Economics Letters 42 (1993) 301–312
0165-1765/93/$06.00 © 1993 Elsevier Science Publishers B.V. All rights reserved

Evolutionary stability and efficiency

Joel Sobel*

Department of Economics, D-008, University of California–San Diego, La Jolla, CA 92093, USA

Abstract

'One of the advantages of moral philosophy over game theory is that moralists give sensible advice to moral agents while game theory can give stupid advice to game theorists' – Ian Hacking.

1. Introduction

Theory can lag behind common sense. In some games there appears to be a clear prediction of behavior, but that prediction does not follow unambiguously from a direct application of common solution concepts. In this paper I discuss how evolutionary arguments can lead to sensible predictions.

In games where the players have common interests, one expects that they will be able to coordinate on an efficient outcome. Standard solution concepts do not yield this conclusion in games with multiple, strict, Pareto-ranked equilibria. I will describe arguments based on evolutionary stability that do guarantee efficiency in these games.

The central message is that evolutionary pressures tend to destabilize inefficient outcomes. Common-interest games are central to the study because only in these games can I identify a small set of outcomes as stable. In all of my applications, the essential argument is the same. If the population ever reaches a state in which players obtain inefficient payoffs, then evolutionary pressures introduce a strategy that works as well as existing strategies against the current population, and has the flexibility to achieve an efficient payoff when playing with an appropriate partner.

I will discuss three types of games. In section 4 I add a single round of pre-play communication to a given finite, two-player game. When I add payoff irrelevant pre-play communication to the game, these equilibria are no longer strict, and evolutionary pressures force the population away from inefficient patterns of behavior. In common-interest games, these pressures lead the population to the efficient outcome, and once players coordinate on an efficient outcome, they do not ever obtain less.

In section 5 I discuss cheap-talk games with private information. In these games there always is a Nash equilibrium in which the uninformed player ignores what the informed player says to him. Inefficient outcomes fail to be evolutionarily stable, however. As in complete-information games,

* I prepared this paper for the session 'Recent Advances in Evolutionary Economics' held at the January 1993 meetings of the American Economic Association. Special thanks to Eric Maskin, who presented the paper in my absence. I also thank Andreas Blume, Antonio Cabrales, Vincent Crawford, Eddie Dekel-Tabak, Yong-Gwan Kim, Matthew Rabin, and Larry Samuelson for many useful conversations about evolution and communication in games. They are not responsible for the shortcomings of this paper. I am grateful to the NSF for supporting this research.

intuition suggests that players will be able to obtain the best outcome when there are no conflicts of interest. The evolutionary approach validates this intuition.

The same evolutionary arguments that destabilize inefficient outcomes in cheap-talk games also destabilize these outcomes in repeated games. In the first games invading strategies could use costless messages to signal an intention to cooperate, and ultimately use this flexibility to move the population away from any inefficient outcome. In repeated games with no discounting, what players do in the first portion of the game has no impact on the average payoffs. Hence, by playing a sequence of actions that no other member of the population plays, an invading strategy can identify itself in finite time and be able to achieve a higher payoff.

I present these results using a new, non-equilibrium, set-valued, evolutionary stability concept. I explain the concept in section 2 and its central implication in section 3. The advantage of my formulation is that it makes the central results clear. In the process it demonstrates that without the common-interest assumption, evolutionary arguments do not help to narrow the set of predictions in the games that I discuss.

The qualitative results that I describe are not new. They have been obtained (in slightly different forms using different stability notions) by others. I describe the implications of different modeling assumptions in section 7.

2. The framework

I consider games that have two identified roles. There is a finite population of $2N$ players, with N individuals assigned to each role. These players use pure strategies. Players are randomly matched to play the game, and do so repeatedly and anonymously. On rare occasions, one member of the population changes the strategy it uses. I will not provide a dynamic specification of this process. One may think that at regular intervals someone who chooses an optimal (or at least better) response to the existing population profile replaces the member of the population that is doing the worst. The stability condition that I describe below is meant to capture the idea that bad strategies die out, new strategies have the potential to enter the population, and that successful strategies may be adopted by more members of the population. The behavior that I describe is extremely naive.

