Evolutionary Stability in the Repeated Prisoner's Dilemma

JOSEPH FARRELL

Department of Economics, University of California,
Berkeley, California 94720

AND

ROGER WARE

Department of Economics, University of Toronto,
Toronto, Ontario, Canada M5S 1A1

Received June 27, 1988

Recently, biologists have explored evolutionary explanations of apparently altruistic behavior in situations of conflict, often modeled as the "Prisoner's Dilemma." Certain simple cooperative strategies, notably TIT-FOR-TAT, have been successful in computer simulations of the evolution of populations of individuals who interact according to the Prisoner's Dilemma. Some attempts to formalize this have used the concept of "evolutionary stability." But Boyd and Lorberbaum (1987, *Nature* (*London*) **327**, 58–59) recently showed that no single pure strategy (such as TIT-FOR-TAT) can be evolutionarily stable. We extend their argument to derive a more powerful result, which implies, first, that no finite mixture of pure strategies can be evolutionarily stable, and, second, that no mixture of TIT-FOR-*n*-TATS can be evolutionarily stable. We interpret our negative results to suggest that evolutionary stability is too demanding a criterion. © 1989 Academic Press, Inc.

In an influential book (Axelrod, 1984) and a series of articles, Axelrod has argued that social cooperation may evolve through natural selection. This possibility is clearly of great importance for the understanding of animal and human behavior. Axelrod and others have focussed on the repeated Prisoner's Dilemma game as a simple framework in which to study the evolution of cooperation. Of particular interest is whether there exist cooperative strategies for this game which are *evolutionarily stable*, that is, which cannot be invaded by other strategies or combinations of strategies (Maynard Smith, 1982).

Recently, Boyd and Lorberbaum (1987) showed that no pure strategy

0040-5809/89 $3.00

could be evolutionarily stable. Their proof consisted of showing that, given any pure strategy, a combination of strategies can always be found which will successfully invade a population consisting of that single strategy. We extend their argument to show that in any evolutionarily stable mixture of strategies, every finite history occurs with positive probability. This implies in particular that no *finite mixture* of pure strategies can be evolutionarily stable. It also implies that certain apparently plausible infinite mixtures cannot be evolutionarily stable. Our result may suggest that evolutionary stability is too strong a solution concept.

The idea behind our result can be summarized as follows. In any population consisting of a finite mixture of pure strategies, some sequences of choices by the players ("histories") are impossible. In game-theoretic terms, there are nodes in the game tree that are reached only with zero probability: we call such nodes "unreached." For example, consider a population consisting only of TIT-FOR-TAT (the strategy defined by "cooperate on the first move, and thereafter each period play the same move as your opponent's previous move") and ALLC (the strategy that always cooperates). In such a population, the two-period history [(C, C), (D, D)] in which both cooperate in the first period and both defect in the second period can never occur. Adding finitely more pure strategies could easily make that history possible (for instance, add the strategy "alternate between C and D"), but could not (as we show is required) make *every* history possible.

We take such an unreached node, and construct a *mutation* strategy that mimics one of the given strategies up to a certain node just before the unreached node, and then does the opposite, so that it sometimes reaches the formerly unreached node when it interacts with the given population. Then we design an *invading* strategy that mimics one of the given strategies until it reaches the previously unreached node (as it will sometimes do in a population augmented by the "mutation" we constructed), but then behaves differently. It is easy to design the mutation so that it yields a higher expected fitness for the invading strategy than for the strategy which the latter imitates. Since the invading strategy plays exactly like one of the original strategies against all the original strategies, it will have higher expected fitness against the population (since it does better against the mutation), and hence the original mixture of strategies is not evolutionarily stable.

We first illustrate the idea of invasion with some familiar strategies. First, as Boyd and Lorberbaum show, TIT-FOR-TAT (TFT) can be invaded by TIT-FOR-TWO-TATS (TF2T, which always cooperates except immediately after two successive defections by its opponent), if mutation maintains a low but positive frequency of SUSPICIOUS-TIT-FOR-TAT (STFT, which is like TFT except that it defects on the first move). It is easy to see why. TFT and TF2T play identically against themselves and against each other (in Axelrod's characterization, they are both "nice") but against STFT, TFT gets locked into an alternating cycle of mutual retaliation (a vendetta), whereas TF2T is sufficiently tolerant to induce STFT to cooperate.

