p(\tilde{\mu} \mid b) = \emptyset$, then $B_p^\delta(\tilde{\mu}) \neq \emptyset$, and $\Pi_\theta(\tilde{\mu} \mid \tilde{b}) \geq \Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b})$ for all $\tilde{b} \in B_p^\delta(\tilde{\mu})$ and $\theta \in C(\mu)$.

An outcome x is stable if there exists a stable configuration with that outcome, *i.e.* there exists a stable (μ, b) with $x = x(\mu, b)$. A preference distribution μ is stable if (μ, b) is stable for some b.

Finally, as mentioned in the Introduction, the definition of stability is demanding, and hence, like in many static evolutionary concepts, existence is not guaranteed.

3. OBSERVABLE PREFERENCES

When preferences are observable, any pair of incumbents are playing a complete information game given their preferences, isolated from the rest of the population. Our stability definition requires the incumbents to do as well as any mutant (in small proportions) in all post-entry focal equilibria, that is, those equilibria in which any pair of incumbents continue to play as they were playing prior to the entry of the mutant. Note that the set of focal post-entry equilibria is non-empty by definition in the observable preferences case.

We call a symmetric strategy profile *efficient* if its fitness is highest among all symmetric strategy profiles.

Definition 4. (σ^*, σ^*) is *efficient* if $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma, \sigma)$ for all $\sigma \in \Delta$.

When (σ^*, σ^*) is efficient, we refer to σ^* as the efficient strategy and $\pi(\sigma^*, \sigma^*)$ as the efficient fitness. Our first result specifies a sufficient condition for stability: if a pure-strategy profile (a^*, a^*) is efficient as well as a strict Nash equilibrium of G, then it is stable.[13] The reason is straightforward: consider a population consisting of types for which a^* is a strictly dominant strategy, and consider any entrant type. If the entrants' play against the incumbents puts zero weight on a^*, 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. The proof adds to these observations by finding a uniform barrier ε that would work for all possible mutants with population shares less than ε, even those with actions arbitrarily close to a^* against the incumbents.

Proposition 1. *If (a^*, a^*) is both efficient and a strict Nash equilibrium of G, then it is stable.*

We now turn our attention to necessary conditions for stability. We will show that if an outcome is stable, then all incumbent types in the stable distribution receive the same fitness in *each* of their interactions, including those with their own types. (Stability only implies that they do as well on average.) Therefore, the average fitness of every type in a stable distribution must equal the pay-off of some symmetric strategy profile in G. Moreover, this average fitness must be efficient.

[13] All proofs are in Appendix.

Efficiency is a necessary condition for stability. The idea behind this result is simple, and best demonstrated with monomorphic populations, where its "secret handshake" flavour is clear.[14] Suppose the incumbent's fitness is less than the efficient pay-off. We can always find a mutant that would do better than the incumbent in the post-entry population. Consider, for example, a coordination game. The outcomes of the "bad equilibria" are not stable, because a mutant whose preferences coincide with the fitness function can invade by playing, as part of a post-entry equilibrium, the bad action against the incumbent and the good one against itself. Consider next a PD game. The defection outcome is not stable. Any population where defection is played can be invaded by a mutant who has "coordination" preferences, where defection (respectively, cooperation) is the unique best response to itself. There is a post-entry equilibrium in which the mutant and the incumbent both defect whenever they are matched and the mutant cooperates against itself. Our necessity result shows that these arguments can be generalized. We use entry by an indifferent type in the proof of this result; we discuss this further in Section 6.

Proposition 2. *If an outcome x^* is stable with configuration (μ^*, b^*), then*

$$\Pi_\theta(\mu^* \mid b^*) = \pi(b^*_\theta(\theta'), b^*_{\theta'}(\theta)) = \pi(\sigma^*, \sigma^*),$$

for all $\theta, \theta' \in C(\mu^*)$, where (σ^*, σ^*) is efficient.

By combining Proposition 1 and Proposition 2, we obtain a unique prediction in terms of stability for a class of games which includes much-studied coordination games: if the only efficient profile is a strict Nash equilibrium, then its outcome is the only stable one.

3.1. 2×2 *Games*

In this subsection we focus on 2×2 games, which attracted considerable attention from the evolutionary game theory literature. We are able to give a characterization of both stable outcomes and the stable distributions of preferences for this class of games. In Proposition 2 we showed that efficiency was necessary for stability. It turns out that efficiency of a pure strategy is sufficient for the corresponding outcome to be stable in 2×2 games. Moreover, games with a mixed efficient strategy do not have stable outcomes with the exception of a non-generic class of Hawk–Dove games. Hence, existence of an efficient pure strategy is both necessary and sufficient for a stable outcome to exist in generic 2×2 games. We then characterize the stable distributions of preferences. Types that make cooperation a stable outcome in PD games are particularly interesting. They all belong to an equivalence class that has a "secret handshake" flavour: they cooperate in equilibrium only if their opponent is cooperating with probability one.

In order to simplify our exposition in this subsection, we now introduce some notation while making a basic observation about 2×2 games. Consider any 2×2 (normal) game form with the strategy set $\{A, B\}$. In terms of equilibrium behaviour, all possible preference relations belong to one and only one of the following equivalence classes: $\mathcal{AA}$, $\mathcal{AB}_\alpha$, $\mathcal{BA}_\alpha$, $\mathcal{BB}$, and θ^0, where $\alpha \in [0, 1]$, and

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

[14] In a *monomorphic* population all individuals have the same preferences; in a *polymorphic* population preferences may differ.

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

All players with preferences belonging to the same equivalence class will have the same set of equilibria in any game (defined by the set of players and their pay-off functions, in addition to the game form), and hence are referred to as a type. $\mathcal{AA}$ represents all preference relations where A strictly dominates B, and any player whose preferences belong to $\mathcal{AA}$ will play A in any equilibrium of any game. Similarly, a player whose preferences are in $\mathcal{AB}_\alpha$, 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 only if her opponent plays A with probability α. For example, in any game involving an $\mathcal{AB}_\alpha$ player and an $\mathcal{AA}$ player, the unique equilibrium is (A, A). Similarly, when an $\mathcal{AB}_\alpha$ is matched with an $\mathcal{AB}_\beta$, where $\alpha, \beta \in (0, 1)$, there are three equilibria: (A, A), (B, B), and the mixed one in which $\mathcal{AB}_\alpha$ (respectively, $\mathcal{AB}_\beta$) plays A with probability β (respectively, α).

We next present a result that, together with Proposition 2, characterizes stable outcomes in 2 × 2 games. Without loss of generality assume $a \geq d$ and let G be

$$\begin{array}{c|c|c|} & A & B \\ \hline A & a, a & b, c \\ \hline B & c, b & d, d \\ \hline \end{array}\,.$$

Proposition 3.