I will denote the set of pure strategies of players assigned to role one by S; T will denote the set of strategies available to players in the second role. A population strategy profile is a list $(s_1, s_2, \ldots, s_N; t_1, t_2, \ldots, t_N)$ of (not necessarily distinct) pure strategies for each of the $2N$ players in the population with $s_i \in S$ and $t_i \in T$ for $i = 1, \ldots, N$; the strategy profile $(s_1, s_2, \ldots, s_N; t_1, t_2, \ldots, t_N)$ is homogeneous if $s_i = s_j$ and $t_i = t_j$ for all i and j. Denote by u (for player 1) and v (for player 2) the payoff functions for the two players in the game. Given a strategy profile, the payoff functions induce the population payoffs for each player in the population, $U_i(s_1, s_2, \ldots, s_N; t_1, t_2, \ldots, t_N) = \Sigma_j\, u(s_i, t_j)/N$ and $V_j(s_1, s_2, \ldots, s_N; t_1, t_2, \ldots, t_N) = \Sigma_i\, u(s_i, t_j)/N$.

The strategy profile $\theta' = (s_1, \ldots, s_i', \ldots, s_N; t_1, \ldots, t_N)$ [or $(s_1, \ldots, s_N; t_1, \ldots, t_j', \ldots, t_N)$] can *replace* the strategy profile $\theta = (s_1, s_2, \ldots, s_N; t_1, t_2, \ldots, t_N)$ if $U_i(\theta) \le U_i(\theta')$ [or $V_j(\theta) \le V_i(\theta')$]. In order for one strategy profile to be able to replace another, it must differ from the original strategy in the behavior of only one player and the player that changes weakly gains by doing so.

Definition. A set Θ of strategy profiles is a *non-equilibrium evolutionarily stable* (NES) set if it is a minimal non-empty set with respect to the property:
(R) if $\theta \in \Theta$ and θ' can replace θ, then $\theta' \in \Theta$.

This definition suggests how the population might evolve over time. If a given strategy profile is an element of a stable set, then so is a profile of strategies obtained by changing one individual's strategy in a way that makes that individual weakly better off (in the short run). Condition (R) describes how one strategy can enter the population by replacing another strategy. If a NES set consists of a single strategy profile, then the population strategy is homogeneous and a strict Nash equilibrium. In the games with common interests that I discuss in this paper, the NES set consists of Nash equilibria that rise to the same outcome. In general, however, there may be non-equilibrium elements in the set.

The following argument establishes the existence of NES sets for finite games. For each strategy profile θ let $A(\theta)$ be the set of all strategy profiles attainable from θ after a sequence of replacements [that is, $\theta' \in A(\theta)$ if there exists a sequence of strategy profiles θ_i, $i = 1, \ldots, k$, such that $\theta = \theta_1$, $\theta' = \theta_k$, and, for each $i > 1$, θ_i can replace θ_{i-1}]. The NES sets are the minimal subsets of strategy profiles Θ for which $A(\Theta) = \Theta$. Since there exists a set X such that $A(X) = X$ (namely the set of all strategy profiles), at least one minimal subset satisfying $A(X) = X$ must exist. I can extend the argument to infinite strategy spaces (as in section 6) using Zorn's Lemma.

The definition of non-equilibrium evolutionarily stable sets differs from other proposed evolutionary stability conditions in several ways. It does not require every individual assigned to play the same role to use the same strategy. Allowing this type of polymorphism permits the population to exhibit strategy frequencies that would otherwise require individual randomization. The solution concept is set valued. The set-valued solution concept permits behavior to drift off the equilibrium path. If one strategy profile is an element of a NES set, then any strategy profile obtained by modifying a strategy at unreached information sets is also an element of the NES set. By including all specifications of behavior at unreached information sets, NES avoids the trivial non-existence problems associated with the concept of ESS.[1] It is the ability of the population to drift to different reponses out of equilibrium that enables me to give simple arguments that show NES sets must contain efficient payoffs. Another prominent way to handle the problem is to assume that the players tremble when they implement their strategies so that there are no unreached information sets [as in Selten (1983) and Fudenberg and Maskin (1990, 1991)].

Since the static evolutionary stability condition is meant to summarize a dynamic process, it should describe outcomes that arise when the dynamic does not necessarily reach a rest point. Ignoring an outcome because it fails to satisfy a static evolutionary equilibrium condition could lead to significant problems in interpretation because evolutionary equilibria need not exist; it is also a problem when there is a unique strategy that satisfies static evolutionary equilibrium conditions. The unwary may treat uniqueness as a sign that there is no other behavior consistent with the dynamic process implicitly being modeled, when in fact the evolutionary process could approach a limit cycle (or a more complicated set of limit points) rather than a unique strategy. Elements of a NES set are to be thought of as limit points of evolutionary processes. By not providing an explicit dynamic process, I leave doubts about whether or not there actually is a dynamic process associated with these limit sets. Furthermore, investigating an explicit process might provide more detailed information about limiting behavior, for example it may permit one to compute a probability distribution over the strategies in the limit.

