The Evolution of Altruism: Correlation, Cost, and Benefit[1]

ELLIOTT SOBER

Philosophy Department
University of Wisconsin
Madison 53706, U.S.A.

ABSTRACT: A simple and general criterion is derived for the evolution of altruism when individuals interact in pairs. It is argued that the treatment of this problem in kin selection theory and in game theory are special cases of this general criterion.

KEY WORDS: Altruism, evolution, game theory, group selection, kin selection, prisoners' dilemma.

1. INTRODUCTION

In principle, group selection, kin selection, and reciprocity (as modeled in iterated prisoners' dilemmas) provide three mechanisms that allow altruism to evolve. However, it is a matter of controversy how these three mechanisms are related. Many biologists will not touch hypotheses of group selection with a stick, but feel that kin selection theory and game theory are acceptable frameworks (e.g., Trivers 1972; Dawkins 1976; Maynard Smith 1982). Others argue that kin selection is a kind of group selection (in which the groups are composed of relatives) and/or that game theoretic models involve group selection in which the groups are of size 2 (Wilson 1980; Michod and Sanderson 1985; Wilson and Dugatkin, in press; Wilson and Sober 1988; Wilson 1990). Quite clearly, this dispute turns in part on how the concept of group selection is defined.

It is not the purpose of this paper to try to resolve this disagreement. Rather, I will describe a simple and general criterion for the evolution of altruism. I then will show how kin selection theory and game theory each explore special circumstances in which that general criterion may be satisfied. Quite apart from the taxonomy one adopts of how these mechanisms are related, it is important to identify a conceptual element that group selection theory, kin selection theory, and game theory have in common.

Most of the main ideas in this paper can be found in the biological literature (see, especially, Michod and Sanderson 1985). However, the biological literature is mathematically difficult compared with the very simple analysis I provide in what follows. Also, these sophisticated treatments have not prevented a variety of confusions from persisting. I hope that my very simple exposition will dispel some misunderstandings.

Biology and Philosophy **7**: 177–187, 1992.

When two individuals interact and one acts altruistically while the other acts selfishly, the immediate result of their interaction is that the former individual does worse than the latter. How it is nonetheless possible for the altruistic behavior to evolve is the fundamental puzzle that game theory, kin selection theory, and theories of group selection attempt to solve.

Assume that the individuals in a population are either altruists (A) or selfish (S). The behavior of altruists has two fitness consequences; altruists donate a benefit b to others and incur a cost to themselves of c. Selfish individuals make no such donations, although they can be recipients. Notice that an altruist *must* be a donor and *may* be a recipient, while a selfish individual *can't* be a donor and *may* be a recipient.

Altruism will evolve when the fitness of altruists [w(A)] exceeds the fitness of selfish individuals [w(S)]; this is the standard assumption of heritability that is used in phenotypic models of natural selection. We will see that the criterion for $w(A) > w(S)$ depends on two quantities: the correlation between interactors and the cost/benefit ratio.

2. THE UNSTRUCTURED CASE: EVERYONE INTERACTS WITH EVERYONE

Suppose that the population is composed of n altruists and some number of selfish individuals. Suppose that each altruist donates a benefit b *to every other individual in the population.* In this case, the fitnesses of the two traits are:

$w(A) = (x\text{-}c) + (n\text{-}1)b$
$w(S) = x + nb.$

In these expressions, x is a "baseline fitness." When altruists reduce their fitness by c units, this is a reduction from the fitness they would have had if they had not donated. In both expressions, the first addend describes the effect on the individual of its own phenotype, while the second describes the effect on the individual of the behaviors of others. An altruist receives donations from the (n-1) *other* altruists; a selfish individual receives donations from *all* n altruists.

Simple algebra shows that

(1) $w(S) > w(A)$ if and only if $c + b > 0$.

This means that if donation confers a genuine benefit on the recipient ($b > 0$) and if it entails a genuine cost to the donor ($c > 0$), then altruism cannot evolve. Proposition (1) also allows for altruism to decline in frequency when $c < 0$. A trait that benefits the donor will decline in frequency if it benefits the recipient *more* (Wilson 1990).

