J. theor. Biol. (1990) **144**, 379-396

Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake

ARTHUR J. ROBSON

Department of Economics, University of Western Ontario, London, Ontario, Canada N6A 5C2

(*Received on* 19 *May* 1989, *Accepted in revised form on* 3 *January* 1990)

This paper considers any evolutionary game possessing several evolutionarily stable strategies, or ESSs, with differing payoffs. A mutant is introduced which will "destroy" any ESS which yields a lower payoff than another. This mutant possesses a costless signal and also conditions on the presence of this signal in each opponent. The mutant then can protect itself against a population playing an inefficient ESS by matching this against these non-signalers. At the same time, the mutants can achieve the more efficient ESS against the signaling mutant population itself. This construction is illustrated by means of the simplest possible example, a co-ordination game. The one-shot prisoner's dilemma is used to illustrate how a superior outcome which is not induced by an ESS may be temporarily but not permanently attained. In the case of the repeated prisoner's dilemma, the present argument seems to render the "evolution of co-operation" ultimately inevitable.

1. Introduction

The title "*The Origin of Species*" seems to suggest, at least to the unwary, that Darwin intended a group to be the unit of selection. The subtitle of Darwin's treatise reinforces this casual impression—"*... or the Preservation of Favoured Races in the Struggle for Life*". However, Darwin explicitly recognized that "the struggle for life (is) most severe between individuals and varieties of the same species", (Darwin, 1859; Mayr, 1982: 484-485, attributes Darwin's recognition of the importance of intra-species competition to his reading of Malthus, 1798, but see also pp. 491-493). It is still not uncommon to hear popular explanations of animal behavior which rely upon an appeal to the interest of a species or, occasionally, even to the collective interest of several. A few modern biologists believe that there exist phenomena which cannot be explained without invoking a group selection mechanism (see e.g. Wynne-Edwards, 1962). Most other modern biologists, on the other hand, seem to find such explanations to be unattractive in the light of the logic of natural selection. Dawkins, (1976: 8-9), for example, argues that natural selection operates most forcefully below the level of the group and, far from promoting the interest of the group, might entail its extinction. Indeed, Dawkins argues that the unit of selection is, in a certain sense, below even the level of the individual, is the gene (see, in particular Dawkins, 1982*a*: 45-64).

0022-5193/90/011379+18 $03.00/0

Views of evolutionary biology which emphasize the role of the individual are, of course, highly congenial to economists. Further strengthening both the formal and substantive links between the two disciplines is that game theory has now also been applied to evolutionary biology (see Maynard Smith & Price, 1973; Maynard Smith, 1982). Consider the interactions of the individuals of one particular species with one another. Suppose that these interactions involve two individuals at a time contesting some scarce resource, such as food. With a large population, the equilibrium outcome introduced by Maynard Smith & Price is a special kind of symmetric Nash equilibrium, involving what is designated an "evolutionarily stable strategy", or ESS. A population playing such a strategy is immune to invasion by a mutant population playing any alternative strategy. It need not be the case, of course, that such an ESS is unique, and indeed ESSs may have different payoffs. That is, an ESS can exist which yields an outcome which is inferior to that obtained under some other ESS. Such a situation has a well-known analog within pure game theory. That is, it is often possible to Pareto rank some of the Nash equilibria of a game. A long-standing problem is how to construct a theory which predicts Pareto-efficiency within the set of Nash equilibria.

The intention of the present work is to suggest that "Mother Nature" might be less baffled by this problem than are game theorists. For there is a simple mechanism, based on the *individual* as the unit of selection, whereby such inefficient ESSs can be destroyed. That is, a mutation can be defined which will successfully invade a population which is playing any ESS which is payoff-dominated by another. This mutation must, of necessity, involve more than simply a different choice from the original set of strategies, for it is just these latter mutants which are considered in the definition of an ESS. Indeed, the mutation here entails the possession of a signal, that is, an observable characteristic which can be taken to have zero inherent cost. Mutants recognize the presence or absence of this signal in the other individual and condition their choice of strategy on this.

It should be emphasized that there is little doubt that animals actually use signals, for certain purposes at least. Thus, for example, Maynard Smith (1982: 82–86) discusses the Harris sparrow, individuals of which vary in the color of their plumage and also in aggression towards other birds, with the darker birds being more aggressive. Dark birds painted pale continued to behave aggressively but were involved in a larger number of fights than normal dark birds. Pale birds painted dark tended to be persecuted by normal dark birds and were sometimes forced to feed away from the flock. It is not asserted that this will precisely fit the model here. What it does demonstrate is the reality of signaling as a natural phenomenon.

Consider then a population for some evolutionary game which is in a low-level trap, that is an ESS which is inferior to another ESS. The appropriate mutant uses the kind of signal discussed above as follows. Against the old non-signaling population the mutant plays the old inefficient EES, thus protecting itself from the consequences that would otherwise occur. On the other hand, mutants recognize the signal in other mutants and then can attain the more efficient ESS. It is assumed that the old non-signaling population remains blind to the signal. (This motivates the phrase the "secret handshake" of the subtitle. This assumption of blindness is without loss of generality in the sense that it is not-possible for non-signalers to do better than by playing the old ESS against such mutants, given the mutants' behavior.)

The argument can better be understood with the aid of the simplest possible example, an evolutionary game with two pure strategies, each of which is an ESS. This is presented in section 3.1. The effect of the mutation is to add a third pure strategy which plays the inferior strategy against either of the original two strategies and the superior strategy against itself. Thus the entries in the new 3×3 matrix are derived in a straightforward way from those in the original 2×2 matrix. It is then readily shown that the only ESSs now are the old superior strategy and the new mutant strategy, which are equivalent in payoffs. Furthermore, it is shown that the population must converge to one of these equivalent ESSs, given a generic initial point, under the pure-strategy dynamic. Initial points near the original inefficient ESS converge in particular to the new efficient mutant ESS.