I assume that asymmetric roles exist. The traditional context for evolutionary game theory has been symmetric games played by a single population. Concentrating on a game in which roles are already distinguished is consistent with the standard setting of economic models. In symmetric

[1] Evolutionarily stable strategies will not exist in games in which there are non-trivial unreached information sets because any strategy that plays as the population strategy does on the equilibrium path but differently off the path is able to enter. For a further discussion of this problem see Selten (1983) or Van Damme (1987, ch. 9).

models a new strategy can enter the population if it is an optimal response to the existing population and if the new strategy responds to itself at least as well as the existing population responds to it. The second condition does not arise in asymmetric contexts, hence the replacement condition is easier to satisfy than the corresponding condition in single-population games. Selten (1983) has noted that, because of this difference, ESSs must be strict Nash equilibria in asymmetric games.

There are possible variations in the definition. First, I could have presented the stability conditions under the assumption that an individual enter the population rather than holding the population size fixed. Second, there are variations on the replacement condition. For example, a strategy can remain in the population if it does better than the average strategy currently being used. This condition is an analog to the replicator dynamic of evolutionary biology. Alternatively, I could make the entry condition more restrictive, for example by requiring that the new strategy does better than the best performing strategy currently available. Changing the entry condition in either of these ways will not change the results of this paper. Still another modification of the entry condition would be to require that the new strategy continue to perform well after other changes in the population's strategy profile take place. I discuss versions of this assumption that do not lead to changes in the results in section 7.

3. Games of common interest

Associated with any game is the feasible set of payoffs, $F = \{(u(s,t), v(s,t)) \colon (s,t) \in S \times T\}$. I say that the game has *common interests* if F has a unique weakly Pareto-efficient point. Denote this point (when it exists) by (u^*, v^*). Proposition 1 is a simple implication of the definition of NES sets.

Proposition 1. In any game with common interests, there exists a NES set in which each individual obtains its highest payoff in each element of the NES set.

Proposition 1 states that in common-interest games the efficient payoff is evolutionarily stable.

Proof. Let Θ be the set of all strategy profiles that give rise to efficient payoffs. That is, $(s_1, \ldots, s_N;\ t_1, \ldots, t_N) \in \Theta$ if and only if $(u(s_i, t_j), v(s_i, t_j)) = (u^*, v^*)$ for all i and j. Θ is non-empty and no strategy outside of Θ could satisfy (R). Since the intersection of sets that satisfy (R) also satisfy (R), there exists a NES set that is contained in Θ.

Taken by itself, Proposition 1 contains the routine observation that the efficient payoff in common-interest games is stable. My interest is in demonstrating that this is the only stable payoff for a class of common-interest games. The result is not general; I must make some assumption on the structure of the game. In the simple 2×2 coordination game of Fig. 1 with two strict, Pareto-ranked equilibria, it is apparent that a homogeneous population that coordinates on the (DOWN, RIGHT) equilibrium is (taken as a singleton) a NES set. The problem in this example is that there is no way for an individual to play to make the efficient outcome possible without being punished by a population that coordinates on the inefficient strict equilibrium. Enlarging the game to include repetitions or pre-play communication permits a drift to a situation in which a player who is willing to move towards the efficient equilibrium can signal this intention without hurting himself against the population as a whole.

In the next sections I will describe situations where evolutionary pressures do lead to efficiency.

The basic argument will be to demonstrate that in all of the games that I examine (either with free talk or infinite repetition), every NES set contains a strategy profile that leads to efficient payoffs, whether or not the players have common interests. Combined with Proposition 1, this result guarantees that in common-interest games the only stable payoffs are efficient.

	LEFT	RIGHT
UP	2,2	0,0
DOWN	0,0	1,1

Fig. 1

4. Pre-play communication

Begin with a finite two-player game, called the underlying game. Add to that game a single round in which both players simultaneously select a message from a given language and then play the underlying game based on what has been said. Assume that the set of available messages contains more than N (the number of pairs of players in the population) elements. Communication is free; payoffs depend directly on only the second-round actions. Strategies for the communication game consist of what to say in the first round, followed by a rule that determines how to act in the underling game as a function of all possible first-round statements. Any Nash equilibrium in the underlying game will induce a Nash equilibrium in the communication game in which players send arbitrary signals in the first round, and then use the same equilibrium strategies from the underlying game no matter what was said.

The ability to communicate freely prior to the underlying game creates pressure that destabilizes any inefficient payoff. In my formulation the pressure is not biased: the unique NES set of a game with pre-play communication always contains strategies that attain every efficient payoff of the underlying game.