Alternatively, suppose that each altruist provides a *public good.* This means that the altruist's donation benefits everyone in the group, including the altruist who makes the donation. In this case, the fitnesses will be slightly different, but the prospects for the evolution of altruism are not much improved. The fitnesses now are:

$w(A) = (x\text{-}c) + nb$
$w(S) = x + nb.$

It follows that

$w(S) > w(A)$ if and only if $c > 0$.

3. A GENERAL CRITERION WHEN THERE IS PAIR-WISE INTERACTION

Now let's introduce some structure into this population. Let the individuals pair up; whether this happens at random, or there is a tendency for similar individuals to pair with each other, I will discuss shortly. The paired individuals then engage in some sort of behavioral interaction that affects their fitnesses. In this interaction, an individual's fitness is influenced both by its own phenotype and by the phenotype of the individual with which it is paired. The payoffs to row are as follows:

		You are paired with	
		A	S
You are	A	$x\text{-}c+b$	$x\text{-}c$
	S	$x+b$	x

When selfish individuals pair with each other, each receives a baseline fitness of x. When altruists pairs with selfish individuals, the altruists suffer the cost (c) of donation, and so their fitness is reduced to $x\text{-}c$; however, the selfish individuals in such pairs receive the benefit (b) donated by their associates and do not incur the cost of donation themselves. So selfish individuals who are paired with altruists have a fitness of $x+b$. Lastly, we must consider the fitness that altruists have when they pair with each other. In this case, altruists pay the cost of donation but also receive a benefit from the donation of their associates; so altruists paired with altruists have a fitness of $x\text{-}c+b$.

Notice that whenever an altruist and a selfish individual pair, the altruist always does worse than the selfish individual ($x\text{-}c < x+b$, on the assumption that $b+c > 0$). However, we now will see that this fact does not settle whether altruism is overall less fit than selfishness when individuals interact after forming into pairs. A population with this structure differs in a fundamental way from the unstructured everyone-interacts-with-everyone arrangement described by (1).

Given the payoffs just described, the fitnesses $w(A)$ and $w(S)$ are as follows:

$$(2)\quad \begin{aligned} w(A) &= (x\text{-}c+b)P(A/A) + (x\text{-}c)P(S/A) \\ w(S) &= (x+b)P(A/S) + (x)P(S/S). \end{aligned}$$

Here P(S/A) is the probability that one individual in the pair is S, given that the other is A. Simple calculation shows that

(3) $w(A) > w(S)$ if and only if $P(A/A) - P(A/S) > c/b$.

P(A/A)-P(A/S) is a familiar statistical quantity; it is the *correlation* of the phenotypes of the interacting individuals. If pairs form at random, then $P(A/A) = P(A/S) = P(A)$ and the inequality reduces to $0 > c/b$. This means that one or the other of b and c must be negative if altruism is to evolve. This would happen if the "benefit" given away were actually a harm, or if the "cost" to self were negative (i.e., the effect on self were positive). But if there is genuine cost to donor and genuine benefit to recipient, altruism cannot evolve when pairs form at random.

At the other extreme is the case in which like always associates with like. If $P(A/A) = 1$ and $P(A/S) = 0$, the inequality becomes $b > c$. In this case, altruism evolves precisely when the benefit to recipient exceeds the cost to donor.

Between these two extremes are different degrees of positive association. How should these be represented? Borrowing an idea from the theory of inbreeding, we can describe the frequencies of the three sorts of pairs as follows:

$P(AA) = fp + (1-f)p^2$
$P(AS) = (1-f)2pq$
$P(SS) = fq + (1-f)q^2$.

Here $p = P(A)$ and $q = P(S)$. Just as in the theory of inbreeding, we imagine that f of the individuals with a given trait pair with individuals like themselves, and the remaining $(1-f)$ individuals with that trait pair up at random. If $f = 1$, we have the case in which like always pairs with like. If $f = 0$, then pair formation is at random.

Using these expressions for the frequencies of the pairs, we can rewrite (3) as follows:

$P(A/A) - P(A/S) = P(AA)/P(A) - [1-P(SS)/P(S)]$
$= [fp + (1-f)p^2]/p - 1 + [fq + (1-f)q^2]/q > c/b$.

This simplifies to

(4) $f > c/b$.