We can then ask whether a mixture of TF2T and STFT is evolutionarily stable. Boyd and Lorberbaum show that a mixture with fractions p and $1-p$ of TF2T and STFT is "locally stable" if p is such that, given the p: $1-p$ mixture, the two strategies have equal expected fitness. Such an equilibrium can equivalently be regarded either as a mixture of pure strategies within the population occurring with *frequencies* p and $1-p$, or, in the language of game theory, as a symmetric pair of mixed strategies (Luce and Raiffa, 1957) played with *probabilities* p and $1-p$.

But the stability of this equilibrium is "local": it depends on restricting attention to the *given* two strategies. It is easy to construct a third strategy that will invade this mixture. Consider for instance the following strategy: play D on the first move, then always cooperate (ALLC) in response to a D on the rival's first move, and alternate C and D in response to a C on the rival's first move. This strategy will exploit the tolerant TF2T and generate cooperation with STFT, thus successfully invading the mixture. (Intuitively, it "tests" its opponent, and behaves differently according to what it learns.) Thus the mixture is not evolutionarily stable.

More generally, we show that a simple extension of Boyd and Lorberbaum's argument establishes that *no* finite mixture of pure strategies can be evolutionarily stable; indeed, we prove the stronger result:

PROPOSITION. *In any evolutionarily stable mixture of strategies in the infinitely repeated Prisoner's Dilemma game, where the probability of further interaction is sufficiently high, every finite history occurs with positive probability.*

Proof. Consider a population mixture of strategies, P, such that at least one finite history occurs only with zero probability. We show how to construct a pair of strategies, an "invading strategy" I and a sustained "mutation strategy" M, such that when one-way mutation sustains M at a low frequency in the population, I can successfully invade the given mixture P.

By assumption, there exists some finite history that cannot be generated by the interaction of strategies in P. Let H be a shortest such history, or in game-theoretic terms, a first unreached node in the game tree. Thus, the previous node, $H-1$, *is* reached by the interactions of some pairs of strategies in P. Label one such pair (S_1, S_2).

First, we claim that we can always pick S_1, S_2 and H so that, if we changed the move chosen by S_2 at $H-1$, then H would result. This point

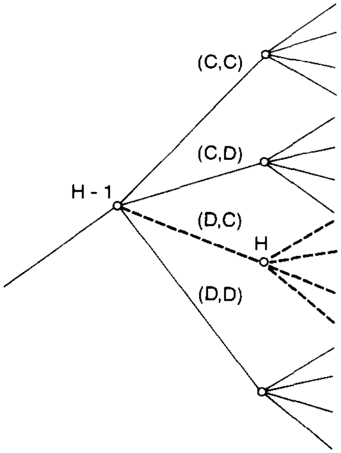

FIG. 1. Branches of the game tree at H, a first unreached node.

is illustrated with the aid of Fig. 1. Suppose for instance that, from $H-1$, the branch (D, C) leads to H. If the branch (D, D) is followed by some pair of strategies, choose these as S_1 and S_2: then changing S_2 will imply H. If the node corresponding to (D, D) is not reached by any pair of strategies, call this node H. Provided that (C, D) is reached, the corresponding strategies can be labelled S_2 and S_1, respectively, and the claim follows. Finally, if (C, D) is not reached, then (C, C) must be, and the claim follows.

Now we describe a mutation strategy M whose sustained presence makes it possible to design a successful invading strategy. Let M imitate S_2 at all nodes *except* at $H-1$ and after histories that begin with $H-1$. At $H-1$, M plays the opposite of S_2's choice, thus reaching node H if it happens that its opponent is S_1. At H, M cooperates, and thereafter it plays ALLC or ALLD as described below.