Proposition 2. For any game with pre-play communication, there is a unique NES set. For any efficient payoff in the underlying game, there exists a homogeneous strategy profile contained in the NES set that yields the efficient payoff.

A corollary of Propositions 1 and 2 is that in common-interest games, there is a unique NES set containing only strategy profiles that lead to efficient payoffs.

Proof. The proof makes repeated use of the fact that if θ is an element of a NES set Θ, and θ' is obtained from θ by making changes in how players respond to unsent messages, then $\theta' \in \Theta$. This property, which I call drift, follows from condition (R).

Let (a_1^*, a_2^*) be strategies in the underlying game that give rise to the highest feasible payoff to player one; denote by (w_1^*, w_2^*) the corresponding payoff. Let θ be an arbitrary element of a NES set Θ. Since the cardinality of the message space is greater than the population size, there exist messages m_1 and m_2 such that no player assigned to role i in the population uses message m_i in θ. By drift, the strategy profile in which each individual in the role of player two behaves exactly as in θ unless m_1 is sent, and responds to m_1 with a_2^* must be an element of Θ. Consequently, it follows from repeated applications of (R) (once for each individual in role one) that the strategy profile in which each role one individual signals m_1 and then plays a_1^* (and the role two players play as in θ') is an element of Θ. Call this strategy θ''. Moving from θ' to θ'' requires each individual assigned to role one to change his/her strategy on the equilibrium path. Since the change causes the role one individual to obtain his/her highest feasible payoff, condition (R) holds. Continuing in this way, I can conclude that the profile in which role two players always use m_2 is an element of Θ. Call this strategy profile θ^*. Without loss of generality (by invoking the drift property), it follows that a homogeneous strategy profile that uses only the messages m_1 and m_2 and that attains the efficient payoff preferred by the role one individuals must be in Θ.

I have shown that every NES set contains a homogeneous strategy profile in which individuals in role one obtain their highest feasible payoff. From this result it is straightforward to show that any homogeneous strategy profile in which individuals in role one obtain their highest feasible payoff is an element of the NES set. Hence the NES set is unique. Furthermore, given that $\theta^* \in \Theta$, only a small modification of the argument is needed to demonstrate that any feasible payoff of the underlying game in which player two obtains more than w_2^* must also be attained by a homogeneous element of Θ.[2]

By limiting attention to homogeneous strategy profiles I am only able to describe a subset of the possible NES payoffs. It is apparent nonetheless that the NES set will generally be large. NES outcomes need not be equilibrium outcomes of the communication game nor individually rational. The reason that the NES set is so large is that the entry condition is so easy to satisfy. Efficiency results similar to Proposition 2 hold with more restrictive entry conditions.

Unless the underlying game has common interests, there will necessarily be multiple NES payoffs. It follows directly from Proposition 1, however, that in a game with common interests, if the population ever reaches an efficient strategy profile (one in which every member of the population obtains his/her greatest feasible payoff), then it will never again reach an inefficient outcome.

5. Cheap-talk games with incomplete information

In this section I briefly describe how the results in section 4 apply to another class of games with communication. In simple Sender–Receiver games, nature first informs the Sender of her type (selected from a finite set of types), after which she chooses a message to send to the otherwise uninformed Receiver. After the Receiver hears the Sender's message, he takes a payoff relevant action. I maintain the assumption that talk is free in this game; the Sender's message does not enter either player's payoff function directly. I also assume that the set of possible messages is large enough so that there always is one unused message for each type in any strategy profile.[3]

The ability to communicate in Sender–Receiver games expands the set of equilibria. If there were no communication, then the Receiver could not condition his action on the Sender's information. With communication, such conditioning is feasible, although, as in the complete information games of section 4, it is always a part of an equilibrium for the Receiver to make his action choice independent of the Sender's message.

For Sender–Receiver games, versions of the efficiency and existence results of section 4 hold without augmenting the strategy sets to include additional opportunities to communicate.

[2] The modification is to first allow role one individuals' response to out-of-equilibrium messages to drift to a response that allows the second player to obtain the target payoff by choosing the appropriate response.

[3] If there are M types of Sender, then there should be at least $(N + 1)M$ messages.

Proposition 3. Every Sender–Receiver game has a unique NES set. There exists a homogeneous strategy profile in the NES set in which each type of Sender attains her highest feasible payoff.

