Canadian Journal of Philosophy
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rcjp20

What Is Evolutionary Altruism?

ELLIOTT SOBER [a]
[a] University of Wisconsin, Madison , Madison , WI , 53711 , U.S.A.
Published online: 01 Jul 2013.

To cite this article: ELLIOTT SOBER (1988) What Is Evolutionary Altruism?, Canadian Journal of Philosophy, 18:sup1, 75-99

To link to this article: http://dx.doi.org/10.1080/00455091.1988.10715945

CANADIAN JOURNAL OF PHILOSOPHY
Supplementary Volume 14

What Is Evolutionary Altruism?

ELLIOTT SOBER
University of Wisconsin, Madison
Madison, WI 53711
U.S.A.

In this paper I want to clarify what biologists are talking about when they talk about the evolution of altruism. I'll begin by saying something about the common sense concept. This familiar idea I'll call 'vernacular altruism.' One point of doing this is to make it devastatingly obvious that the common sense concept is very different from the concept as it's used in evolutionary theory. After that preliminary, I'll describe some features of the evolutionary concept. Then I'll conclude by briefly considering what explanatory relation might obtain between vernacular altruism and evolutionary altruism.

Although the points I'll make are rather elementary ones, their interest is not restricted to those who have never heard of the evolutionary problem. The reason for this is that there is some amount of confusion about evolutionary altruism among evolutionary biologists themselves. Sociobiologists sometimes confuse vernacular and evolutionary altruism, as when they argue that people cannot really be altruists in the vernacular sense, on the grounds that evolutionary altruism cannot be a reality.[1] It also is common for biologists to think that Trivers' (1971) concept of reciprocal altruism describes a form of evolutionary altruism. My view is that Trivers' concept does not describe a form of evolutionary altruism at all. The idea that 'reciprocal altruism isn't altruism' may sound like a contradiction, but it is an idea I will defend in what follows. And lastly, there

1 See Kitcher (1985), chapter 11 for discussion of this error.

is a paradox that is absolutely central to the evolutionary concept, one which has not been widely appreciated.

Another reason for reviewing some of these ideas is that they directly parallel an idea that social scientists have thought about a great deal. Although vernacular and evolutionary altruism are quite separate matters, their similarities are very much in evidence when we consider what students of game theory call the tragedy of the commons (or the prisoners' dilemma). So besides separating biology from the social sciences in one sense, I want to bring them together in another.

I Vernacular altruism

The first and most obvious difference between the vernacular and the evolutionary concept of altruism is this: To be a vernacular altruist, you have to have a mind. But biologists can discuss the question of evolutionary altruism for any organism you please, whether it has a mind or not.

The reason I say that a mind is essential for the common sense concept is that vernacular altruism has to do with motives. Doing someone a good turn is not definitive of this sort of altruism. If I aim at harming you but by mistake do you some good, that does not make me an altruist. Likewise, if I aim at helping you but my plans get messed up, I nevertheless may be an altruist. So altruism, whatever else it is, has to do with the motive of benefitting others.

The second simple feature of the common sense concept that we should note is that the aimed for benefits do not have to be reproductive benefits. If I know that you love to play the piano, I may give you a volume of Beethoven sonatas out of the goodness of my heart. I am an altruist here, but the good I have done you does not enhance your evolutionary fitness. In fact, it may be true that time at the piano is time away from reproduction; so in love are you with the piano, that you would rather play the piano than make babies. If so, my gift diminishes your prospects for reproductive success. But I may have been a vernacular altruist nonetheless.

The third component of this familiar concept is a little less obvious. If I give you the volume of sonatas out of the goodness of my heart, I may thereby count as an altruist. Now suppose that unbeknownst to me someone else gives you *two* volumes of sonatas. This donor has given away more than I have. We might want to say that he behaved *more* altruistically than I did. Notice that this is a comparative judgment. My present point, though, is that this comparative claim does not show that I am not an altruist.

Vernacular altruism is an 'absolute' concept, not a comparative one. An altruist is someone who acts from certain sorts of motives. It follows that whether *I* am an altruist does not conceptually depend on what *you* do or on what your motives are. Altruism is an intrinsic property. It's more like the concept of being a millionaire than it is like the concept of being rich.

I have noted three properties of our common sense concept. It is essentially psychological. It does not essentially involve reproduction. And it is not essentially comparative. This last point, recall, does not mean that we never say that some people are more altruistic than others. Rather, the idea is that in calling people altruists, we are making a comment on their motives, not comparing their motives with those of others.