The analysis of the above example suggests that a similar mutant would be successful in invading a population playing some ESS if only there exists an *outcome* yielding a superior payoff. That is, the argument does not appear to use directly the assumption that the superior outcome be itself an ESS. In order to discuss this issue, the one-shot "prisoner's dilemma" is presented as a second example, in section 3.2. It is noted firstly that, indeed, the addition of an appropriate mutant to the standard 2×2 prisoner's dilemma game results in a 3×3 game in which the group preferred "co-operative" outcome is generated by the unique ESS. Even if this is credible as a short-run outcome, there is an obvious objection to this in the long-run. That is, there is another mutant waiting impatiently in the wings, a mutant which avails itself of the signal and plays the old "fink" ESS against the old population, but also "finks" against the first mutant. This second mutant thus "finks" unconditionally. It is desirable, then, to consider the 4×4 game with both mutants. It is noted firstly that this 4×4 game does not, strictly speaking, possess any ESS. However, this is simply a technicality rather than a deep difficulty. It is shown, indeed, that the path of the population over time generally converges to some mixture between the old "fink" strategy and the mutant which signals but "finks" always anyway, under the pure strategy dynamic. The limiting payoff is uniquely determined, in general, as the usual payoff for the 2×2 game. In essence then, the conventional equilibrium prediction for the prisoner's dilemma is ultimately restored.

The next section defines the notion of an ESS and the dynamical systems. Section 3, as noted above, discusses two key examples. Section 4 discusses related work and the implications of the present results. The Appendix shows how the analysis of the first example can be generalized whenever there are at least two ESSs with differing payoffs.

2. Definitions

The description of the game and motivation of the ESS definition follow Hines (1987):

Definition 1: *Basic evolutionary game*

There is a large (strictly infinite) population of individuals who interact two at a time. (Riley, 1979, discusses the complications needed in order to analyze small populations). In each interaction, each individual has pure strategies $i = 1, \dots, n$. These yield the associated mixed strategies denoted by x, say, where $x_i \geq 0$ and $\sum_{i=1}^{n} x_i = 1$ and so $x \in \Delta^{n-1}$, the $(n-1)$ simplex in R^n. If a given individual chooses pure strategy i and his opponent chooses j, the payoff to the first individual is a_{ij}. The payoff to the second individual, given symmetry between the two individuals, is therefore a_{ji}. (These payoffs measure "fitness" in the biological sense of determining the number of descendants, as is modeled explicitly in the dynamics below.) This symmetry between individuals means that it is enough to specify $A = (a_{ij})$, a single $n \times n$ matrix.

The payoff to choice of mixed strategy x against a mixed strategy x^* is then readily seen to be:

$$x^T \mathbf{A} x^*.$$

The requirements for an ESS can then be motivated as follows. Consider a "monomorphic" population every member of which plays the mixed strategy x^*. This is to be resistant to invasion by a monomorphic mutant playing any alternative mixed strategy x, say. Suppose indeed that the fraction $(1-\varepsilon)$ of the total population plays x^* and the fraction ε plays x, where each opponent is randomly selected. Effectively, then, each individual faces the mixed strategy $(1-\varepsilon)x^* + \varepsilon x$, where ε is small. The payoff to x^* is then;

$$x^{*T}\mathbf{A}[(1-\varepsilon)x^* + \varepsilon x] = (1-\varepsilon)x^{*T}\mathbf{A}x^* + \varepsilon x^{*T}\mathbf{A}x,$$

whereas the payoff to x is;

$$x^T\mathbf{A}[(1-\varepsilon)x^* + \varepsilon x] = (1-\varepsilon)x^T\mathbf{A}x^* + \varepsilon x^T\mathbf{A}x.$$

This suggests the following, as in Maynard Smith (1982):

Definition 2: *Evolutionarily stable strategy, ESS*

An ESS is $x^* \in \Delta^{n-1}$ such that, for all $x \in \Delta^{n-1}$, where $x \neq x^*$, *either*;

(i) $x^T\mathbf{A}x^* < x^{*T}\mathbf{A}x^*$

or

(ii) $x^T\mathbf{A}x^* = x^{*T}\mathbf{A}x^*$ and $x^T\mathbf{A}x < x^{*T}\mathbf{A}x$.

The above conditions can be paraphrased in game theoretic terms as follows. It is required that the strategy x^* be a best-reply to itself and, further, that any other best-reply to x^*, x, say, be a worse reply to x than is x^*. In particular, then, an ESS yields a symmetric Nash equilibrium. Not every symmetric mixed strategy Nash equilibrium induces an ESS, however, as taking $\mathbf{A}$ to be a matrix of zeroes shows. Indeed, this also shows an ESS need not exist.

If each individual were restricted to the use of the pure strategies $i = 1, 2, \dots,$ the vector $x^* \in \Delta^{n-1}$ can then be reinterpreted as the distribution of a polymorphic population over these pure strategies. If x^* is an ESS as above, its immunity to invasion by any mixed strategy mutant implies immunity, in particular, to any pure strategy mutant. However, it is not true that immunity to invasion by any pure strategy mutant implies x^* is an ESS (see Maynard Smith, 1982: 185).

For some purposes it is useful to consider not only an appropriate static equilibrium concept, as above, but the evolution of the population over time. The first specifications of dynamics were based on pure strategies (see e.g. Taylor & Jonker, 1978; Zeeman, 1979). For simplicity of exposition, the present paper will also confine attention to such pure strategy dynamics. They are given as:

Definition 3: *Pure strategy dynamics*

The fraction of the population playing a particular pure strategy is taken to evolve according to the difference between the "fitness" of this pure strategy and the average fitness of the population. This yields a concrete interpretation of the payoff matrix, **A**. Thus,

$$\frac{\dot{x}_i}{x_i} = (\mathbf{A}x)_i - x^T\mathbf{A}x$$

or

$$\dot{x}_i = x_i[(\mathbf{A}x)_i - x^T\mathbf{A}x], \qquad i = 1, \ldots, n,$$

where the r.h.s. is cubic in x. Clearly, if $x_i = 0$, for any i, at $t = 0$, then $x_i = 0$ always. Furthermore, it is easily seen that every solution path remains on the unit simplex, given that it starts there. Hence the dimensionality of the system is $n - 1$ rather than n.