Proof. By assumption, for every strategy profile θ there exists at least one unused message for each type of Sender. If θ is an element of a NES set Θ, then so is θ', the strategy profile obtained from θ by letting, for each type of Sender, the response that players in the role of Receiver make following an unsent message drift to an action that leads to the highest payoff for that type. From θ', repeated application of (R) guarantees that the NES set contains one homogeneous strategy profile in which each Sender type obtains its highest payoff. Given that such a strategy is an element of Θ it is straightforward to show that homogeneous strategy profiles that give rise to the same payoffs using different messages must also be in Θ. Uniqueness of the NES set follows.

Propositions 1 and 3 combine to demonstrate that the only NES payoff in a common-interest Sender–Receiver game is the efficient payoff.

6. Infinitely repeated games

Now consider infinitely repeated games in which players do not discount. Players seek to maximize the limit of the average of their stage-game payoffs. Assume that the stage game is non-trivial in the sense that each player has at least two pure strategies. Exploiting the finiteness of the population and the restriction to pure strategies, there always is a finite sequence of actions that can distinguish an individual from any other strategy in the population. For example, given a population strategy profile, consider the strategy for a player in the first role that for each history of length k or less, chooses an action that players in the first role select least often. So, in the first period, the strategy selects an action that is not used, if possible, but must specify an action that no more than one-half of the population uses. In each successive period, the new strategy choice distinguishes itself from at least one-half of the strategies that could have played consistently with the history observed thus far. Hence if k is the smallest integer greater than $\log_2 N$, then the new strategy would have revealed itself in no more than k periods. Given this revealing strategy profile, there will exist revealing histories, after which an individual from the other role could be certain that there is no possible existing strategy playing in this way. To get efficiency, modify role two players' strategies so that they respond to revealing histories using their actions that support the efficient payoff most preferred by role one players. Now modify the role one players' behavior so that they play the revealing strategy until they are revealed, and then coordinate on their good equilibrium. This argument shows that the dominating payoff is reachable. It is straightforward to modify the argument so that once you reach the frontier you can reach any point in the frontier.

Proposition 4. For any undiscounted, infinitely repeated game between patient players, there is a unique NES set. For any efficient payoff in the repeated game, there exists a homogeneous strategy profile contained in the NES set that yields the efficient payoff.

The assumption that the players do not discount plays a role in the proof. It permits individuals to 'waste' a finite sequence of stage-game actions without cost in order to signal a willingness to play to a new outcome. When players discount future payoffs, I can show that given any positive ϵ there exists a $\delta < 1$ such that if players use a discount factor at least as great as δ, then each NES set contains a strategy profile that yields payoffs within ϵ of the set of efficient stage-game payoffs.

Proposition 4 is quite similar to Proposition 2. In the repeated game context the early stages of the repeated game serve to communicate in the way that cheap talk does prior to a one-shot game. There is one substantive difference. In the repeated game setting at least one knows that efficient, individually rational outcomes are equilibria.

7. Related work

I have shown how evolutionary pressures can move players away from inefficient outcomes in three types of game that lack strict inefficient equilibria. Others have used variations of the methods on the same problems. I use this section to discuss these contributions.

NES permits a large amount of drift, which I exploit in the proofs of Propositions 2, 3, and 4. This type of drift is not appropriate in all contexts. In situations where players make mistakes implementing strategies, evolutionary pressures should discipline the behavior of players off the equilibrium path. Furthermore, it may seem strange to assume, as my stability notion implicitly does, that players would change behavior at unreached information sets. Similar results are possible even if the solution concept does not permit drift.

I begin by discussing the implications of static variants of evolutionarily stable strategies (ESSs) that are not set valued; no drift can take place if stable outcomes must be single strategies.

Binmore and Samuelson (1992), Fudenberg and Maskin (1990), Kim (1990), and Robson (1990) apply evolutionary arguments to select efficient outcomes in repeated games. These papers look at repeated games in situations where one expects any individually rational, feasible, stage-game payoff to be the average payoff of a subgame-perfect equilibrium of the repeated game, and provide conditions under which only a subset of the subgame-perfect equilibrium payoffs survive.

Binmore and Samuelson (1992) study repeated games without discounting. They assume that strategies must be implemented by finite automata and that there is a cost (infinitesimal relative to the payoffs of the underlying game) to adding states to the machine that describes a strategy. They do not consider the possibility of mistakes. As a result, an evolutionarily stable strategy cannot have unused states. This property reduces the punishments that are available to support equilibria. Consequently stable outcomes must satisfy a particularly strong efficiency property; they maximize the sum of the players' payoffs.

Fudenberg and Maskin (1990) look at undiscounted games, restrict attention to strategies of finite complexity, and assume that players make small mistakes. They demonstrate that no player does worse than his/her least preferred payoff in the set of payoffs that maximize the sum of the players' utilities at an ESS of a symmetric repeated game.