(a) *If (A, A) is efficient, then it is stable.*
(b) *If (A, A) is not efficient, then the efficient (σ^*, σ^*) is stable iff $b = c > a$ (otherwise there is no stable outcome).*

We next turn our attention to stable distributions, that is, to the preferences that are selected by the evolutionary forces. In the PD game where A (cooperate) is efficient, (A, A) is stable with a monomorphic population of $\mathcal{AB}_1$, a type that is indifferent between A and B when the opponent plays A and strictly prefers B when the opponent plays B. This type has a secret handshake flavour, and moreover, it is immune to "sucker punches": it never cooperates unless its opponent is cooperating with probability 1. Therefore, no type can enter and obtain a strictly higher average fitness than $\mathcal{AB}_1$ in any post-entry equilibrium. It turns out that this stable distribution is unique. We illustrate this with monomorphic populations and also discuss why it is not unreasonable to consider the cooperation outcome to be stable with $\mathcal{AB}_1$ even though for this type cooperation is a weakly dominated strategy.

It is clear why $\mathcal{AA}$, a type that always cooperates, cannot be stable: $\mathcal{BB}$ enters, and in the unique equilibrium it defects while the incumbent is cooperating. Can a cooperating $\mathcal{AB}_\alpha$ type be stable? These types have the secret handshake flavour: in equilibrium, they defect when the opponent defects and cooperate when the opponent cooperates. The problem arises from the mixed-strategy equilibrium against the mutant $\mathcal{AB}_\beta$, where $\beta > \alpha$. In this equilibrium $\mathcal{AB}_\alpha$ cooperates with probability β and $\mathcal{AB}_\beta$ cooperates with probability α, that is, $\mathcal{AB}_\alpha$ is cooperating more than $\mathcal{AB}_\beta$ and hence obtaining lower fitness. Heuristically, selection leads cooperating $\mathcal{AB}_\alpha$ types to be replaced with $\mathcal{AB}_\beta$ types, with $\beta > \alpha$, who are also cooperating among themselves.[15]

[15] If these types are not cooperating among themselves, then they will be replaced by types who will. As we argued before, efficiency is a necessary condition for stability.

The limit of these types is $\mathcal{AB}_1$, and cooperation occurs with probability 1 in the limit of the mixed-strategy equilibria that enable the mutants to destabilize incumbent populations. Moreover, $\mathcal{AB}_1$ itself does not have the problem of cooperating more than its opponent in equilibrium. Against any type, it cooperates with positive probability only if its opponent is cooperating with probability 1. Hence, a monomorphic population of $\mathcal{AB}_1$ is the unique stable distribution. Note that $\mathcal{AB}_1$ does not represent generic preference relations for a normal form game. Furthermore, stability requires $\mathcal{AB}_1$ to play (A, A) when matched with itself, which is a non-perfect equilibrium, since A is weakly dominated for $\mathcal{AB}_1$. However, as our discussion above illustrates, both the preferences and the equilibrium played given these preferences are endogenous and arise from evolutionary selection.

We also show that in coordination games any type for which (A, A) is an equilibrium when matched against itself can be found in a stable distribution. In Hawk–Dove games where the efficient strategy is mixed, a monomorphic population of $\mathcal{AB}_{\alpha^*}$ is the unique stable distribution, where α^* is the weight that the efficient strategy puts on A. This is curious, since the fitness function of these games is given by $\mathcal{BA}_{\alpha^*}$. However, $\mathcal{BA}_{\alpha^*}$ cannot be in any stable distribution, because a type for which "Hawk" is a dominant strategy $(\mathcal{AA})$ can enter and obtain a strictly higher average fitness than $\mathcal{BA}_{\alpha^*}$, since $\mathcal{BA}_{\alpha^*}$ plays "Dove" in the unique equilibrium when it is matched with $\mathcal{AA}$.

Proposition 4. *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.*

1. *For games in which (A, A) is efficient, that is, $a > \pi(\sigma, \sigma)\ \forall \sigma \neq a$:*
   (a) *If $a > c$ and $a > b$, then $M(G) = \{\mathcal{AA}, \mathcal{AB}_\alpha, \mathcal{BA}_1, \theta^0\}$, $\alpha \in [0, 1]$.*
   (b) *If $a > c$ and $b > a$, then $M(G) = \{\mathcal{AA}, \mathcal{AB}_\alpha\}$, where $\alpha < \frac{a-d}{b-d}$.*
   (c) *If $c > a$, then $M(G) = \{\mathcal{AB}_1\}$.*
2. *For games in which (A, A) is not efficient, if $b = c > a$, then $M(G) = \{\mathcal{AB}_{\alpha^*}\}$, where σ^* is efficient and $\alpha^* = \sigma^*(A)$.*

4. UNOBSERVABLE PREFERENCES

Given the symmetric nature of the interaction in the population, only outcomes of symmetric strategy profiles of the fitness game G are relevant in the unobservable preferences case: the outcome induced by any strategy profile in $\Gamma_0(\mu)$ will be (σ, σ), for some $\sigma \in \Delta$. The first part of our result with unobservable preferences revives the "stable only if Nash" folk theorem: Nash behaviour (relative to the objective fitness function) is a necessary condition for stability. The reason is intuitive. Consider a monomorphic population and suppose that the incumbent type is taking an action $\sigma \in \Delta$, where (σ, σ) is not a Nash equilibrium of G. This means that there exists a pure action $a_i \in A$ that is a strictly better response, in terms of fitness, to σ than σ itself. Therefore, a mutant type for which a_i is strictly dominant can obtain a strictly higher fitness than the incumbent type in all focal post-entry equilibria as long as its population share is small enough. Moreover, a_i is a strictly better response to actions that are "close enough" to σ as well. We conclude that (σ, σ) is not stable. If the population is polymorphic the argument is extended by noting that then some type in the population does worse than the entrant, and hence the configuration is not stable.

While Nash equilibrium is necessary for stability, not all Nash equilibria are stable. The mutations introduced in the evolutionary process induce trembles, and hence it is easy to show by example that they will refine the set of equilibria. This is why the second part of our result, presenting sufficient conditions for stability, must focus on a refinement of Nash equilibrium.

While the precise refinement would depend on the fine details of the stability notion, the arguments used suggest that the necessity of Nash equilibrium for stability, and similarly the sufficiency of strict equilibrium discussed next, are robust.[16]

The argument that every strict Nash equilibrium outcome is stable is also intuitive. Fix a strict Nash equilibrium (a^*, a^*), and consider an incumbent type for which a^* is strictly dominant. The incumbent will play a^* in any equilibrium given any distribution of preferences, so every post-entry equilibrium is focal. Since (a^*, a^*) is a strict Nash equilibrium, there does not exist any mutant that can obtain a strictly higher fitness than the incumbent in any post-entry equilibrium, as long as its population share is small enough. Therefore, (a^*, a^*) is stable.

Proposition 5.

(a) (σ, σ) *is stable only if it is a Nash equilibrium of* G.
(b) *If* (a_i, a_i) *is a strict Nash equilibrium of* G, *then it is stable.*

This result illustrates the close relation between evolution of preferences and the standard evolutionary models when preferences are unobservable. The basic intuition for this is straightforward: if others cannot observe one's preferences, and hence condition their behaviour on that, there is no advantage in having preferences that differ from the fitness function, which is the determinant of evolutionary success.

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 her opponent's preferences with probability $p \in (0, 1)$ independent of what her opponent observes. Given the distribution of preferences in the population, the interaction can be analysed as a symmetric two-player Bayesian game. We assume that the aggregate play in the population corresponds to a Bayesian–Nash equilibrium of this game.