A point $x^* \in \Delta^{n-1}$ is said to be a "point attractor" if it is the limit of the solution path of the dynamical system for all initial values in some neighborhood of x^*. Zeeman (1979) shows that any ESS must be a point attractor for the pure strategy dynamic, but also gives an example of a point attractor which is not an ESS.

It should be noted, however, that it is, in some ways, more mathematically complete to allow mixed strategies throughout. Thus Hines (1980, 1987) obtains the following:

Definition 4: *Mixed strategy dynamics*

Suppose that the distribution of the population over mixed strategies is given by the probability measure F on Borel subsets of Δ^{n-1}. Define then the mean mixed strategy as;

$$\mu = \int x \, \mathrm{d}F(x),$$

and the associated covariance matrix as;

$$\mathbf{C} = \int (x - \mu)(x - \mu)^T \, \mathrm{d}F(x).$$

If $x^* \in \operatorname{Int} \Delta^{n-1}$ is a fixed point, Hines (1980) shows that the appropriate mixed strategy dynamic satisfies;

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mu - x^*) = \mathbf{CA}(\mu - x^*).$$

Hines can then show that $x^* \in \text{Int}\,\Delta^{n-1}$ is always a "sink" for the mixed strategy dynamic system for μ if and only if x^* is an ESS as above. (To be a sink here is to require the eigenvalues of **CA**, suitably restricted, all to have negative real parts. This is to hold for all appropriate **C**). Note that convergence of μ-x^* does not imply that the limiting distribution itself is uniquely determined (see Hines, 1987: 208, for further complications).

Altogether, then, the definition of an ESS is most readily understood in terms of mixed strategies. Furthermore, being an ESS is, in a sense, necessary and sufficient for the stability of the mixed strategy dynamic. On the other hand, whereas being an ESS is sufficient to be immune to invasion by any pure strategy, it is not necessary. Again, being an ESS is sufficient to be an attractor of the pure strategy dynamic, but it is not necessary.

3. Two Key Examples

3.1. A CO-ORDINATION GAME

This is, it seems, the simplest possible game yielding two ESSs which have different payoffs. It is given in Fig. 1.

	u	*d*
u	1	0
d	0	2

FIG. 1. A co-ordination game.

It is clear that both "*u*" and "*d*" are ESSs and it is not difficult to see that there is no other ESS (this follows also from Bishop & Cannings, 1978: 91). Clearly the ESS "*d*", which yields a payoff of 2, is better for the population as a whole than is "*u*", which yields only a payoff of 1. Notice however that "*u*" is certainly immune to invasion by a small group of mutants playing "*d*". Indeed, suppose that the mutant "*d*" comprises a fraction ε of the total population with the remaining fraction $(1-\varepsilon)$ being still "*u*". In this case, each mutant obtains an average payoff of 2ε against the whole population, whereas the original "*u*" strategy yields $(1-\varepsilon)$. Thus the mutant will die out if $\varepsilon < 1/3$.

Now suppose that the mutation discussed in the introduction is introduced. This mutant carries a signal which is assumed to cost nothing to produce. Furthermore the mutant recognizes the presence of the signal in its opponent and conditions its choice of strategy on this. Suppose that the mutant here plays "*u*" against the non-signaling original population but plays "*d*" against other signaling mutants. The effect of this is to enlarge the original 2×2 game to the 3×3 game given in Fig. 2.

In Fig. 2, the mutant signaling strategy is labeled "*m*". It should be emphasized that this enlarged 3×3 game still has just "*u*" and "*d*" as its underlying choices

	u	*d*	*m*
u	1	0	1
d	0	2	0
m	1	0	2

FIG. 2. The co-ordination game with a signaling mutant.

and that the payoff consequences of a given pair of these underlying choices are also as in the 2×2 game.

It is easy to check that "u" is no longer an ESS, although (u, u) remains a Nash equilibrium. That is, "m" is also a best reply to "u" but "m" does better against itself than "u" does against "m". The only ESSs are now easily seen to be "d" and "m".

In order to more fully characterize the behavior of the population with these three strategies, the appropriate pure strategy dynamic system derived from the above matrix, **A**, is treated. This is:

$$\frac{\dot{x}_1}{x_1} = x_1 + x_3 - W$$

$$\frac{\dot{x}_2}{x_2} = 2x_2 - W$$

$$\frac{\dot{x}_3}{x_3} = x_1 + 2x_3 - W,$$

where, x_1, x_2, and x_3 are the components of the vector x, corresponding to "u", "d" and "m" respectively, and where,

$$W = x_1(x_1 + x_3) + 2x_2^2 + x_3(x_1 + 2x_3),$$

now denotes the average fitness of the entire population. Using the relation that $x_1 = (1 - x_2 - x_3)$ to eliminate x_1 and simplifying yields the following equations for x_2 and x_3:

$$\frac{\dot{x}_2}{x_2} = -1 + 4x_2 - 3x_2^2 - x_3^2$$

$$\frac{\dot{x}_3}{x_3} = x_2 + x_3 - 3x_2^2 - x_3^2.$$

The phase diagram for these equations is readily derived and is sketched in Fig. 3. (When there are two dimensions it is not possible for exotic behavior such as "chaos" to arise).