The general condition for the evolution of altruism when individuals pair is that the degree of positive association between the phenotypes of the paired individuals must exceed the cost/benefit ratio (Eshel and Cavalli-Sforza 1982).

This means that when an altruistic behavior implies a particular cost/benefit ratio, conditions (3) and (4) describe how much correlation there must be between interactors if the behavior is to evolve. The more costly the donation (for a fixed benefit), the harder it is for the trait to evolve. And if the cost exceeds the benefit, c/b will be greater than 1. Since correlations have 1 as their maximum value, it is impossible for this sort of *hyperaltruism*, as we might call it, to evolve. Few of us would die to make someone smile. If our behavior were under the control of the kind of selection process described by (3) (a controversial assumption, to say the least), the reluctance to engage in hyperaltruism would be perfectly intelligible.

Let's compare criteria (1) and (3). If c and b are both positive, then altruism cannot evolve if everybody interacts with everybody. However, altruism may be able to evolve when populations are subdivided into interacting pairs in which the interactors are correlated. Population structure makes all the difference.

4. KIN SELECTION

One way to effect the correlation between interactions needed for altruism to evolve is to have relatives form up into groups. This is the basic idea in Hamilton's (1964) theory of kin selection. When groups are composed of relatives, altruism can evolve because (or to the extent that) relatives resemble each other. It is not the common pedigree of the interactors that is decisive; rather, what is crucial is that kin pairing with kin can ensure that like pairs with like.

Condition (3) describes the general criterion for altruism to evolve when the population is structured into pairs. Now I shall explore the special case in which the pairs are full sibs. Hamilton characterized the coefficient of relatedness between full sibs as $r = 1/2$. Hamilton's inequality for the evolution of altruism in kin selection when the paired individuals are full sibs is $1/2 > c/b$. We now will see that interactions exclusively between full sibs are characterized by P(A/A)-P(A/S) = 1/2, if the population is randomly mating and inheritance follows a simple symmetrical pattern.

Let us suppose that AxA parents produce 100% A offspring, AxS parents produce 50% A and 50% S offspring, and that SxS parents produce 100% S offspring. This is sometimes called a haploid sexual mode of inheritance. If p is the frequency of A among the parents and mating is at random, then the frequency (before selection) of the three types of offspring sib groups is

$$P(AA) = p^2 + pq/2$$
$$P(AS) = pq$$
$$P(SS) = q^2 + pq/2.$$

This departure from Hardy-Weinberg frequencies is the result of the fact that sibs tend to resemble each other. However, note that the frequencies of A and S among the sibs (before selection) does not depart from the parental frequencies (P(A) = P(AA) + P(AS)/2 = p and P(S) = P(SS) + P(AS)/2 = q. These unconditional probabilities allow us to define the following conditional probabilities:

$$P(A/A) = P(AA)/P(A) = (p^2 + pq/2)/p = p + q/2$$
$$P(S/S) = P(SS)/P(S) = (q^2 + pq/2)/q = q + p/2.$$

Substituting these conditional probabilities into (2), we obtain

$$w(A) = (x+b-c)(p + q/2) + (x-c)(q/2)$$
$$w(S) = (x+b)(p/2) + x(q + p/2).$$

Simple algebra then entails that

(5) When interactions are exclusively between full sibs, $w(A) > w(S)$ if and only if $1/2 > c/b$.

It is worth noting that the evolution of altruism in the model just considered is governed by a *frequency independent criterion.* With random mating and symmetric inheritance, P(A/A)-P(A/S)=1/2, regardless of whether A is common or rare. If x, c, and b have fixed values, then either A cannot increase under selection, or it goes all the way to fixation.[2] For the sake of contrast, I describe in the Appendix a simple genetic model of this situation in which the criterion for the evolution of altruism is *frequency dependent.*

Informal discussion of Hamilton's inequality $r > c/b$ for the case of full sibs often includes remarks like "full sibs share half their genes." Although this will be true in special cases, it is not true in general. A population in which natural selection has destroyed lots of variation may be such that all individuals are *very* similar to each other, and full sibs are even more so (Dawkins 1979). In addition, for altruism to evolve, it really does not matter how overall similar full sibs are to each other. What matters is the quantity P(A/A)-P(A/S), where A and S are the two phenotypes (or the genes coding for them); the rest of the genome is quite irrelevant.