I so far have been working with the idea that altruists are people who act on the basis of their desire to help others. However, a moment's thought shows that this is not sufficient, even if it is necessary. I may give you some money because I want you to have it. But if my want is itself a consequence of some selfish desire, then I will not be an altruist. For example, we do not describe ordinary buying and selling as displays of altruism. Yet notice that in voluntary exchange, each party wants the other to have the goods or the cash. If we interrupt an exchange of this sort and ask – do you really want the other person to have this thing? – each party would sincerely answer 'yes.' But altruism is not involved, because each has this want only because it is a means to the selfish end of getting the cash or the goods.

What, then, is the extra ingredient? An altruist, it would seem, must not just have an other-directed desire, but must have this desire in a noninstrumental way. The good of the other must be an end, not just a means, to some selfish satisfaction. But here we seem

to run up against a banal truism: people want to have their desires satisfied. The altruist wants to help others. The selfish individual wants to keep the cookies for himself. But both, in so far as they engage in rational deliberation, select actions that maximize their chances of getting the most of what they want. Does this mean that vernacular altruism is really an illusion – that the distinction we wish to draw between genuine other-directedness and genuine selfishness dissolves?

This question I will not try to answer here.[2] However, I will note two constraints that an adequate explanation of the difference between vernacular altruism and selfishness must obey. First, the distinction must not run afoul of the truism that people act so as to satisfy the desires they have. That people act on the basis of their own desires is a fact about the *subject* of desires. But this truism about the subject of desires is quite separate from the question of what the *contents* of desires are. Whether I am an altruist concerns *what* I want; the issue is not decided by the obvious fact that it is *I* who does the wanting. The second constraint that an adequate account must respect is that selfish actions can sometimes include motives that involve the welfare of others. This is the point illustrated by the example of buying and selling. We cannot conclude that people are never altruistic because they always act so as to satisfy their own desires; but neither can we conclude that people are sometimes altruistic just because their preferences include benefitting others.

II Darwinian selection

I now want to review some simple facts about Darwinian selection, ones that will allow the issue of evolutionary altruism to emerge clearly. I said in the previous section that vernacular altruism is essentially psychological, not essentially reproductive, and not essentially comparative. Evolutionary altruism is just the opposite:

---

2 I have attempted to do so, however, in Sober (forthcoming).

	Vernacular Altruism	Evolutionary Altruism
Essentially Psychological	YES	NO
Essentially Reproductive	NO	YES
Essentially Comparative	NO	YES

The first two contrasts may be sufficiently obvious. Evolutionary altruism can occur in organisms that don't have minds; and evolutionary altruism involves the donation of reproductive benefits. Evolutionary altruism has to do with the reproductive consequences of behavior, not with the proximate mechanism (psychological or otherwise) that guides that behavior.[3] This is why the concept of evolutionary altruism can apply to creatures with minds as well as to those without.

The third contrast may be a little less transparent. But before it can be clarified, we must review some fundamental facts about how Darwinian selection works.

Let us imagine that there are two kinds of organisms in a single population. We imagine that the two characteristics are heritable. All this means is that parents tend to resemble their offspring. This may be because parents transmit genes to their offspring; or it may be because parents teach their children to be like them. The mechanism of inheritance does not matter; ony the fact of heritability is essential.[4]

3 See Sober (1985) for discussion of the difference between what Ernst Mayr has called 'proximal' and 'ultimate' explanations of biological traits.

4 Here I use 'heritability' in a sense that is broader than that customary in population genetics. The genetical concept is intended to isolate the correlation of parents and progeny attributable to genetic transmission. See Falconer (1981) for discussion.

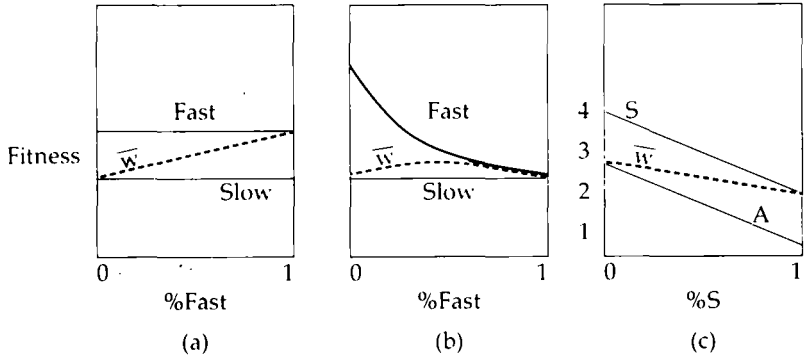
We imagine further that the two traits have different consequences for survival and reproduction. That is to say, we imagine that one of the traits is fitter than the other. We now have the preconditions for a process of Darwinian selection – heritable variation in the fitnesses of organisms.