This diagram shows that generic initial points with strictly positive amounts of all three strategies have solution paths which tend to either "d" or "m", and there

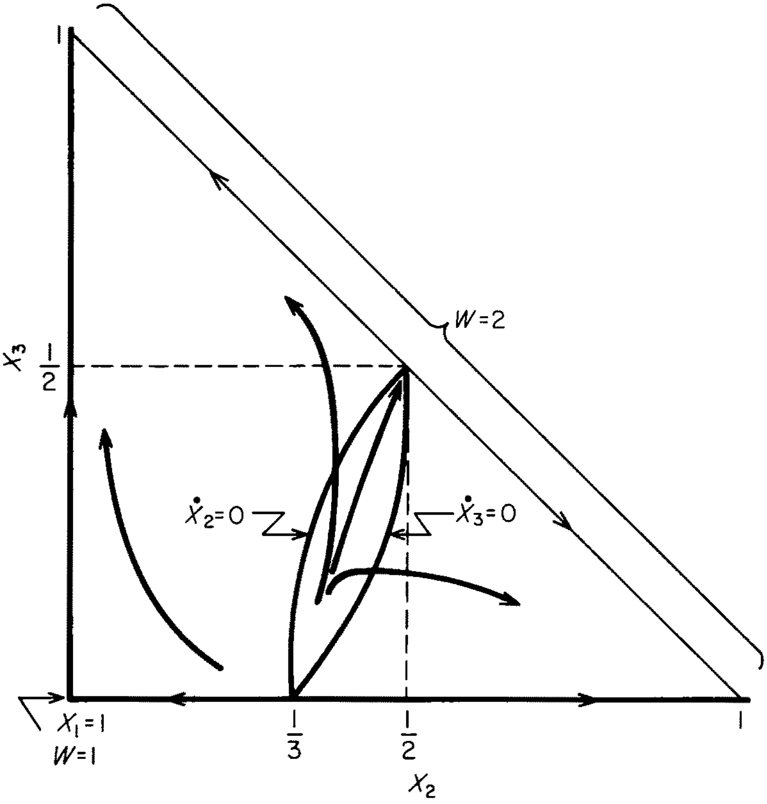

FIG. 3. Phase diagram for the co-ordination game with a mutant.

is a limiting payoff of 2 in all cases. All such initial points near "*u*" have solution paths which tend to "*m*" in particular. That is, the old inefficient ESS at "*u*" has, in this sense, been supplanted by a new efficient ESS at "*m*". Further, this new ESS is clearly immune to the introduction of the further mutant which signals but plays "*u*" against other signalers.

It should be noted that the above process is *not* reversible. That is, suppose instead the mutant plays the *efficient* ESS against non-signalers and plays the *inefficient* ESS against fellow signalers. It is easily shown that "*u*" and "*d*" remain the only ESSs of this augmented game. Hence "*d*", in particular, is not destroyed as an ESS by the introduction of this mutant. Such a mutant can indeed be shown to die out.

3.2. THE ONE-SHOT PRISONER'S DILEMMA

This game is given as Fig. 4. In this case it is easily seen that the unique ESS is "*u*", which yields a payoff of 2, despite the possibility of obtaining 3 by means of

	u	*d*
u	2	4
d	1	3

FIG. 4. One-shot Prisoner's Dilemma.

the entire population playing "*d*". The analysis of the previous example suggests that a mutant playing "*u*" against the old population and "*d*" against itself will be able to successfully invade a population playing "*u*". Indeed, consider Fig. 5.

The only ESS now is easily seen to be "*m*", which entails the payoff of 3. If this mutant and only this mutant were introduced, there is no reason to doubt the merit of this ESS. The difficulty is just that such an ESS is a "sitting duck" for the introduction of still another mutant, one which would prey on the first mutant. The second mutant should carry the signal, but play "*u*" against the first mutant as well as against the non-signaling population. (It would seem that this second mutant could evolve relatively easily from the first since all it involves is a switch in the underlying choice to be played against other signalers). With the introduction of the second mutant in addition to the first the game is as in Fig. 6.

	u	*d*	*m*
u	2	4	2
d	1	3	1
m	2	4	3

FIG. 5. One-shot Prisoner's Dilemma with mutant.

In Fig. 6, the second mutant is labeled "*f*". It is not hard to see that there are now no ESSs, and hence the dynamical system associated with this matrix needs to be analyzed directly. This is a three-dimensional system on the tetrahedron, Δ^3. In general, three-dimensional systems can have markedly more complex behavior than that possible in two dimensions. However, the present example is rather simple. Notice, in fact, that the strategy "*d*" is unambiguously "dominated" not just by

	u	*d*	*m*	*f*
u	2	4	2	2
d	1	3	1	1
m	2	4	3	1
f	2	4	4	2

FIG. 6. One-shot Prisoner's Dilemma with two mutants.

one of the other strategies but by all of them. (A strategy is said to be dominated by another strategy if it yields no more payoff for every choice of the other player and strictly less for some choice.) When strategy "d" is included in the dynamical system, it unambiguously decreases to zero along non-trivial solution paths. It can be shown that the limiting behavior of the full system is then determined by the limiting behavior of the two-dimensional system where "d" and its corresponding fraction "x_2", say, are simply omitted.

The following equations obtain in this case:

$$\frac{\dot{x}_1}{x_1}=2x_1+2x_3+2x_4-W$$

$$\frac{\dot{x}_3}{x_3}=2x_1+3x_3+x_4-W$$

$$\frac{\dot{x}_4}{x_4}=2x_1+4x_3+2x_4-W,$$

where x_1, x_3, and x_4, are the fractions of the population playing strategies "u", "m", and "f" respectively and where;

$$W=x_1(2x_1+2x_3+2x_4)+x_3(2x_1+3x_3)+x_4(2x_1+4x_3+2x_4).$$

is the average fitness of the entire population. Using the fact that $x_1=(1-x_3-x_4)$ to eliminate x_1 on the r.h.s., the system can be expressed as;

$$\frac{\dot{x}_1}{x_1}=-(x_3+x_4)x_3\leq 0$$

$$\frac{\dot{x}_3}{x_3}=-x_4+x_3(1-x_3-x_4)$$

$$\frac{\dot{x}_4}{x_4}=x_3+x_3(1-x_3-x_4)\geq 0.$$

The phase diagram for this essentially two-dimensional system in x_3 and x_4, say, is represented in Fig. 7.