Two other imprecise renditions of Hamilton's inequality bear mention. The first is the suggestion that what is criterial is the probability that one sib will be altruistic if the other one is. I hope it is clear that P(A/A) is *not* the relevant quantity. It is true that in the model we have considered, $P(A/A) = p + q/2$ (if selection is ignored), which approaches 1/2 as p approaches 0. However, P(A/A) increases as A evolves. As noted above, proposition (5) gives a frequency *in*dependent criterion for the evolution of altruism.

The second mistake is the idea that the criterial quantity is the difference between the probability that two sibs will be altruists and the probability that two randomly selected members of the population will be altruists. If altruism is common, it will be impossible for this quantity to be as large as 1/2; yet, for all that, P(A/A)-P(A/S) = 1/2 for full sibs, when mating is random and inheritance is symmetrical.

5. RECIPROCAL ALTRUISM IN ITERATED PRISONERS' DILEMMAS

In the pair-wise interactions considered so far, individuals do *not* pair at random (if altruism is to evolve) and they interact with each other *once.* I now will consider a different situation, one in which the individuals pair at random but interact with each other repeatedly. The pairs of individuals play an n-round iterated Prisoners' Dilemma. On each move they can either cooperate (be altruistic) or defect (be selfish).[3] The payoffs on each move are given in the previous payoff matrix. How well an individual does in this n round game depends on the strategy he or she follows and on the strategy followed by the other individual in the pair.

One possible strategy an individual might follow is to act selfishly on every move. This unconditional strategy is called ALLD ("always defect"). Individuals may follow other, more complicated, strategies. Axelrod (1984)

examines the properties of the strategy TIT-FOR-TAT (TFT). An individual playing TFT will cooperate (i.e., be altruistic) on the first move and then will do on the next move whatever his or her partner did on the previous move. So when the two individuals in a pair both play TFT, both cooperate on every move.

TFT is a strategy that involves *reciprocity* (Trivers 1972). If the opponent cooperates, TFT does the same on the next move; but if the opponent defects, TFT follows suit. TFT involves a form of conditional altruism.

Let us consider a population in which everyone follows either TFT or ALLD. Individuals pair and then play against their partners for n moves. The three sorts of pairs and the sequences of moves that occur within them are as follows:

TFT A A A ...
TFT A A A ...

TFT A S S ...
ALLD S S S ...

ALLD S S S ...
ALLD S S S ...

Notice that even if individuals pair at random, there still is an enormous amount of correlation between altruistic and selfish *behaviors* (Michod and Sanderson 1985; Wilson and Dugatkin, in press). The only time altruistic and selfish behaviors co-occur is during the first round of a game between someone playing TFT and someone playing ALD.

We now can define the fitness of each strategy as follows:

$w(\text{TFT}) = n(x+b-c)\text{P(TFT/TFT)} + [x-c + (n-1)x]\text{P(ALLD/TFT)}$
$w(\text{ALLD}) = nx\text{P(ALLD/ALLD)} + [x+b + (n-1)x]\text{P(TFT/ALLD)}.$

If pairs form at random, P(TFT/TFT) = P(TFT/ALLD) = p and P(ALLD/ALLD) = P(ALLD/TFT) = q. In this case w(TFT) > w(ALLD) if and only if

$$n(x+b-c)p + [(x-c) + (n-1)x]q > nxq + [x+b + (n-1)x]p.$$

This simplifies to

(6) w(TFT) > w(ALLD) if and only if $p(n-1) > c/b$.

Notice that for fixed benefits and costs, whether TFT will be fitter than ALLD depends on the frequencies of the strategies and on the length of the game. In particular, making TFT common (increasing p) and making the game longer (increasing n) both favor the evolution of TFT.

Let's consider an example. Suppose x=1, c=1, and b=4; then the payoff matrix for each move of the n round game becomes:

		You are paired with	
		A	S
You are	A	4	0
	S	5	1

If there are fifteen rounds in each pair-wise interaction (n=15), criterion (6) becomes w(TFT) > w(ALLD) if and only if $p > 1/56$. When TFT is very rare, it cannot evolve, but once it crosses the threshold of $p = 1/56$, it goes all the way to fixation.