Binmore and Samuelson's prediction for games without common interests is more precise than Fudenberg and Maskin's. The difference between these two studies is primarily in the specification of preferences. Binmore and Samuelson's assumption that additional states are costly limits the nature of punishment strategies that may be used. Fudenberg and Maskin's assumption that players want to perform well in the event of a mistake imposes subgame perfection on stable strategies.

Proposition 4 implies that the NES set of a game without common interest will contain homogeneous strategy profiles that give rise to many different payoffs. The set of payoffs is strictly larger than that obtained in Fudenberg and Maskin or Binmore and Samuelson, and it is qualitatively different, as their work suggests that the population can settle down at a particular outcome, while mine suggests that the population will drift from one outcome to another. The essential reason for the difference in results is the different solution concepts. For an outcome to be stable in my framework, it must be able to resist invasions no matter how one specifies out-of-equilibrium behavior. More work is needed to determine when dynamic specifications will impose the discipline on strategies at unreached information sets needed to conclude that individual strategies will be evolutionarily stable in repeated games.

Both Binmore and Samuelson and Fudenberg and Maskin assume that the game is played by a single population (if the original game is not symmetric, then they assume that individuals play each of the two roles with equal probability). When roles are distinguished (and an asymmetric version of ESS is applied), any efficient payoff will be evolutionarily stable in Binmore and Samuelson's framework; any payoff in which each player does at least as well as its least preferred efficient payoff will be evolutionarily stable in Fudenberg and Maskin's framework.

Kim (1990) examines limit ESSs in finitely repeated games and shows that the set of limit-ESS payoffs is strictly contained in the set of subgame-perfect equilibrium payoffs, although generally contains inefficient payoffs.

Robson (1990) uses the device of extending the game to change outcomes. He considers the possibility of creating extra strategies in repeated games. These strategies play exactly the same role as communication does. Robson demonstrates how the additional strategies force cooperation in coordination games and that they destabilize the inefficient outcome in the prisoner's dilemma. As in this paper, Robson's analysis does not provide a clear prediction when the original game does not have common interests.

There are also papers that provide alternative evolutionary approaches to communication games. Wärneryd (forthcoming) shows that every neutral ESS[4] in a Sender–Receiver game must be efficient in a game in which both players receive a positive payoff if the Receiver correctly guesses the Sender's type, and they receive zero otherwise. These pure-coordination games are games of common interest, so Wärneryd's results are consistent with the results in section 5. Wärneryd's propositions do not extend to the broader class of common-interest games that we study. Strategies are not permitted to drift freely off the equilibrium path and an inefficient outcome may be stable if there were an action that led to especially low payoffs (for example, less than the pooling equilibrium payoffs) for all types. Selten's (1983) notion of limit ESS adds trembles to the game so that there are no unreached information sets. Limit ESSs must exist and be efficient for the games Wärneryd studies, but not for general common-interest games for the same reason that Wärneryd's results do not extend to pure coordination games.

Wärneryd (1991) applies the solution concept of his forthcoming paper to pure-coordination games with pre-play communication and complete information. Players must use pure strategies. In 2 × 2 games, he obtains the efficiency result. In larger games the result does not hold: neutrally stable strategies that support an inefficient equilibrium exist provided that the population is able to punish invading strategies by switching to an even less efficient equilibrium.

Bhaskar (1992) examines the same class of games, but permits randomization at the individual level. He shows that unless there is a countably infinite set of messages, neutrally stable strategies need not lead to efficiency, even in pure-coordination games. The problem is that when individuals randomize, there is no guarantee that unused messages are available. It is often straightforward to show, as I have done, that an invading strategy will be able to use an unsent message as a secret handshake [in Robson's (1990) terminology] that identifies the user to others in the population who are willing to cooperate, but leads to no disadvantage otherwise. When the message space has only finitely many elements and individuals can randomize over their signaling strategy, secret handshakes may not be possible. The assumption that there are a countable

[4] To be a neutral ESS (neutrally stable strategy) a strategy σ must have the property that no invading strategy can do strictly better than it when matched with a population that contains a small fraction of individuals playing the invading strategy (and the rest playing σ).

infinity of messages guarantees that there is a pair of rare messages that are used with arbitrarily small probability. An invading strategy that uses one of these rare messages and plays to induce an efficient payoff after its opponent also sends a rare message may not be an optimal response to the population, but it gains enough when it is matched with a similar strategy to compensate.

Bhaskar also obtains Wärneryd's (1991) results (that neutrally stable strategies must be efficient in 2 × 2 common-interest games with pre-play communication, but need not be for larger games) when players make small mistakes in signaling.[5] Bhaskar also studies a model in which players may misinterpret their opponent's message. This type of mistake rules out punishments for generic underlying games and permits Bhaskar to prove a version of Proposition 2.