Our main objective in this section is to study whether stability results in the observable (respectively, unobservable) preferences case continue to hold for high (respectively, low) values of probability of observability. In other words, are the results of previous sections "continuous" in p?

Our first result is, actually, independent of the probability of observability: efficient strict Nash equilibria are stable.

Proposition 6. *If* (a^*, a^*) *is both efficient and a strict Nash equilibrium of* G, *then it is stable for all* $p \in (0, 1)$.

Note that the same result also holds when preferences are observable (Proposition 1) or unobservable (Proposition 5).

We next present a "discontinuity" result, showing the importance of efficiency even for arbitrarily low levels of observability. We know from Proposition 5 that strict Nash equilibrium outcomes are stable when preferences are unobservable. The following coordination game demonstrates that this is not true in the case of imperfectly observable preferences, even for

[16] In an earlier version of this paper (Dekel, Ely and Yilankaya, 2004) we used a slightly different notion of stability under which an outcome was stable if and only if it was induced by a neutrally stable strategy. Here we focus on results that we believe would be robust to other variants of stability.

arbitrarily small values of p. In this example, (B, B) is not stable for any $p > 0$, despite being a strict (and risk-dominant (Harsanyi and Selten, 1988)) Nash equilibrium.

Example 1. Consider the following game:

	A	B
A	$6, 6$	$0, 5$
B	$5, 0$	$2, 2$

If entrants observed one another perfectly, it is clear they could invade the B-playing incumbent population, obtaining 6 when matched against themselves and 2 otherwise. With almost no observability, if entrants play A (only) upon observing another entrant, the observing entrant almost surely obtains 0. However, the observed entrant obtains 5, so on average the entrants obtain 2·5, which is better than incumbents. The insight here is that an entering population with coordination preferences can do better on average than the inefficient incumbents: entrants who observe that their opponents are also entrants, do worse than incumbents, but as they are observed just as often as they observe, this behaviour on average is beneficial.

Formally, suppose that (B, B) is stable with distribution μ^*. Notice that every incumbent must be a preference relation for which B is a best response to itself. Consider an entrant type with "coordination" preferences AB_α, where $\alpha \in (0, p]$, that is, preferences where A is a best response to any mixed strategy in which A is chosen with probability more than α. There is a focal post-entry equilibrium in which the entrant plays A if it observes itself and plays B otherwise, and the incumbents continue to play B regardless of what they observe. To see this, note that all of the incumbents' opponents are always playing B, so it is a best response for the incumbents to play B. When the entrant observes an incumbent, the opponent is playing B. When the entrant does not observe anything, it is matched with an incumbent (that always plays B) with very high probability, since the share of the entrant in the population is only ε. 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 is a best response, since p is greater than α. For the focal post-entry equilibrium just specified, incumbents' and the entrant's fitnesses are given by, respectively,

$$\Pi_\theta(\cdot \mid \cdot) = 2, \ \forall \theta \in C(\mu^*)$$

and

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

It is straightforward to check that $\Pi_{\mathcal{AB}_\alpha}(\cdot \mid \cdot) > 2$ for all $p > 0$. Therefore, (B, B) is not stable for any $p > 0$.

On the other hand, a version of the "stable only if efficient" result of the observable preferences case (Proposition 2) does hold for a high enough probability of observability. Specifically, the outcome of a symmetric pure-strategy profile is not stable for p high enough, if that strategy is not efficient.

Proposition 7. *If (a_i, a_i) is not efficient, then there exists a $\overline{p} \in (0, 1)$ such that it is not stable for any $p \in (\overline{p}, 1)$.*

The selection issue between the risk-dominant and the pay-off-dominant equilibria in coordination games has been studied extensively using evolutionary models.[17] The risk-dominant equilibrium is selected in the models of Ellison (1993), Kandori, Mailath and Rob (1993), and Young (1993); the pay-off-dominant equilibrium is favoured in Robson and Vega-Redondo (1996), Ely (2002), and in studies of cheap talk, for example, Matsui (1991), Kim and Sobel (1995), Bhaskar (1998); in Binmore and Samuelson (1997) either can be selected.

We next summarize all our results so far in this section for the special case of coordination games. Our model of evolution of preferences provides support for selecting the efficient equilibrium in coordination games in the following sense:

Corollary 1. *Consider (strict) coordination games. The outcome of the risk-dominant equilibrium is not stable for large enough p, unless the equilibrium is also pay-off-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 pay-off-dominant equilibrium is stable for all $p \in [0, 1]$.*

As our last result, we show that a version of the "stable only if Nash" result (Proposition 5 *a*) of the unobservable preferences case holds for low enough probability of observability. That is, the outcome of a pure-strategy profile that is not a Nash equilibrium will not be stable for low enough p.

Proposition 8. *If (a_i, a_i) is not a Nash equilibrium of G, then there exists a $\overline{p} \in (0, 1)$ such that it is not stable for any $p \in (0, \overline{p})$.*

It is important to note the distinction between Proposition 2 and Proposition 5(*a*) that concern the polar cases, and the "robustness-confirming" Proposition 7 and Proposition 8 that consider the nearby models. The former hold for all outcomes, but we only show that the necessary conditions are preserved for *pure* outcomes. We do not know whether it is possible to obtain stronger results that are robust to various modelling choices we have made. This issue is especially relevant for the PD game. We showed that the cooperation outcome was stable when $p = 1$. Perfect observability is crucial for this result: cooperation is not stable for any $p < 1$.[18] This result and Proposition 7 suggest that there is no stable outcome in the PD when p is close to 1. However, as we mentioned above, such a conclusion does not immediately follow.[19]

6. DISCUSSION

We study a model of preference evolution for general preferences and allowing (a form of) partial observability. When preferences are (almost) fully observable, this evolution is a force towards efficient outcomes, as in the literature on evolution with secret handshakes and with cheap talk. When preferences are unobservable the "folk" result is that preference evolution does not select among strict Nash equilibria, and this result is confirmed here. However, we show that this conclusion is not robust, as the force towards efficiency can destabilize inefficient equilibria when

[17] Among the non-evolutionary models, Harsanyi and Selten (1988) selects the pay-off-dominant, whereas Carlsson and van Damme (1993) selects the risk-dominant equilibrium.

[18] The proof essentially follows from the discussion preceding Proposition 4 of stable preferences for PD and the observation in footnote 11: $\mathcal{AB}_1$ cannot belong to a stable distribution, since it will defect whenever the entrant does, causing a discrete shift in the outcome. Other cooperating types can be taken advantage of by entrants that "cooperate a little less than them."

[19] It is not difficult to show, however, that there is no stable *monomorphic population* when p is large enough. We do not know whether this holds for polymorphic populations as well, but we suspect that the finding will be sensitive to the exact definition of stability (see footnote 12).

even a small degree of observability is possible. On the positive side, efficient strict equilibria are stable when preferences are observable to any degree.