We can calculate the payoffs to each strategy in the fifteen round game from the above payoff matrix that describes the consequences of each move:

		You are paired with	
		TFT	ALLD
	TFT	60	14
You are			
	ALLD	19	15

The accompanying figure describes the fitness functions for the two strategies entailed by this payoff matrix. Note that there is an (unstable) equilibrium point at $p = 1/56$.

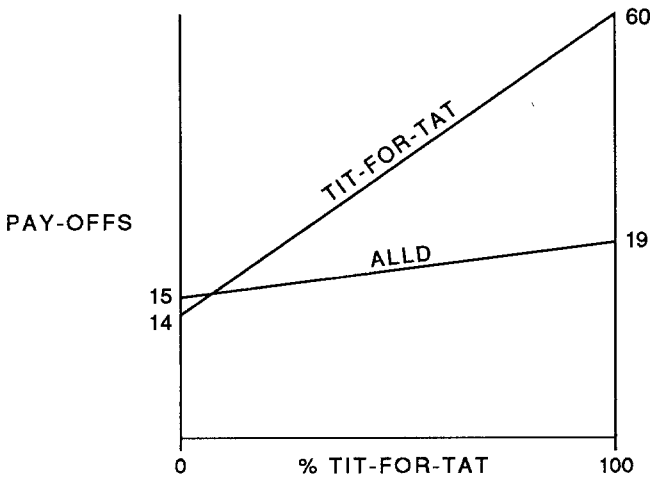

Fig. 1.

Axelrod (1984) interprets the facts represented in Figure 1 to mean that TFT can evolve when a population of ALLD players is invaded by a *cluster* of TFTers.[4] A single mutant TFTer cannot cause the trait to evolve; but if several mutant or migrant TFTers enter the ALLD population together, this may create a frequency of TFT greater than the threshold value of p.

Notice that when n= 1, criterion (6) entails that altruism cannot evolve (since $c > 0$). This describes the case in which there is random pair formation and the individuals interact only once. But when the interactions are repeated, a cooperative strategy like TFT can evolve even when pair formation is at random.

Criterion (3) says that if altruists and selfish individuals pair at random, then altruism cannot evolve (assuming that $c,b > 0$). But criterion (6) says that if

TFTers and ALLDers pair at random, TFT can evolve. There is no contradiction here, since when TFTers and ALLDers pair at random in extended games, there is a correlation between altruistic and selfish *behaviors*. Once again, the degree of positive association between the behaviors of the two interactors is of the essence in determining whether altruism will evolve.

6. CONCLUDING REMARKS

Proposition (3) specifies a general criterion for altruism to evolve when the population is structured into pairs. Proposition (5) describes the special case in which the paired individuals are full sibs. Proposition (6) describes the special case in which individuals pair at random and play TFT or ALLD in an n round prisoners' dilemma.

Proposition (3) easily can be generalized to the case in which groups are of size m (Wilson 1980). And even for the case considered here of $m = 2$, there is more to kin selection and game theory than the two cases described in propositions (5) and (6). In addition, I have assumed that the costs and benefits entailed by a behavior are independent of what the other player in the pair does; an altruistic act costs the donor c units of fitness, regardless of whether the other player behaves altruistically or selfishly; this is the simple case of "additive" rather than "synergistic" payoffs (Maynard Smith 1989, p. 168). In spite of all these limitations, I hope that the preceding exploration has been instructive. They illustrate how the evolution of altruism depends on the costs and benefits of the behaviors considered and on the degree of correlation that obtains among the interacting individuals. This is the fundamental criterion that kin selection, reciprocal altruism, and group selection all must satisfy.

APPENDIX

Section 4 described the evolution of altruism when interactions are between full sibs and inheritance is governed by a set of symmetric phenotypic rules. I now will derive a criterion for the full sib case in which altruism is controlled by a single dominant gene.