All generic solution paths of the dynamical system converge ultimately to a mixture of "u" and "f", and so have limiting payoff of 2 as in the usual equilibrium for the original 2×2 prisoner's dilemma game. However, each such path takes a detour towards the vertex at which the first mutant "m" is the entire population and the average fitness is hence 3. Instead, average fitness at first increases along the path but then decreases as the path heads back to a mixture of "u" and "f". Although the indeterminate nature of the mixture involved implies that there exists no ESS, that there exists indeed no point attractor, this is a technicality in that the payoff is determinate. Note that this example is clearly non-generic in that small *independent* perturbations of the payoffs in the 4×4 matrix here are likely to break the ties in the payoffs. However, these ties are produced endogenously by the signaling mutants and hence should be treated as ties.

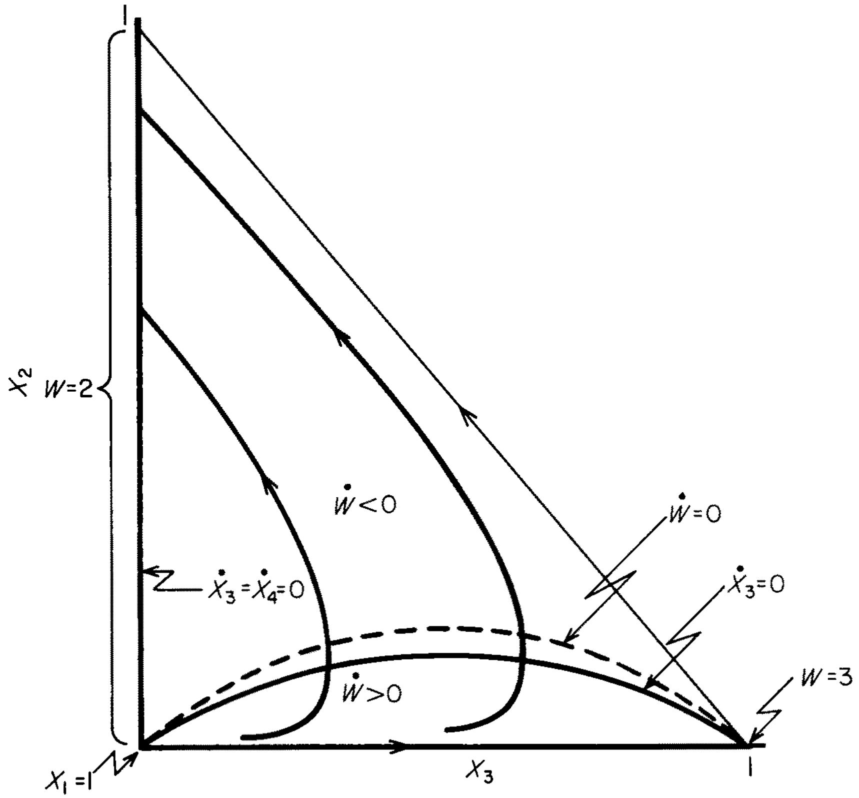

FIG. 7. Phase diagram for reduced form Prisoner's Dilemma with two mutants.

4. Related Works, Implications

A large number of authors from several disciplines have considered how natural selection might be reconciled with a variety of alternative concepts of altruism. Trivers (1971) coined the term "reciprocal altruism" to describe a process for attaining group efficient outcomes by means which essentially respect the individual as the unit of selection. Discussions of altruism, reciprocal or not, have become a central concern of sociobiology (see, e.g. Wilson, 1975: 3).

Within biology, related kinds of signaling mutants to those here are discussed by Dawkins (1982*b*: chapter 8). This comprises a theoretical discussion of how "outlaw genes" might promote their own survival at the expense of other closely related genes. Two of these hypothetical outlaw types discussed use signals to distinguish themselves from other genes. These two types are christened "armpits" and "green beards". Within pure game theory, a related contribution is due to Matsui (1988). He presents an analysis of a two-player infinitely repeated game in which only Pareto-efficient outcomes can be equilibria. It is assumed that there is a small probability that one agent's entire "supergame" strategy is revealed to the other. The proof relies on a construction reminiscent of the signaling strategy here.

For Matsui, of course, the two players are fully rational human beings. Finally, many of the ingredients used here can be found in Binmore (1988). He uses an evolutionary game argument to buttress the utilitarian outcome in a certain situation at the expense of the Nash bargaining one. His argument is analogous of that used here in section 3.2. to analyze the one-shot prisoner's dilemma.

The game theoretic issue addressed here arises repeatedly in the previous literature. Dawkins (1976: 197-202), for example, outlines a game-theoretic approach to mutual grooming. He finds two ESSs with differing payoffs, thus fitting the model here. Axelrod & Hamilton (1981) and Axelrod (1984) contain a detailed evolutionary game theoretic analysis of co-operation (see also Maynard Smith, 1984: chapter 13, for a summary of this). Much of this work assumes that the two individuals play the prisoner's dilemma not once, but repeatedly, and there is some given probability of termination at each repetition. In this context, different pure strategies turn out to generate ties in payoffs in a manner which creates difficulties with the notion of an ESS as in Definition 2. Boyd (1989), however, shows that pure strategies can be reinstated as ESSs if individuals can make mistakes with certain probabilities. For such a "supergame", indeed, always finking remains an ESS in this extended sense. A modified version of the strategy "tit-for-tat", which co-operates initially but thereafter matches the opponent's last move, can also be an ESS. If so it will yield a group preferred outcome. This model then essentially fits the framework developed here. In the original analysis the "evolution of co-operation" was hampered because "tit-for-tat" mutants were at a disadvantage playing against an initial population which always finked relative to this initial population itself. Thus Axelrod (1984: 175), for example, emphasized the need for geographical clustering of these mutants to provide a friendlier initial environment for themselves. The present analysis suggests how mutants might arise which are not at a disadvantage relative to the initial population, while still obtaining higher payoffs against one another. The "evolution of co-operation" would then by ultimately inevitable.

Finally, it should be noted that a full genetic model of the above "secret handshake" mutation is ultimately desirable. It is difficult to imagine such a mutation as being likely to arise from a change in a gene at a single locus. It is more plausible, perhaps, that the gene controlling an appropriate change in appearance is at one locus and that controlling the recognition of this change is at another. Sexual reproduction is generally believed to facilitate such compound mutations. However, the full implications of genetic recombination for the "secret handshake" mutation are not clear.