There are several modelling issues that deserve further discussion. Partial observability could also be modelled as having noisy signals, with full observability being the limit as the noise disappears. Sethi and Somanathan (2001) and Heifetz, Shannon and Spiegel (2004) looked at such models for restricted sets of preferences and classes of games. It would be worthwhile to extend our general analysis to such a specification.

We think that allowing for all possible preferences to evolve is natural. But by allowing all possible preferences, we make it easy to destabilize outcomes. For example, we used indifferent types as potential entrants to prove that efficiency is necessary for stability when preferences are observable. It seems of interest to explore the extent to which restricting attention to various subclasses of preferences would impact our conclusions.[20]

Finally, our approach treats the observability of preferences as exogenous and we analyse the consequences for stability of varying degrees of observability as a sort of comparative-statics exercise. However, in the long run, observability itself is subject to evolutionary forces. A more general model would incorporate the evolution of observability.

A related question that arises when assuming observable preferences is why we do not allow evolution of mutants who mimic the observable types in appearance, but not in behaviour. Allowing such mutants would not effect our negative results: the force towards efficiency would remain. The mutants that we currently allow are sufficient to destabilize the inefficient outcomes that we identify as unstable, and adding new mutants would not change this. However, intuition suggests that this additional level of mutation would prevent non-Nash outcomes from being stable. The idea is similar to the secret handshake results and the discussion in Robson (1990). On the other hand, our results could be interpreted as a medium-run analysis when this ability to mimic existing types evolves at a slower rate than the types themselves. For related ideas, see Wiseman and Yilankaya (2001).

APPENDIX

Proposition 1. *If (a^*, a^*) is both efficient and a strict Nash equilibrium of G, then it is stable.*

Proof. Let (a^*, a^*) be both efficient and a strict Nash equilibrium of G. Consider a monomorphic population consisting of θ^* for which a^* is strictly dominant. The outcome of the unique equilibrium among incumbents, b^*, is (a^*, a^*). We will show that there exists $\varepsilon \in (0,1)$ such that for all $\varepsilon' \in (0,\varepsilon)$, θ^* would obtain (weakly) higher average fitness than any mutant $\tilde{\theta}$ with a population share of ε', in any post-entry equilibrium, which will all be focal, since θ^* will play a^*, which is strictly dominant for θ^*, regardless of the opponent's type. Note that since the set of focal equilibria is non-empty (which is always the case when $p = 1$), the second part of the stability definition does not apply, so δ is irrelevant and we can select ε independently from δ.

Suppose that $\tilde{\theta}$ plays $\sigma' \in \Delta$ when matched with θ^*, and $\sigma'' \in \Delta$ when matched against itself in $\tilde{b} \in B_p(\tilde{\mu} \mid b^*)$, where $\tilde{\mu} = (1-\varepsilon')\theta^* + \varepsilon'\tilde{\theta}$. We will show that the stability condition is satisfied for all $\sigma', \sigma'' \in \Delta$. The average fitnesses of the incumbent and the mutant are

$$\Pi_{\theta^*}(\tilde{\mu} \mid \tilde{b}) = (1-\varepsilon')\pi(a^*, a^*) + \varepsilon'\pi(a^*, \sigma') \tag{4}$$

and

$$\Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b}) = (1-\varepsilon')\pi(\sigma', a^*) + \varepsilon'\pi(\sigma'', \sigma''). \tag{5}$$

[20] The indifferent type enables making a general argument with one entrant rather than using different entrants for different incumbent populations or for different games. For most cases this argument is not needed, and "coordination types" could be used instead. For example, a pure-strategy outcome that is not efficient can be destabilized with these types. Any inefficient mixed outcome can be similarly destabilized as long as its support does not coincide with the support of the efficient strategy, at least when we restrict attention to monomorphic populations. If the supports are identical, it is possible to find examples where the indifferent type is actually needed.

If $\sigma' \neq a^*$, then $\pi(a^*,a^*) > \pi(\sigma',a^*)$, since (a^*,a^*) is a strict Nash equilibrium. If $\sigma' = a^*$ on the other hand, $\pi(a^*,\sigma') = \pi(a^*,a^*) \geq \pi(\sigma'',\sigma'')$ for all $\sigma'' \in \Delta$, since (a^*,a^*) is efficient.

It remains to show that there exists a uniform ε that works for all σ', including those that are arbitrarily close to a^*. (The potential concern is that for σ' close to a^* the advantage of a^* gets smaller, leaving room for σ'' against itself to be better than a^* against σ'. But as σ' gets close to a^*, a^* against σ' is almost efficient. We now provide a detailed calculation.)

Let $\sigma' = qa^* + (1-q)\sigma$, where $\sigma(a^*) = 0$ and $q \in [0,1]$. Using (4), (5), and $\pi(a^*,a^*) \geq \pi(\sigma'',\sigma'')$,

$$\begin{aligned}\Pi_{\theta^*}(\cdot \mid \cdot) - \Pi_{\tilde{\theta}}(\cdot \mid \cdot) &= \left[(1-\varepsilon')\pi(a^*,a^*) + \varepsilon' q\pi(a^*,a^*) + \varepsilon'(1-q)\pi(a^*,\sigma)\right] \\ &\quad - \left[(1-\varepsilon')q\pi(a^*,a^*) + (1-\varepsilon')(1-q)\pi(\sigma,a^*) + \varepsilon'\pi(\sigma'',\sigma'')\right] \\ &\quad \geq (1-q)\left[\pi(a^*,a^*) - \pi(\sigma,a^*) - \varepsilon'(2\pi(a^*,a^*) - \pi(\sigma,a^*) - \pi(a^*,\sigma))\right].\end{aligned}$$

Since (a^*,a^*) is a strict Nash equilibrium, there exists $k > 0$ such that $\pi(a^*,a^*) - \pi(\sigma,a^*) \geq k > 0$ for all σ such that $\sigma(a^*) = 0$. On the other hand, $2\pi(a^*,a^*) - \pi(\sigma,a^*) - \pi(a^*,\sigma) \leq l$ for some l. Therefore, there exists an $\varepsilon \in (0,1)$ such that $\Pi_{\theta^*}(\cdot \mid \cdot) - \Pi_{\tilde{\theta}}(\cdot \mid \cdot) \geq (1-q)(k - \varepsilon' l) \geq 0$ for all $\varepsilon' \in (0,\varepsilon)$, $\sigma'' \in \Delta$, $q \in [0,1]$, and $\sigma \in \Delta$ such that $\sigma(a^*) = 0$, proving the stability of (a^*,a^*). ‖

Proposition 2. *If an outcome x^* is stable with configuration (μ^*, b^*), then*

$$\Pi_\theta(\mu^* \mid b^*) = \pi(b^*_\theta(\theta'), b^*_{\theta'}(\theta)) = \pi(\sigma^*,\sigma^*),$$

for all $\theta, \theta' \in C(\mu^*)$, where (σ^*,σ^*) is efficient.