Fudenberg and Maskin (1991) also study the effect of imposing evolutionary stability on games with pre-play communication. They assume that the underlying game is finite and symmetric, that there is a potentially unlimited number of rounds of pre-play communication in which the players speak simultaneously. They further assume that players make mistakes with small probability. Talk is not completely free, but its cost is infinitesimal relative to the probability of mistakes.

Fudenberg and Maskin obtain an efficiency result. Any evolutionarily stable payoff must give each player at least as much as he/she gets in his/her least favorite strongly efficient outcome (where an outcome is strongly efficient if it maximizes the sum of payoffs) in any evolutionarily stable payoff. This result corresponds to Proposition 2. In contrast to their work on infinitely repeated games, where strongly efficient payoffs can be supported as equilibria, existence is not assured in this setting. The central difference is that strongly efficient payoffs always can be supported as Nash equilibria in repeated games between patient players. They need not be Nash equilibria without repetition. Fudenberg and Maskin are able to prove existence if there is a symmetric, strict, strongly efficient equilibrium in the underlying game. This condition holds automatically in symmetric games with common interests. They also prove that if there exists a strict symmetric equilibrium that is better than a strict strongly efficient equilibrium, then it is evolutionarily stable.

There are also alternative set-valued ideas that, like NES sets, attempt to describe limiting outcomes in population games subject to evolutionary forces. These ideas all incorporate some form of drift in the solution.

The entry resistant (ER) sets introduced in Blume et al. (forthcoming) and Gilboa and Matsui's (1991) notion of cyclically stable sets [and the variants studied in Matsui (1991, 1992)] are qualitatively similar to NES sets. Entry resistant sets and cyclically stable sets are both minimal non-empty sets of strategies that are closed with respect to an entry condition. The entry condition is more restrictive than (R): both of the conditions implicitly assume an infinite population and require that new strategies respond not only to the population prior to the arrival of the new strategy, but also to a perturbed population that contains a tiny fraction of individuals who use the new strategy. Neither of these ideas requires elements of the stable set to be Nash equilibria. A result corresponding to Proposition 1 of this paper, that the efficient set of outcomes in common-interest games is stable, holds for these definitions. Matsui (1991) shows that all elements of the only cyclically stable set in 2 × 2, complete-information, common-interest games augmented by a round of pre-play communication are efficient. More general versions of the efficiency result are not true because Matsui permits individuals to use mixed strategies, and there is no guaran-

[5] Unlike Wärneryd (1991), Bhaskar (1992) must assume that players tremble in order to obtain the efficiency result in the 2 × 2 case. Otherwise one can support an inefficient outcome with a randomized punishment.

tee that there will be the unused messages needed to move to an efficient outcome for general games.

Swinkels (1992) and Thomas (1985a,b) have proposed set-valued evolutionary equilibrium notions. These ideas differ from NES in two ways: they require elements of the stable set to be Nash equilibria, and they impose a more restrictive Sentry condition than (R). An equilibrium evolutionarily stable (EES) set, due to Swinkels (1992), is a non-empty, closed set of Nash equilibria that is minimal with respect to an entry condition: any strategy σ that is an optimal response to the perturbed environment that arises when a small fraction of the population also plays according to σ can enter. The notion of an evolutionarily stable (ES) set due to Thomas (1985a,b) has a weaker entry condition than in the definition of an EES set: any strategy that does better than the population strategy in a perturbed environment can enter (the entrant need not be an optimal response to the perturbed environment). Due to my finite population assumption, the entry condition in the definition of NES is weaker still. Consequently any NES set that consists entirely of equilibria must be an ES set which in turn must be an EES set.

Kim and Sobel (1992) show that the payoffs of EES sets must coincide with the efficient payoffs in common-interest games that satisfy the additional property of equilibrium common interest (players have the same rankings over equilibria). Like Bhaskar (1992), they permit individuals in the population to randomize; without assuming that players have the same preferences over Nash equilibria of the underlying game there is no guarantee that stability implies efficiency.

Blume et al. (forthcoming) uses EES sets in a study of Sender–Receiver games. They prove versions of Propositions 1 and 3. Furthermore, they show that the non-equilibrium ER sets rule out inefficient non-communicative equilibria in a class of games that is slightly more general than common-interest games.