I am indebted to Robert Axelrod, W. G. S. Hines, John Maynard Smith, Eric van Damme and an anonymous referee for helpful comments.

REFERENCES

AXELROD, R. (1984). *The Evolution of Co-operation.* New York: Basic Books.
AXELROD, R. & HAMILTON, W. D. (1981). *Science* **211**, 1390-1396.
BINMORE, K. (1988). University of Michigan, CREST Working Paper 89-05.
BISHOP, D. T. & CANNINGS, C. (1978). *J. theor. Biol.* **70**, 85-124.

BOYD, R. (1989). *J. theor. Biol.* **136,** 47-56.
DARWIN, C. (1859). *The Origin of Species.* London.
DAWKINS, R. (1976). *The Selfish Gene.* Oxford: Oxford University Press.
DAWKINS, R. (1982*a*). *Current Problems in Sociobiology.* Cambridge: Cambridge University Press.
DAWKINS, R. (1982*b*). *The Extended Phenotype.* Oxford: Oxford University Press.
HINES, W. G. S. (1980). *J. appl. Prob.* **17,** 333-340.
HINES, W. G. S. (1987). *Theor. pop. Biol.* **31,** 195-272.
MALTHUS, T. R. (1798). *An Essay on Population.* London: J. Johnson.
MATSUI, A. (1988). Information leakage forces co-operation. Northwestern University, MEDS: Preprint.
MAYNARD SMITH, J. (1982). *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.
MAYNARD SMITH, J. & PRICE, G. R. (1973). *Nature, Lond.* **246,** 15-18.
MAYR, E. (1982). *The Growth of Biological Thought.* Cambridge, MA: Harvard University Press.
RILEY, J. G. (1979). *J. theor. Biol.* **76,** 109-123.
TAYLOR, P. D. & JONKER, L. B. (1978). *Math. Biosci.* **40,** 145-156.
TRIVERS, R. L. (1971). *Q. Rev. Biol.* **46,** 35-57.
VAN DAMME, E. (1987). *Stability and Perfection of Nash Equilibria.* New York: Springer.
WILSON, E. O. (1975). *Sociobiology.* Cambridge, MA: Harvard University Press.
WYNNE-EDWARDS, V. C. (1962). *Animal Dispersion in Relation to Social Behaviour.* London: Oliver and Boyd.
ZEEMAN, E. C. (1979). *Global Theory of Dynamical Systems. Lecture Notes in Math.* **819,** New York: Springer.

APPENDIX General Case

As a matter of notation:

Definition 5. *Support, best-replies*

Consider an evolutionary game as in Definition 1. Suppose $x \in \Delta^{n-1}$, then define;

$$R(x) = \{i \in \{1, \ldots, n\} | x_i > 0\} = \text{support of } x$$

and

$$S(x) = \{i \in \{1, \ldots, n\} | (\mathbf{A}x)_i = \max_j (\mathbf{A}x)_j\}$$

$$= \text{set of pure strategy best-replies to } x.$$

Clearly, if x^* is an ESS as in Definition 2, $R(x^*) \subset S(x^*)$.

Suppose the following holds:

Assumption 1 *two EESs*

Consider an evolutionary game as in Definition 1 with a non-zero number of ESSs as in Definition 2. (This number must be finite. See van Damme, 1987: 214). Take p to be any ESS which yields the maximum over ESSs of average fitness, and suppose further that there is an ESS q with strictly lower average fitness:

$$q^T\mathbf{A}q < p^T\mathbf{A}p.$$

It is intended, then, to introduce plausible mutant(s) which destroy any such inefficient q as an ESS. The simplest way of doing this might be to introduce a single monomorphic mutant which played the ESS q, as a mixed strategy, against the old non-signaling population but played the ESS p, also as a mixed strategy, against fellow signalers. This would undoubtedly achieve the desired result. Indeed, if the ESS q had previously been achieved by a monomorphic population playing it as a mixed strategy, it seems reasonable that a mutant could also use this mixed strategy against non-signalers. What seems less plausible is that the mutant would immediately stumble upon the exact mixed strategy, p, to be used in attaining the new ESS. It is perhaps more plausible to hypothesize instead the introduction of a number of mutants, each of which chooses a pure strategy against fellow signalers:

Assumption 2 *form of mutants*

Suppose the evolutionary game given in Definition 1 is augmented by the introduction of n signaling mutants. Mutant i plays the *pure strategy* i against fellow signalers, $i = 1, \dots, n$. Against non-signaling players, however, every mutant plays the inferior ESS q, as a *mixed strategy.*

Remark:

It is easy to see, given the proof below, that the number of mutants can be reduced to $|R(p)|$, that is, the number of elements in the support of p.

The effect of Assumption 2 is to convert the original game, represented by the matrix $\mathbf{A}$, as in Definition 1, into the game represented by the augmented matrix;

$$\underset{2n\times 2n}{\tilde{\mathbf{A}}} = \begin{bmatrix} \underset{n\times n}{A} & \underset{n\times n}{B} \\ \underset{n\times n}{C} & \underset{n\times n}{D} \end{bmatrix}$$

Now $B = (b_{il})$ where;

$$b_{il} = \text{payoff to old strategy } i \text{ against mutant } l$$

$$= \sum_{j=1}^{n} a_{ij} q_j = (\mathbf{A}q)_i = b_i, \quad \text{say.}$$

Furthermore, $C = (c_{kj})$ where;

$$c_{kj} = \text{payoff to mutant } k \text{ against old strategy } j$$

$$= \sum_{i=1}^{n} a_{ij} q_i = (q^T \mathbf{A})_j = c_j, \quad \text{say.}$$

Finally, $D = (d_{kl})$ where;

$$d_{kl} = \text{payoff to mutant } k \text{ against mutant } l$$

$$= a_{kl}.$$

Hence;

$$\tilde{\mathbf{A}} = \begin{bmatrix} A & b & \cdots & b \\ c^T & & & \\ \vdots & & A & \\ c^T & & & \end{bmatrix}, \quad \text{where } b = \mathbf{A}q, \text{ and } c^T = q^T \mathbf{A}.$$

The main result now follows in a purely formal fashion from Definition 2. The interpretation intended is given below. Thus:

Theorem 1: *Elimination of an inferior ESS*

Suppose an evolutionary game is described by $\tilde{\mathbf{A}}$ as above, where Definition 1 applies to $\tilde{\mathbf{A}}$. Now $[q^T, 0^T]$ is not an ESS of $\tilde{\mathbf{A}}$. However, both $[p^T, 0^T]$ and $[0^T, p^T]$ are ESSs of $\tilde{\mathbf{A}}$ (ESSs are as in Definition 2).