Proof. Suppose that x^* is stable with (μ^*, b^*). Let $m(\theta) \in \underset{\theta' \in C(\mu^*)}{\arg\max}\, \pi(b^*_{\theta'}(\theta), b^*_\theta(\theta'))$, that is, $m(\theta)$ is the incumbent, which gets the highest (equilibrium) fitness against θ. Let θ^0, a type that is indifferent between all actions against any action of the opponent, be the mutant. Consider the focal post-entry equilibrium $\tilde{b}$, where $(\tilde{b}_{\theta^0}(\theta), \tilde{b}_\theta(\theta^0)) = (b^*_{m(\theta)}(\theta), b^*_\theta(m(\theta)))$ for all $\theta \in C(\mu^*)$, and $\tilde{b}_{\theta^0}(\theta^0) = \sigma^*$, where (σ^*,σ^*) is efficient. In other words, the mutant's fitness against an incumbent θ is at least as high as the fitness that any incumbent obtains against θ. It must be the case that

$$\pi(b^*_{\theta'}(\theta), b^*_\theta(\theta')) = \pi(b^*_{\theta'}(\theta), b^*_\theta(\theta'')) \;\forall \theta, \theta', \theta'' \in C(\mu^*), \tag{6}$$

that is, every incumbent type obtains the same fitness against a given incumbent θ, since if this were not the case, for small enough ε', θ^0 would obtain a strictly higher fitness in $\tilde{b}$ than at least one incumbent. Given that, for any θ, every type (including θ) obtains the same fitness against θ, we can, without loss of generality, choose $m(\theta) = \theta$. The average fitnesses of any $\theta \in C(\mu^*)$ and θ^0 in $\tilde{b}$ are, respectively,

$$\Pi_\theta((1-\varepsilon')\mu^* + \varepsilon'\theta^0 \mid \tilde{b}) = (1-\varepsilon') \sum_{\theta' \in C(\mu^*)} \pi(b^*_\theta(\theta'), b^*_{\theta'}(\theta))\mu^*(\theta') + \varepsilon'\pi(b^*_\theta(\theta), b^*_\theta(\theta))$$

and

$$\Pi_{\theta^0}((1-\varepsilon')\mu^* + \varepsilon'\theta^0 \mid \tilde{b}) = (1-\varepsilon') \sum_{\theta' \in C(\mu^*)} \pi(b^*_\theta(\theta'), b^*_{\theta'}(\theta))\mu^*(\theta') + \varepsilon'\pi(\sigma^*,\sigma^*).$$

Since x^* is stable and (σ^*,σ^*) is efficient, it follows that

$$\pi(b^*_\theta(\theta), b^*_\theta(\theta)) = \pi(\sigma^*,\sigma^*) \;\forall \theta \in C(\mu^*). \tag{7}$$

Combining (6) and (7), we conclude that every incumbent type must obtain the efficient fitness in each and every one of its interactions within the incumbent population, hence its average fitness must be efficient as well. ‖

Proposition 3.

(a) *If (A, A) is efficient, then it is stable.*
(b) *If (A, A) is not efficient, then the efficient (σ^*,σ^*) is stable iff $b = c > a$ (otherwise there is no stable outcome).*

Proof.

(a) $((A, A)$ 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 1 implies that (A, A) is stable.

(ii) $a \leq c$: In this case, which consists of PD and Hawk–Dove games, (A, A) is stable with a monomorphic population of $\mathcal{AB}_1$ (the type for which both A and B are best responses to A, and B is the unique best response to B) playing (A, A) in equilibrium. Let $\tilde{\theta} \in \Theta$ be an arbitrary mutant and $\tilde{b}$ be any focal post-entry equilibrium. Suppose that in $\tilde{b}$, $\mathcal{AB}_1$ and $\tilde{\theta}$ play σ and σ' respectively when they are matched, and $\tilde{\theta}$ plays σ'' when matched with itself. We will show that there exists an $\varepsilon \in (0,1)$ such that for all $\varepsilon' \in (0,\varepsilon)$ and $\sigma, \sigma', \sigma'' \in \Delta$, the incumbent will obtain (weakly) higher average fitness than the mutant with a population share of ε'. The average fitnesses of the incumbent and the mutant are, respectively,

$$\Pi_{\mathcal{AB}_1}((1-\varepsilon')\mathcal{AB}_1 + \varepsilon'\tilde{\theta} \mid \tilde{b}) = (1-\varepsilon')a + \varepsilon'\pi(\sigma,\sigma')$$

and

$$\Pi_{\tilde{\theta}}((1-\varepsilon')\mathcal{AB}_1 + \varepsilon'\tilde{\theta} \mid \tilde{b}) = (1-\varepsilon')\pi(\sigma',\sigma) + \varepsilon'\pi(\sigma'',\sigma'').$$

In any equilibrium against any type of opponent, $\mathcal{AB}_1$ plays A with positive probability only if the opponent plays A with probability 1, that is, $\sigma(A) > 0 \Rightarrow \sigma' = A$.
Consider $\sigma' = A$, in which case $\mathcal{AB}_1$ is indifferent between A and B, and consider all possible σ and σ''. The average fitnesses of the incumbent and the mutant are, respectively,

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

and

$$\Pi_{\tilde{\theta}}(\cdot \mid \cdot) = (1-\varepsilon')[qa + (1-q)b] + \varepsilon'\pi(\sigma'',\sigma'),$$

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

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

Also, efficiency of (A, A) implies that

$$qa + (1-q)c \geq a \geq \pi(\sigma'',\sigma'').$$

Hence, $\Pi_{\mathcal{AB}_1}(.) \geq \Pi_{\tilde{\theta}}(\cdot)$ irrespective of ε'.
Now consider $\sigma' \neq A$, which implies that $\sigma = B$. Considering again all possible equilibria, the average fitnesses are,

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

and

$$\Pi_{\tilde{\theta}}(\cdot \mid \cdot) = (1-\varepsilon')[qb + (1-q)d] + \varepsilon'\pi(\sigma'',\sigma^{''}),$$

where $q \in [0,1]$. We have $a \geq qb + (1-q)d$, since $a \geq b$ and $a \geq d$.
For q such that $a = qb + (1-q)d$, we have

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

and hence, $\Pi_{\mathcal{AB}_1}(\cdot \mid \cdot) \geq \Pi_{\tilde{\theta}}(\cdot \mid \cdot)$ irrespective of ε'.
Finally, for q for which $a > qb + (1-q)d$, we can find $\varepsilon \in (0,1)$ such that $\Pi_{\mathcal{AB}_1}(\cdot \mid \cdot) \geq \Pi_{\tilde{\theta}}(\cdot \mid \cdot)$ for all $\varepsilon' \in (0,\varepsilon)$, proving that (A, A) is stable.