So as to make this somewhat abstract formulation more concrete, let us imagine that we are talking about a herd of deer. The two traits are Fast and Slow. Sexual reproduction complicates our simple picture of what heritability means here – if offspring are to resemble their parents, what should happen when one parent is fast and the other is slow? To avoid this complication, let us imagine that the organisms reproduce by asexual cloning. Running speed is unerringly transmitted by the simple rule of like reproducing like.

How are we to compare the fitnesses of the two traits? I am interested in how the two traits allow organisms to avoid being caught and eaten by predators. There are several different fitness relationships we can consider.

First, let's imagine that your chance of being caught is simply determined by whether you are fast or slow. That is, we are imagining that your vulnerability to predators is not affected by whether you live in a fast or a slow herd, or whether the speed you happen to have is common or rare. In this case the fitness relationship of the two traits is frequency independent, as shown in Figure 1a.

Figure 1



What will happen in a population of Slow individuals, if a Fast mutant (or migrant) is introduced? The newcomer will be fitter than the other individuals and so will be more reproductively successful. In consequence, the Fast trait will increase in frequency. In the next generation, it will still be true that Fast individuals are on average fitter than Slow ones, so the trait will increase in frequency once again. This will continue until Fast goes to 100% representation in the population.

At the beginning of the process, all the deer were Slow; at the end, all are Fast. Given our assumption about how the predators behave, the individuals in the population are better off at the end than the individuals were at the beginning. The average fitness of the organisms in the population (called '$\bar{w}$') is represented in Figure 1a by a dotted line. Notice that the process I've just described leads to an increase in this quantity.

This quantity measures how fit, on average, the individuals in a population are. But it also can be taken to measure the welfare of the group itself. Each individual has a probability of being killed by the predator; if all individuals are killed, the group goes extinct. The selection process we have just described, it would seem, has provided the group with an advantage. By increasing the average level of fitness of individuals, selection has also benefitted the group.

We now need to see that increases in $\bar{w}$ and group advantages are not necessary consequences in Darwinian selection. We can see this by asking the following question: what was the essential feature of this selection process that allowed Fast to supplant Slow?

The answer (assuming heritability as we have done all along) is simply that *Fast is fitter than Slow*. This comparative fact suffices. But a few changes in the graph shown in Figure 1a will allow us to see the Fast can replace Slow without $\bar{w}$ ending up higher at the end of the process than it was at the beginning.

Let us suppose, to modify our example, that Fast individuals are always better off than Slow ones, but that the advantage importantly depends on the rarity of Fast individuals. Predators prefer to chase down slow*er* individuals. It isn't that they are too slow to catch the fast ones; it's that they are too lazy to bother, when slower prey present themselves. When Fastness is rare, Fast individuals do enormously better than Slow ones. But when Fastness is very common,

the advantage is slight. And when Fastness has gone to 100%, predators catch them as readily as they caught the Slow ones when the Slow ones were the only things around to eat; the predators just have to run a little faster to do this, but this is something easily within their grasp. This fitness relationship is shown in Figure 1b. Notice that the fitnesses are frequency dependent and that $\bar{w}$ is no higher at the end of the process than it was at the beginning.

In both Figure 1a and 1b, Fast is fitter than Slow. This comparative fact is enough to ensure in both cases that Fast replaces Slow. The figures differ, however, in the question of what happens to average fitness. In Figure 1a, it goes up; in Figure 1b, it rises momentarily, only to fall back to where it began.

A third example will illustrate this point in an even more extreme way. Let us consider two traits S and A, whose fitnesses are depicted in Figure 1c.[5] What will happen when an S individual is dropped into a population of A individuals? Since S is fitter than A, S will increase in frequency. In the next generation, the same fitness relationship obtains, so S continues to increase. The process will take S all the way to 100%. But notice that $\bar{w}$ steadily declines. The organisms at the end of the process are less fit than the organisms in the beginning. It is important to grasp the bleakness of the process depicted in Figure 1c. Natural selection can lead a population right to extinction. The fitter replace the less fit, and the whole process plummets downhill. If Figure 1a portrays an optimistic vision of selection the improver, Figure 1c provides a pessimistic picture of selection the destroyer.