Proof:

As purely a matter of notation, $2n$-vectors z, say $z \in \Delta^{2n-1}$ will be written, in this proof, as $[\alpha r^T, (1-\alpha)s^T]$ where $r, s \in \Delta^{n-1}$, $\alpha \in [0, 1]$.

(a) $[q^T, 0^T]$ is not an ESS of $\tilde{\mathbf{A}}$. Note that;

$$[0^T, p^T]\tilde{\mathbf{A}}\begin{bmatrix} q \\ 0 \end{bmatrix} = [c^T, p^T\mathbf{A}]\begin{bmatrix} q \\ 0 \end{bmatrix} = c^Tq = q^T\mathbf{A}q = [q^T, 0]\tilde{\mathbf{A}}\begin{bmatrix} q \\ 0 \end{bmatrix}.$$

That is, the mutant $[0^T, p^T]$ does exactly as well against $[q^T, 0^T]$ as $[q^T, 0^T]$ does. Further;

$$[0^T, p^T]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ p \end{bmatrix} = [c^T, p^T\mathbf{A}]\begin{bmatrix} 0 \\ p \end{bmatrix} = p^T\mathbf{A}p,$$

whereas;

$$[q^T, 0]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ p \end{bmatrix} = [q^T, 0]\begin{bmatrix} b \\ \mathbf{A}p \end{bmatrix} = q^Tb = q^T\mathbf{A}q < p^T\mathbf{A}p$$

by Assumption 1. Thus the mutant $[0^T, p^T]$ does better against itself than $[q^T, 0]$ does against this mutant. Hence $[q^T, 0]$ is not an ESS of $\tilde{\mathbf{A}}$.

(b) $[p^T, 0^T]$ is an ESS of $\tilde{\mathbf{A}}$. Note that firstly since p and q are both ESSs of $\mathbf{A}$,

$$q^T\mathbf{A}p < p^T\mathbf{A}p,$$

that is, q cannot be a best-reply to p. For suppose;

$$q^T\mathbf{A}p = p^T\mathbf{A}p.$$

Then since p is an ESS,

$$q^T\mathbf{A}q < p^T\mathbf{A}q,$$

which contradicts q being an ESS. (This result is implied by Bishop & Cannings, 1978: 91.) Hence, $\forall s \in \Delta^{n-1}$

$$[0^T, s^T]\tilde{\mathbf{A}}\begin{bmatrix} p \\ 0 \end{bmatrix} = [c^T, s^T\mathbf{A}]\begin{bmatrix} p \\ 0 \end{bmatrix} = c^Tp = q^T\mathbf{A}p < p^T\mathbf{A}p.$$

Also, of course, $\forall r \in \Delta^{n-1}$,

$$[r^T, 0^T]\tilde{\mathbf{A}}\begin{bmatrix} p \\ 0 \end{bmatrix} = r^T\mathbf{A}p \leq p^T\mathbf{A}p,$$

since p is an ESS of $\mathbf{A}$. Altogether, then, $\forall \alpha \in [0, 1]$

$$[\alpha r^T, (1-\alpha)s^T]\tilde{\mathbf{A}}\begin{bmatrix} p \\ 0 \end{bmatrix} = \alpha r^T \mathbf{A}p + (1-\alpha)q^T \mathbf{A}p \leq p^T \mathbf{A}p = [p^T, 0^T]\tilde{\mathbf{A}}\begin{bmatrix} p \\ 0 \end{bmatrix},$$

and equality is only possible if $\alpha = 1$. In this case, then,

$$r^T \mathbf{A}p = p^T \mathbf{A}p.$$

Since p is an ESS of $\mathbf{A}$

$$r^T \mathbf{A}r < p^T \mathbf{A}r,$$

so that;

$$[r^T, 0^T]\tilde{\mathbf{A}}\begin{bmatrix} r \\ 0 \end{bmatrix} < [p^T, 0^T]\tilde{\mathbf{A}}\begin{bmatrix} r \\ 0 \end{bmatrix},$$

exactly as required in order that $[p^T, 0^T]$ be an ESS of $\tilde{\mathbf{A}}$.

(c) $[0^T, p^T]$ is also an ESS of $\tilde{\mathbf{A}}$. Note that, of course,

$$[0^T, p^T]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ p \end{bmatrix} = p^T \mathbf{A}p.$$

Now $\forall r, s \in \Delta^{n-1}$ and $\alpha \in [0, 1]$,

$$[\alpha r^T, (1-\alpha)s^T]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ p \end{bmatrix} = \alpha r^T b + (1-\alpha)s^T \mathbf{A}p = \alpha r^T \mathbf{A}q + (1-\alpha)s^T \mathbf{A}p,$$

where,

$$r^T \mathbf{A}q \leq q^T \mathbf{A}q < p^T \mathbf{A}p,$$

since q is an ESS of $\mathbf{A}$ and by Assumption 1. Further;

$$s^T \mathbf{A}p \leq p^T \mathbf{A}p,$$

since p is an ESS of $\mathbf{A}$. Hence;

$$\alpha r^T \mathbf{A}q + (1-\alpha)s^T \mathbf{A}p \leq p^T \mathbf{A}p,$$

with equality implying $\alpha = 0$. In this case, then,

$$s^T \mathbf{A}p = p^T \mathbf{A}p,$$

and since p is an ESS of $\mathbf{A}$,

$$s^T \mathbf{A}s < p^T \mathbf{A}s,$$

so that;

$$[0^T, s^T]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ s \end{bmatrix} < [0^T, p^T]\tilde{\mathbf{A}}\begin{bmatrix} 0 \\ s \end{bmatrix},$$

exactly as required for $[0^T, p^T]$ to be an ESS of $\tilde{\mathbf{A}}$.