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

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

and $\sigma^*(B) = 1-\alpha^*$. Note that $\pi(\sigma^*,\sigma^*) = d + \frac{(b+c-2d)^2}{4(b+c-a-d)}$. Since $\pi(\sigma^*,\sigma^*) > a$, σ^* is unique, and hence Proposition 2 implies that if an outcome is stable, then (σ^*,σ^*) must be played in each interaction within the stable distribution. So the support of any stable distribution must be a subset of $\{\mathcal{AB}_{\alpha^*}, \mathcal{BA}_{\alpha^*}, \theta^0\}$. We now consider four classes of 2×2 games in turn:

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

(ii) $a \geq c$ and $b \geq d$: If $c \geq b$, then (A, A) is efficient. So, let $b > c$. Suppose that (σ^*, σ^*) is stable with μ^*. We will show that $\mathcal{AB}_0$ can enter and obtain strictly higher average fitness against incumbents than the incumbents obtain against themselves in the focal post-entry equilibrium $\tilde{b}$ where $\mathcal{AB}_0$ mixes between A and B (playing A with probability α^*) and the incumbents play B whenever they are matched. We have, for any $\theta \in \mu^*$,

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

and

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

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 $d \geq b$: Since $a \geq d$, we have $c \geq b$. If $c = b$, then (A, A) is efficient. So, let $c > b$. Suppose that μ^* is a stable distribution. Let $\mathcal{AB}_1$ be the mutant, and consider the focal post-entry equilibrium 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 mutant's average fitness from its interactions with the incumbents is $\alpha^* a + (1-\alpha^*)c$, which is greater than $\pi(\sigma^*, \sigma^*)$, showing 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 argument in case (ii) (respectively, (iii)) applies, so there is no stable outcome. We will show that if $b = c$, then (σ^*, σ^*) is stable with a monomorphic population of $\mathcal{AB}_{\alpha^*}$ playing σ^*. Note that when $b = c$, (σ^*, σ^*) is a Nash equilibrium of G. Let $\tilde{\theta}$ be an arbitrary entrant and $\tilde{b}$ be any focal post-entry equilibrium. Suppose that in $\tilde{b}$, $\mathcal{AB}_{\alpha^*}$ and $\tilde{\theta}$ play σ and σ', respectively, when they observe each other, and $\tilde{\theta}$ plays σ'' when matched with itself. The average fitnesses are

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

and

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

Since the incumbent is $\mathcal{AB}_{\alpha^*}$, $\sigma' \in \{A, B, \sigma^*\}$. If $\sigma' = A$ (respectively, B), then $\sigma = A$ (respectively, B). In either case, since $\pi(\sigma^*, \sigma^*) > \pi(\sigma', \sigma) = \pi(\sigma, \sigma')$ and $\pi(\sigma^*, \sigma^*) \geq \pi(\sigma'', \sigma'')$ for all $\sigma'' \in \Delta$, $\Pi_{\mathcal{AB}_{\alpha^*}}(\cdot \mid \cdot) > \Pi_{\tilde{\theta}}(\cdot \mid \cdot)$ for all $\varepsilon' < \frac{1}{2}$.

For $\sigma' = \sigma^*$, we have

$$\pi(\sigma', \sigma) = \pi(\sigma^*, \sigma) = \pi(\sigma, \sigma^*) = \pi(\sigma^*, \sigma^*),$$

for all $\sigma \in \Delta$, where the second equality follows from the fact that in G both players obtain the same fitness by definition (since $b = c$), and the last equality follows from (σ^*, σ^*) being a Nash equilibrium of G. Independent of ε',

$$\Pi_{\mathcal{AB}_{\alpha^*}}(\cdot \mid \cdot) = \pi(\sigma^*, \sigma^*) \geq (1-\varepsilon')\pi(\sigma^*, \sigma^*) + \varepsilon'\pi(\sigma'', \sigma'') = \Pi_{\tilde{\theta}}(\cdot \mid \cdot).$$

Therefore, $\Pi_{\mathcal{AB}_{\alpha^*}}(\cdot \mid \cdot) \geq \Pi_{\tilde{\theta}}(\cdot \mid \cdot)$ for all possible $\tilde{\theta}$, $\tilde{b}$, and $\varepsilon' < \frac{1}{2}$, proving the claim. ‖

Proposition 4. *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.*

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

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

2. *For games in which (A, A) is not efficient, if $b = c > a$, then $M(G) = \{\mathcal{AB}_{\alpha^*}\}$, where σ^* is efficient and $\alpha^* = \sigma^*(A)$.*

Proof.

(1) Proposition 2 implies that (A, A) must be played in each match within a stable distribution. So $M(G)$ must be a subset of $\{\mathcal{AA}, \mathcal{AB}_\alpha, \mathcal{BA}_1, \theta^0\}$, $\alpha \in [0, 1]$.

(a) In this case, fitness of (A, A) is greater than fitness of any other strategy profile. Therefore, any type for which A is a best response to itself can be in a stable distribution.

(b) If a stable distribution contains $\mathcal{BA}_1$ or θ^0, an $\mathcal{AA}$ type can enter. In matches against the entrant, both of these incumbent types are willing to play B, thus there is a focal post-entry equilibrium in which they do so. This gives the entrant $b > a$ in matches against $\mathcal{BA}_1$ and θ^0 (and a against other incumbents, if there are any), thereby a higher average fitness than $\mathcal{BA}_1$ and θ^0 independent of its population share. Suppose that a stable distribution contains $\mathcal{AB}_\alpha$, where $\alpha \geq \frac{a-d}{b-d}$. When $\mathcal{AB}_0$ enters, there is a focal post-entry equilibrium where it obtains $\alpha b + (1-\alpha)d \geq a$ when matched with $\mathcal{AB}_\alpha$ (playing A with probability α while $\mathcal{AB}_\alpha$ is playing B), and a when matched with itself or any incumbent other than $\mathcal{AB}_\alpha$, if there are any. Note that $\mathcal{AB}_\alpha$ is obtaining $\alpha c + (1-\alpha)d < a$ in matches against the entrant. Therefore, the entrant's average fitness is strictly higher than that of $\mathcal{AB}_\alpha$, independent of its population share.

Next we show that a distribution is stable if its support is a subset of $\{\mathcal{AA}, \mathcal{AB}_\alpha\}$, where $\alpha < \frac{a-d}{b-d}$. Any entrant, whenever matched with $\mathcal{AA}$, obtains a convex combination of a and c, which is (weakly) less than a, while $\mathcal{AA}$ is obtaining (weakly) more than a (since $b > a$). Therefore, $\mathcal{AA}$ obtains (weakly) higher average fitness than any entrant in every focal post-entry equilibrium, independent of the entrant's population share.

Any entrant, whenever matched with $\mathcal{AB}_\alpha$, $\alpha < \frac{a-d}{b-d}$, obtains at most a. (To see this, note that if the entrant plays A with a probability strictly higher than α, $\mathcal{AB}_\alpha$ will play A, against which the entrant can obtain at most a, since $a > c$. If the entrant plays A with probability α or less the highest fitness it can obtain is $\alpha b + (1-\alpha)d < a$, since $\alpha < \frac{a-d}{b-d}$.) Moreover, whenever the entrant obtains a, so does $\mathcal{AB}_\alpha$. Therefore, any $\mathcal{AB}_\alpha$, $\alpha < \frac{a-d}{b-d}$, obtains a (weakly) higher average fitness than any entrant in any focal post-entry equilibrium, if the entrant's share is small enough. Notice that a uniform barrier ε can be found, since $\alpha b + (1-\alpha)d < a$ for all α such that $\mathcal{AB}_\alpha$ is in the distribution.