Consider a diploid population in which individuals who are *aa* or *as* are altruistic (i.e., they have the *A* phenotype) and individuals who are *ss* are selfish (*S*). As always, the *a* gene will increase in frequency precisely when $w(a) > w(s)$. The allelic fitnesses are as follows:

$$w(a) = P(a)w(aa) + P(s)w(as) = w(A)$$
$$w(s) = P(s)w(ss) + P(a)w(as) = P(s)w(S) + P(a)w(A).$$

This means that $w(a) > w(s)$ if and only if $w(A) > w(S)$. The altruistic gene (a) evolves precisely when the altruistic phenotype (A) has the higher fitness.

It follows that $w(a) > w(s)$ precisely when $P(A/A) - P(A/S) > c/b$. We now

need to evaluate P(A/A) - P(A/S) when interactions are between full sibs. The value of each term is frequency dependent:

P(A/A) = 1 when *a* is common.
P(A/A) = 1/2 when *a* is rare.
P(A/S) = 3/4 when *a* is common.
P(A/S) = 0 when *a* is rare.

P(A/A) - P(A/S) = 1/2 when the *a* gene is rare, but P(A/A) - P(A/S) = 1/4 when *a* is common. So the criterion for altruism to evolve is *frequency dependent* (but compare Dawkins 1976, p. 288).

This means that the model has three possible solutions. If $1/2 < c/b$, selfishness will go to fixation; if $1/2 > c/b > 1/4$, the two traits evolve to a stable polymorphism; *if* $1/4 > c/b$, altruism goes to fixation.

NOTES

[1] My thanks to James Crow, Carter Denniston, Lee Dugatkin, David Wilson, and an anonymous referee of this journal for helpful discussion.

[2] Of course, if c and b change values as A evolves, then the population may evolve to a stable polymorphism. The point is that with constant costs and benefits and the mode of inheritance specified, no stable polymorphism is possible.

[3] Wilson and Dugatkin (in press) observe that there is an equivalence between "cooperation" as discussed in game theory and "altruism" as used in kin selection theory, contrary to the distinction that some authors have drawn between the two concepts (e.g., Trivers 1985).

[4] Axelrod (1984) maintains that TFTers invading a population of ALLDers must interact with each other more often than would occur if interactions were at random. As the above example shows, this is not necessary. See Sober (forthcoming) for further discussion.

REFERENCES

Axelrod, R.: 1984, *The Evolution of Cooperation*, Basic Books, New York.

Dawkins, R.: 1976, *The Selfish Gene*, 2nd edition, 1989, Oxford University Press, Oxford.

Dawkins, R.: 1979, 'Twelve Misunderstandings of Kin Selection', *Zeitschrift für Tierpsychologie* **51**, 184–200.

Eshel, I. and L. Cavalli-Sforza: 1982, 'Assortment of Encounters and Evolution of Cooperativeness', *Proceedings of the National Academy of Sciences USA* **79**, 1331–1335.

Hamilton, W.: 1964, 'The Genetical Evolution of Social Behavior', *Journal of Theoretical Biology* **7**, 1–52.

Maynard Smith, J.: 1989, *Evolutionary Genetics*, Oxford University Press, Oxford.

Michod, R. and M. Sanderson: 1985, 'Behavioral Structure and the Evolution of Cooperation', in J. Greenwood and M. Slatkin (eds.), *Evolution: Essays in Honor of John Maynard Smith*, Cambridge University Press, Cambridge.

Sober, E.: forthcoming, 'Stable Cooperation in Iterated Prisoners' Dilemmas', *Economics and Philosophy*

Trivers, R.: 1972, 'The Evolution of Reciprocal Altruism', *Quarterly Review of Biology* **46**, 35–57.

Trivers, R.: 1985, *Social Evolution*, Benjamin Cummings, San Francisco.
Wilson, D.: 1980, *The Natural Selection of Populations and Communities*, Benjamin Cummings, San Francisco.
Wilson, D.: 1990, 'Weak Altruism, Strong Group Selection', *Oikos* **59**, 135–140.
Wilson, D. and L. Dugatkin, in press, 'Nepotism vs. Tit-for-Tat or, Why Should You Be Nice to Your Rotten Brother?', *Evolutionary Ecology*.
Wilson, D. and E. Sober, 1988, 'Reviving the Superorganism', *Journal of Theoretical Biology* **136**, 337–356.