Interpretation

The form of the mutants given above was motivated by considering a monomorphic population playing the old ESS, q. The above result concerning the ESSs of $\tilde{A}$ also is best interpreted as considering the introduction of monomorphic mutants which mix over all $2n$ of the strategies involved in the construction of $\tilde{A}$. This suggests a motive for analyzing the appropriate mixed strategy dynamical system. For simplicity of exposition, however, only the pure strategy dynamic is studied here.

Remarks

(1) Perhaps not every monomorphic mixed strategy mutant as above is actually possible. For example, it would not be possible for an individual to mix over possession of the signal, if this were a permanent change in appearance. Of course, being an ESS will still suffice to deter invasion by any of the remaining mutants.

(2) On the other hand, there are mutants which are not considered here. For example, take the class of mutants which do not possess the signal but which condition their behavior on its presence in an opponent. It is not claimed that the present result is robust to the exact specification of the mutants.

(3) It is readily checked that the above result is, to some extent, robust to the mixed strategy played by mutants against non-signalers. That is, if mutants use $\tilde{q}$ for this, where $R(\tilde{q}) \subset S(q)$, as in Definition 5, and $\tilde{q}$ is sufficiently close to q, then the Theorem remains true.

Consider, then, the pure strategy dynamic, as in Definition 3, for the matrix $\tilde{A}$. Note, indeed, that the coordination game of section 3.1 has the property that points near the old inefficient ESS are attracted to the new "secret handshake" ESS. This does not follow from Theorem 1. However, the following generalization applies:

Theorem 2: Pure strategy dynamics of augmented game

Suppose the evolutionary game is given by $\tilde{A}$ as above. Consider the pure-strategy dynamical system as in Definition 3. Now $[q^T, 0^T]$ is not even an attractor for there is a path from $[q^T, 0^T]$ leading to $[0^T, p^T]$ in this dynamical system. Indeed, convergence to $[0^T, p^T]$ is guaranteed if the initial point is sufficiently close to this path.

Proof:

It is convenient now to write a $2n$-vector $z \in \Delta^{n-1}$ in the form $[x^T, y^T]$, for suitable $x, y \in R^{n+}$. Now each component x_i of the vector x refers to the number of individuals playing the old pure strategy i against all comers. Each component y_i refers to the number of signaling mutants playing pure strategy i against fellow signalers, where all such mutants play the ESS q as a mixed strategy against non-signalers.

Suppose the initial point is given on the straight line between $[q^T, 0^T]$ and $[0^T, p^T]$ as:

$$[\alpha q^T, (1-\alpha)p^T], \quad \text{where } \alpha \in (0, 1).$$

It is to be shown that the solution path of the dynamical system remains on this line and converges to $[0^T, p^T]$. Now the payoff to the old strategy i is $(\mathbf{A}q)_i$ because *all* other strategies play q against the old population. The payoff to mutant j is;

$$\alpha q^T \mathbf{A} q + (1-\alpha)(\mathbf{A}p)_j,$$

because it obtains the ESS payoff for q a fraction α of the time and plays the mutant population p the remaining fraction of the time. Hence the average payoff overall is;

$$\alpha(q^T \mathbf{A} q) + (1-\alpha)[\alpha q^T \mathbf{A} q + (1-\alpha) p^T \mathbf{A} p] = \alpha(2-\alpha) q^T \mathbf{A} q + (1-\alpha)^2 p^T \mathbf{A} p.$$

It follows that the pure strategy dynamic for $[x^T, y^T]$, as in Definition 3, is given by;

$$\frac{\dot{x}_i}{x_i} = q^T \mathbf{A} q - \alpha(2-\alpha) q^T \mathbf{A} q - (1-\alpha)^2 p^T \mathbf{A} p = -(1-\alpha)^2 [p^T \mathbf{A} p - q^T \mathbf{A} q] < 0,$$

for all $i \in R(q)$. If $i \notin R(q)$ then, of course, $x_i = 0$. Thus the components of x shrink at the same percentage rate, given such an initial point. Similarly,

$$\frac{\dot{y}_j}{y_j} = \alpha q T \mathbf{A} q + (1-\alpha) p^T \mathbf{A} p - \alpha(2-\alpha) q^T \mathbf{A} q - (1-\alpha)^2 p^T \mathbf{A} p$$

$$= \alpha(1-\alpha)[p^T \mathbf{A} p - q^T \mathbf{A} q] > 0,$$

for all $j \in R(p)$. If $j \notin R(p)$ then $y_j \equiv 0$. Hence it is also true that the components of y remain in fixed proportions, given the initial point. Clearly the solution is given by;

$$x(t) = \alpha(t) q \qquad y(t) = [1 - \alpha(t)] p,$$

where,

$$\frac{\dot{\alpha}(t)}{\alpha(t)} = -(1-\alpha(t))^2 [p^T \mathbf{A} p - q^T \mathbf{A} q] < 0,$$

and $\alpha(0) = \alpha$. It follows that $\alpha(t) \to 0$ as $t \to \infty$, that is,

$$[x^T(t), y^T(t)] \to [0, p^T],$$

no matter how close $\alpha(0) = \alpha$ was initially to 1, as was to be shown.

The ESS of $\tilde{\mathbf{A}}$, $[0^T, p^T]$, must be an attractor in this dynamical system. Its "basin of attraction" is then an *open set* which includes the above path (see Zeeman, 1979). That is, ultimate convergence to $[0^T, p^T]$ is guaranteed if the initial point is sufficiently close to the above path. Of course such a result says little about the behavior of paths which start far from this particular path.