(c) We showed in the proof of Proposition 3 that $\mathcal{AB}_1$ is stable. Suppose that a stable distribution contains $\mathcal{AA}$, $\mathcal{BA}_1$, or θ^0. $\mathcal{AB}_1$ can enter. There is a focal post-entry equilibrium in which $\mathcal{AB}_1$ plays B and $\mathcal{AA}$ ($\mathcal{BA}_1$ or θ^0) plays A whenever they are matched, obtaining $c > a$ in these matches, and obtaining a against other incumbents (by playing A). Suppose now that a stable distribution contains $\mathcal{AB}_\alpha$, where $\alpha \in [0, 1)$. Again consider $\mathcal{AB}_1$ as the entrant. There is a focal post-entry equilibrium in which $\mathcal{AB}_1$ mixes between A and B, and $\mathcal{AB}_\alpha$ 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.

(2) Proposition 2 implies that, if an outcome is stable, then (σ^*, σ^*) must be played in each match within the stable distribution. So $M(G)$ must be a subset of $\{\mathcal{AB}_{\alpha^*}, \mathcal{BA}_{\alpha^*}, \theta^0\}$. We showed in the proof of Proposition 3 that $\mathcal{AB}_{\alpha^*}$ is stable. Suppose that a stable distribution contains $\mathcal{BA}_{\alpha^*}$ or θ^0. $\mathcal{AA}$ enters and obtains $b > \pi(\sigma^*, \sigma^*)$ in matches with $\mathcal{BA}_{\alpha^*}$ and θ^0. Hence $\mathcal{AB}_{\alpha^*}$ is the only stable distribution. ‖

Proposition 5.

(a) *(σ, σ) is stable only if it is a Nash equilibrium of G.*

(b) *If (a_i, a_i) is a strict Nash equilibrium of G, then it is stable.*

Proof.

(a) The proof is by contradiction. Suppose that (σ, σ) is stable with the configuration (μ^*, b^*), but (σ, σ) is not a Nash equilibrium of G.

Since (σ, σ) is not a Nash equilibrium, there exists $a_i \in A$ such that

$$\pi(a_i, (1-\varepsilon')\sigma + \varepsilon' a_i) > \pi(\sigma, (1-\varepsilon')\sigma + \varepsilon' a_i) \qquad (8)$$

for small enough ε'. Consider the mutant $\tilde{\theta}$ for which a_i is strictly dominant with a population share of ε'. Let $\tilde{\mu}$ denote the post-entry preference distribution.

First, suppose that there exists a focal post-entry equilibrium. In any focal equilibrium the incumbents' aggregate play is given by σ, so the average fitness of $\tilde{\theta}$ is given by the L.H.S. of (8). Similarly, the average fitness (over all) of incumbents is given by the R.H.S. of (8). Therefore, there exists an incumbent whose average fitness is strictly less than that of $\tilde{\theta}$ whenever ε' is small enough.

Next, suppose that the set of focal equilibria is empty. Note that by the continuity of $\pi(\cdot)$ the same inequality holds as long as the aggregate play is close enough to σ. But this implies that (σ, σ) is destabilized because in any near enough post-entry equilibrium (*i.e.* for equilibria in $B_p^\delta(\tilde{\mu})$ for δ small enough) the entrant outperforms some incumbent type (for any small enough ε').

(b) If (a_i, a_i) is a strict equilibrium, then there exists an $\varepsilon \in (0,1)$ such that

$$\pi(a_i, (1-\varepsilon')a_i + \varepsilon'\sigma) \geq \pi(\sigma, (1-\varepsilon')a_i + \varepsilon'\sigma), \tag{9}$$

for all $\sigma \in \Delta$ and $\varepsilon' \in (0, \varepsilon)$. Consider a monomorphic population of type θ^* that is equivalent (modulus an affine transformation) to the fitness function π, playing the equilibrium b^*, where $b^*_{\theta^*}(\emptyset) = a_i$. Consider any mutant $\tilde{\theta}$ and any $\varepsilon' \in (0, \varepsilon)$. It follows from (9) that the set of focal post-entry equilibria is non-empty, that is, there exists $\tilde{b} \in B_0((1-\varepsilon')\theta^* + \varepsilon'\tilde{\theta})$ such that $\tilde{b}_{\theta^*}(\emptyset) = b^*_{\theta^*}(\emptyset) = a_i$. Moreover, for any focal post-entry equilibrium $\tilde{b}$

$$\begin{aligned}\Pi_{\theta^*}((1-\varepsilon')\theta^* + \varepsilon'\tilde{\theta} \mid \tilde{b}) &= \pi(a_i, (1-\varepsilon')a_i + \varepsilon'\tilde{b}_{\tilde{\theta}}(\emptyset)) \\ &\geq \pi(\tilde{b}_{\tilde{\theta}}(\emptyset), (1-\varepsilon')a_i + \varepsilon'\tilde{b}_{\tilde{\theta}}(\emptyset)) \\ &= \Pi_{\tilde{\theta}}((1-\varepsilon')\theta^* + \varepsilon'\tilde{\theta} \mid \tilde{b}),\end{aligned}$$

where the inequality follows from (9). Therefore, (a_i, a_i) is stable with the configuration (θ^*, b^*). ||

Proposition 6. *If* (a^*, a^*) *is both efficient and a strict Nash equilibrium of* G*, then it is stable for all* $p \in (0,1)$.

Proof. Fix $p \in (0,1)$. Let (a^*, a^*) be a strict Nash equilibrium of G that is efficient. Consider a monomorphic population consisting of θ^* for which a^* is strictly dominant. The outcome of the unique Bayesian–Nash equilibrium is (a^*, a^*). Moreover, after any mutant's entry, in all post-entry equilibria the incumbent θ^* will always play a^* (regardless of what is observed), since a^* is strictly dominant for θ^*. It follows that all post-entry equilibria will be focal, and so the set of focal equilibria is non-empty. Since the incumbent is always playing a^*, and (a^*, a^*) is a strict Nash equilibrium of G, mutants that do not play a^* when they are matched with θ^* (both when they observe θ^* and when they do not observe anything) will obtain strictly less fitness than incumbents if their population share is sufficiently small. But for mutants that play a^* whenever they are matched with θ^*, the incumbent's average fitness is given by $\pi(a^*, a^*)$, and since no mutant can obtain a fitness strictly higher than this when it is matched with itself (since (a^*, a^*) is efficient), it cannot obtain a strictly higher average fitness either. Finding a uniform barrier ε for all possible mutants is straightforward, since (a^*, a^*) is a strict Nash equilibrium. (See, *e.g.* the proof of Proposition1.) We conclude that (a^*, a^*) is stable. ||

Proposition 7. *If* (a_i, a_i) *is not efficient, then there exists a* $\overline{p} \in (0,1)$ *such that it is not stable for any* $p \in (\overline{p}, 1)$.

Proof. Suppose that (a_i, a_i) is stable with μ^* when the probability of observability is p. Let $\tilde{\theta}$ be the "coordination type" mutant for which a_i is a strict best response to itself and σ^*, the efficient strategy, is a best response to $p\sigma^* + (1-p)a_i$. When the share of the mutant ε' is small enough there exists a focal post-entry equilibrium in which incumbents play a_i regardless of what they observe, and $\tilde{\theta}$ plays a_i if it observes any incumbent or if it observes nothing, and plays σ^* if it observes $\tilde{\theta}$. In this equilibrium,

$$\Pi_\theta(\cdot \mid \cdot) = \pi(a_i, a_i), \ \forall \theta \in C(\mu^*)$$

and

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

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

Proposition 8. *If* (a_i, a_i) *is not a Nash equilibrium of* G*, then there exists a* $\overline{p} \in (0,1)$ *such that it is not stable for any* $p \in (0, \overline{p})$.