The invading strategy I is as follows. Let I imitate S_1 at all nodes except H and histories that begin with H. Since H is never reached in the given population P, which contains S_1, H is never reached by the augmented population P^* with I added. Define $V(S_i \mid S_j)$ to be the expected fitness of a strategy S_i against a strategy S_j, and define $V(S_i \mid S)$ to be $\sum_j p_j V(S_i \mid S_j)$, where p_j are the frequencies of each strategy S_j in the population S. Since I always emulates S_1 in interacting with P, it follows that $V(I \mid P^*) = V(I \mid P) = V(P \mid P^*) = V(P \mid P)$. Accordingly, if M is the only (or the most significant) addition to P^*, I can invade if and only if $V(I \mid M) > V(P \mid M)$. We complete our constructions of M and I so that this is true.

At H, let I do the opposite of what S_1 would do. Thereafter let I play

ALLD. After H, let M play ALLD if its opponent acted like S_1 at H, and play ALLC otherwise.

From our construction, both S_1 and I reach H when paired with M. After H, I does strictly better than S_1, since M plays ALLC against I and plays ALLD against S_1. Although at H, I may do worse than S_1, nevertheless $V(I|M) > V(S_1|M)$ provided that w, the probability of a further interaction, is large enough (see Boyd and Lorberbaum for a more detailed explanation, including a proof that the lower bound on w is independent of H). Therefore the original mixture was not evolutionarily stable. ▌

It is worth noting that, while the set of nodes is countably infinite, it is nevertheless mathematically possible to put non-zero probability on each. This is most easily seen by imagining that each player simply randomizes (with equal probabilities) between C and D at every move. For simplicity, and because it is traditional in the biological literature, we restrict attention to "pure strategies" in which players do not randomize, but the reader should note that the effects of such strategies are generally replicated by assuming that some players play the pure strategy C and others the pure strategy D at the given node.

We can rephrase our Proposition by saying that the behavior of any evolutionarily stable strategy mixture can never be completely predictable, even after any finite history of the game. Thus, we can never make (non-probabilistic) predictions about how a given pair of agents will behave with one another, even after observing them together for an arbitrarily long time.

Our proposition has two corollaries of interest. First, no finite mixture of pure strategies can be evolutionarily stable. For any finite mixture of pure strategies can follow only finitely many paths with positive probability, and therefore must leave infinitely many unreached nodes, which cannot be evolutionarily stable.

A second corollary bears on the observations of May (1987), who suggests that TFT's failure to be evolutionarily stable may be repaired if there is some probability of not defecting immediately in response to a defection by a rival. If we represent his suggestion by a mixture of the strategies TF*n*T, where TF*n*T retaliates only after n defections, then our result shows that any such mixture would *not* be evolutatationarily stable, since histories such as [(C, C), (D, D)] would never occur in such a population. It is not sufficient for evolutionary stability, though it is necessary, that infinitely many pure strategies coexist in the population.

What should be concluded from these observations? The possibility of finding an infinitely mixed set of strategies that is evolutionarily stable is unsatisfactory since it could yield no firm predictions. A more promising direction for theoretical research might be to restrict the allowable mixtures and mutant strategies in ways suggested by empirical observation. Our argument, like Boyd and Lorberbaum's, relies on constructing somewhat unnatural mutants, and reasonable restrictions might restore the existence of evolutionarily stable strategies or mixtures. For example, it might be sensible to restrict strategies to a short memory, i.e., actions can only be based on, say, the two previous moves, and not on the entire history of interaction (Kalai and Stanford, 1988.) The particular environment under study may also indicate restrictions on the kinds of strategies and mutations that can arise, which may in turn restore a restricted evolutionary stability. In game-theoretic terms, we might say that evolutionary stability is too strong or too demanding as a solution concept: outcomes that are not evolutionarily stable may nevertheless be perfectly plausible, provided that the invading mixtures are not likely to arise.

References

AXELROD, R. 1984. "The Evolution of Cooperation," Basic Books, New York.

BOYD, R., AND LORBERBAUM, J. P. 1987. No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game, *Nature* (*London*) **327**, 58–59.

KALAI, E., AND STANFORD, W. 1988. Finite rationality and interpersonal complexity in repeated games, *Econometrica* **56**, 397–410.

LUCE, D., AND RAFFIA, H. 1957. "Games and Decisions," Wiley, New York.

MAY, R. 1987. More evolution of cooperation, *Nature* (*London*) **327**, 15–17.

MAYNARD SMITH, J. 1982. "Evolution and the Theory of Games," Cambridge Univ. Press, Cambridge.