Finally, there remains the issue of whether there are *any* explicitly dynamic results corresponding to the analysis of this paper. Canning (1992) and Nöldeke and Samuelson (1992) present adaptive dynamic models of Sender–Receiver games. Using the techniques of Freidlin and Wentzell (1984), which were first used in game-theoretic models by Foster and Young (1990) and Kandori et al. (1993), these papers prove that the only limiting outcomes are efficient in a subset of common-interest games with cheap talk.[6] (Canning assumes that the Sender and the Receiver have the same preferences over actions. Nöldeke and Samuelson study the same set of coordination games that Wärneryd analyzed.) While the details of these dynamic specifications differ, they share important similarities. They assume that the population of players is finite; that players change their strategy randomly; and that mistakes or mutations occur and cause the models to have a unique ergodic distribution, which they can partially characterize using results of Freidlin and Wentzell. These techniques select between different basins of attraction of the dynamic based on the relative difficulty in moving from one state to another. As in this paper, movements arise because with positive probability the population will select an optimal response to its current configuration. The dynamics of these models exhibit the same kind of drift that occurs in NESs.

These dynamic arguments have been applied to complete-information games. Kandori and Rob (1991) use the model of Kandori et al. (1993) to provide a dynamic selection of the efficient equilibrium in pure-coordination games. Matsui and Rob (1991) and Nöldeke et al. (1991) have shown that only efficient outcomes arise as limits of an evolutionary dynamic process in pure-coordination games with pre-play communication.

[6] Canning also shows that without mistakes his dynamic need not abandon inefficient outcomes even in pure coordination games.

References

Bhaskar, V., 1992, Noisy communication and the evolution of cooperation, Delhi University.

Binmore, K. and L. Samuelson, 1992, Evolutionary stability in repeated games played by finite automata, Journal of Economic Theory 57, 278–305.

Blume, A., Y.-G. Kim and J. Sobel, forthcoming, Evolutionary stability in games of communication, Games and Economic Behavior.

Canning, D., 1992, Learning language conventions in common interest signaling games, Columbia University, mimeo.

Foster, D. and P. Young, 1990, Stochastic evolutionary game dynamics, Theoretical Population Biology 38, 219–232.

Freidlin, M. and A. Wentzell, 1984, Random perturbations of dynamical systems (Springer-Verlag, New York).

Fudenberg, D. and E. Maskin, 1990, Evolution and cooperation in noisy repeated games, American Economic Review 80, 274–279.

Fudenberg, D. and E. Maskin, 1991, Evolution and communication in games, preliminary notes.

Gilboa, I. and A. Matsui, 1991, Social stability and equilibrium, Econometrica 59, 859–867.

Kandori, M. and R. Rob, 1991, Evolution of equilibria in the long run: A general theory and applications, Princeton, mimeo.

Kandori, M., G. Mailath and R. Rob, 1993, Learning, mutation, and long run equilibria in games, Econometrica 61, 29–56.

Kim, Y.-G., 1990, Evolutionary analysis of two person finitely repeated coordination games, University of Iowa, mimeo.

Kim, Y.-G. and J. Sobel, 1992, An evolutionary approach to pre-play communication, UCSD, mimeo.

Matsui, A., 1991, Cheap-talk and cooperation in a society, Journal of Economic Theory 54, 245–258.

Matsui, A., 1992, Best response dynamics and socially stable strategy, Journal of Economic Theory 57, 343–362.

Matsui, A. and R. Rob, 1991, The role of public information and pre-play communication in evolutionary games, Pennsylvania, mimeo.

Nöldeke, G. and L. Samuelson, 1992, The evolutionary foundations of backward and forward induction, University of Wisconsin, mimeo.

Nöldeke, G., L. Samuelson and E. Van Damme, 1991, The evolution of communication, mimeo.

Robson, A., 1990, Efficiency in evolutionary games: Darwin, Nash and the secret handshake, Journal of Theoretical Biology 144, 379–396.

Selten, R., 1983, Evolutionary stability in extensive two-person games, Mathematical Social Sciences 5, 269–363.

Swinkels, J., 1992, Evolutionary stability with equilibrium entrants, Journal of Economic Theory 57, 306–332.

Thomas, B., 1985a, On evolutionarily stable sets, Journal of Mathematical Biology 22, 105–115.

Thomas, B., 1985b, Evolutionarily stable sets in mixed-strategist models, Theoretical Population 28, 332–341.

Van Damme, E., 1987, Stability and perfection of Nash equilibria (Springer-Verlag, Berlin).

Wärneryd, K., 1991, Evolutionary stability in unanimity games with cheap talk, Economics Letters 36, 375–378.

Wärneryd, K., forthcoming, Cheap talk, coordination, and evolutionary stability, Games and Economic Behavior.