Proof. Suppose that (a_i, a_i) is stable with (μ^*, b^*). If (a_i, a_i) is not a Nash equilibrium, then there exists $a_j \in A$ such that $\pi(a_j, a_i) > \pi(a_i, a_i)$. Consider the mutant $\tilde{\theta}$ for which a_j is strictly dominant, and let $\tilde{\mu} = (1-\varepsilon')\mu^* + \varepsilon'\tilde{\theta}$. For any focal post-entry equilibrium $\tilde{b} \in B_p(\tilde{\mu} \mid b^*)$

$$\Pi_\theta(\tilde{\mu} \mid \tilde{b}) = (1-\varepsilon')\pi(a_i, a_i) + \varepsilon'[p\pi(\tilde{b}_\theta(\tilde{\theta}), a_j) + (1-p)\pi(a_i, a_j)], \ \ \forall \theta \in C(\mu^*)$$

and

$$\Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b}) = (1-\varepsilon')\left[p \sum_{\theta \in C(\mu^*)} \pi(a_j, \tilde{b}_\theta(\tilde{\theta}))\mu^*(\theta) + (1-p)\pi(a_j, a_i)\right] + \varepsilon'\pi(a_j, a_j).$$

Let a_* be such that $\pi(a_j, a_*) \leq \pi(a_j, a)$ $\forall a \in A$. Since $\pi(a_j, a_i) > \pi(a_i, a_i)$, there exists a $\overline{p} \in (0,1)$ such that for all $p \in (0, \overline{p})$

$$p\pi(a_j, a_*) + (1-p)\pi(a_j, a_i) > \pi(a_i, a_i), \tag{10}$$

so that, for small enough ε', $\Pi_{\tilde{\theta}}(\tilde{\mu} \mid \tilde{b}) > \Pi_{\theta}(\tilde{\mu} \mid \tilde{b})$ for all $\tilde{b} \in B_p(\tilde{\mu} \mid b^*)$ and for all $\theta \in C(\mu^*)$. This shows that (a_i, a_i) is destabilized when there exists a focal equilibrium. The continuity of $\pi(.)$ and (10) ensure that there exists $\theta \in C(\mu^*)$ such that the same conclusion holds for all $\tilde{b} \in B_p(\tilde{\mu})$ with outcomes close enough to (a_i, a_i). Therefore even in cases where there is no focal equilibrium, (a_i, a_i) is again destabilized. ||

Acknowledgements. We thank seminar participants at the Econometric Society World Congress (Seattle, 2000), Conference on Economic Design (Istanbul, 2000), Koc, Michigan, Northwestern, Queen's, Rochester, Simon Fraser, Southampton, UBC, two anonymous referees, and the Managing Editor. Financial support from NSF (Dekel) and SSHRC (Yilankaya) is gratefully acknowledged.

REFERENCES

BECKER, G. S. (1976), "Altruism, Egoism, and Genetic Fitness", *Journal of Economic Literature*, **14**, 817–826.

BHASKAR, V. (1998), "Noisy Communication and the Evolution of Cooperation", *Journal of Economic Theory*, **82**, 110–131.

BINMORE, K. and SAMUELSON, L. (1997), "Muddling Through: Noisy Equilibrium Selection", *Journal of Economic Theory*, **74**, 235–265.

CARLSSON, H. and VAN DAMME, E. (1993), "Global Games and Equilibrium Selection", *Econometrica*, **61**, 989–1018.

DEKEL, E., ELY, J. C. and YILANKAYA, O. (2004), "Evolution of Preferences" (http://www.econ.ubc.ca/discpapers/dp0412.pdf).

ELLISON, G. (1993), "Learning, Local Interaction, and Coordination", *Econometrica*, **61**, 1047–1072.

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

ELY, J. C. and YILANKAYA, O. (2001), "Nash Equilibrium and Evolution of Preferences," *Journal of Economic Theory*, **97**, 255–272.

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.

GÜTH, W. (1995), "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives", *International Journal of Game Theory*, **24**, 323–344.

GÜTH, W. and YAARI, M. (1992), "Explaining Reciprocal Behavior in a Simple Strategic Game", in *Explaining Process and Change-Approaches to Evolutionary Economics* (Ann Arbor: University of Michigan Press) 23–24.

HARSANYI, J. C. and SELTEN, R. (1988), *A General Theory of Equilibrium Selection in Games* (Cambridge, MA and London: MIT Press).

HEIFETZ, A., SHANNON, C. and SPIEGEL, Y. (2004), "What to Maximize if You Must" (Mimeo, UC Berkeley, Department of Economics, http://elsa.berkeley.edu/users/cshannon/wp/what.pdf).

HIRSHLEIFER, J. (1977), "Economics from a Biological Viewpoint", *The Journal of Law and Economics*, **20**, 1–52.

KANDORI, M., MAILATH, G. and ROB, R. (1993), "Learning, Mutation, and Long-Run Equilibria in Games", *Econometrica*, **61**, 29–56.

KIM, Y.-G. and SOBEL, J. (1995), "An Evolutionary Approach to Pre-Play Communication", *Econometrica*, **63**, 1181–1194.

MATSUI, A. (1991), "Cheap Talk and Coordination in Society", *Journal of Economic Theory*, **54**, 245–258.

OK, E. and VEGA-REDONDO, F. (2001), "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario", *Journal of Economic Theory*, **97**, 231–254.

ROBSON, A. J. (1990), "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake", *Journal of Theoretical Biology*, **144**, 379–396.

ROBSON, A. J. and VEGA-REDONDO, F. (1996), "Efficient Equilibrium Selection in Evolutionary Games with Random Matching", *Journal of Economic Theory*, **70**, 65–92.

RUBIN, P. and PAUL, C. (1979), "An Evolutionary Model of Taste for Risk", *Economic Inquiry*, **42**, 585–596.

SAMUELSON, L. (2001), "Introduction to Evolution of Preferences", *Journal of Economic Theory*, **97**, 225–230.

SETHI, R. and SOMANATHAN, E. (2001), "Preference Evolution and Reciprocity", *Journal of Economic Theory*, **97**, 273–297.

VON WIDEKIND, S. (2004), "Evolution of Non-Expected Utility Preferences" (IMW Working Paper #370, Bielefeld University).

WISEMAN, T. and YILANKAYA, O. (2001), "Cooperation, Secret Handshakes, and Imitation in the Prisoner's Dilemma", *Games and Economic Behavior*, **37**, 216–242.

YILANKAYA, O. (1999), "On the Evolution of Preferences" (Ph.D. thesis, Northwestern University, Evanston).

YOUNG, P. (1993), "The Evolution of Conventions", *Econometrica*, **61**, 57–84.