The three figures have in common the thing that is fundamental to Darwinian selection – *comparative* fitness determines the population's trajectory. This leaves totally unspecified what happens to *absolute* fitness along the way; it is with respect to this quantity ($\bar{w}$) that the three graphs differ.

Figure 1c depicts the essentials of the concepts of evolutionary selfishness (S) and altruism (A). We can interpret this graph as showing that there are two causal factors that affect an individual's fitness.

---

5 Ignore the numbers labelling the y-axis in Figure 1c for now.

First, it is better to be selfish than to be altruistic. Second, it is better to live among altruists than among selfish individuals. Altruists thus provide a group advantage – they benefit those with whom they live, even though altruists would be better off being selfish.

So it is nice to have altruists around. But the fact of the matter is that Darwinian selection predicts that there should be no such thing. Selfish spitefulness will triumph: a trait that makes things worse for everyone will spread to fixation, as long as it makes things worse for nonbearers of the trait than it does for bearers of the trait. Imagine for example, a trait in a plant population that causes its bearer to leach a toxic chemical into the soil. As long as the poison hurts nonbearers of the trait more than it hurts bearers of it, the trait will spread. The mirror image is that a trait that boosts everyone's reproductive prospects cannot evolve, if it benefits nonbearers more than it benefits bearers. Imagine a trait that causes the plants that have it to leach an insecticide into the soil. If nonbearers of the trait are benefitted more – either because the chemical makes them more immune or because nonleachers do not incur the energetic cost of providing the chemical – the trait cannot evolve by Darwinian selection.

The definition of altruism I have given is essentially comparative. An altruistic trait is one that is related to the alternative trait (which we call 'selfish') by the fitness function shown in Figure 1c. Within a group, selfish individuals do better than altruists, but everybody benefits in a group by having lots of altruists around.

In this respect, evolutionary altruism differs from the vernacular variety. Consider a trait that leads individuals who have it to give away one unit of benefit to each of the individuals with whom they live. Is this trait an instance of evolutionary altruism? No answer can be given until the alternative traits are specified. If the other individuals in the population give away no benefits at all, then the single unit donors are altruists. If, on the other hand, the other individuals give away two units of benefit, than the single unit donor is selfish.

An immediate consequence of this example is that we should not equate altruism with donation. In a population of single unit donors and double unit donors, both traits involve donation, but only one

of them is altruistic. In a sense, every altruist is a donor, but not every donor is an altruist.

This helps show why Trivers' (1971) idea of reciprocal altruism really does not involve evolutionary altruism at all. Let's imagine a population of beavers who cooperate to build a dam. The dam is very important to the beaver way of life, but what is to prevent cheating beavers from enjoying the benefits of the dam without helping to build it? As stated so far, the answer is *nothing*. If the population consists of two types of individuals – one helps build and the other does not – and both can enjoy the benefits of the dam once it exists, we have an example of altruism and selfishness. Darwinian selection should eliminate the builders, perhaps to the detriment of builders and nonbuilders alike.

But suppose the traits present in the population are different. Let us imagine that the builders are able to prevent the nonbuilders from enjoying the benefits of the dam. Builders assassinate cheaters, we might imagine. The game is now different because the players are different. In this case, the builders will be fitter than the nonbuilders, so Darwinian selection will maintain the building behavior.

In this example, the builders cooperate. Nonbuilders, we are imagining, do not. But the builders are not evolutionary altruists; and the nonbuilders are not evolutionarily selfish.

The vengeful builders are reciprocal altruists, in Trivers' sense. They do things that benefit others, but punish individuals who do not reciprocate. The point to focus on is that within the single beaver population, vengeful builders are fitter on average than the individuals who do not build. Vengeful building is just a variety of Darwinian selfishness. Given the choice between being a vengeful builder and an atomistic nonbuilder, an individual would quite selfishly prefer to be a builder. This is why reciprocal altruism is not altruism.

I want to emphasize that I have no interest in quibbling over words here. My reason for saying that reciprocal altruism is not altruism is motivated by a desire to clearly distinguish different kinds of causal processes. Individual selection can produce reciprocal altruists, but it cannot produce altruism in the sense defined in Figure 1c. We should recognize this fact about individual selection, not obscure it by lumping together two quite different kinds of characters. In

saying this, I think I am following Trivers' (1971) own observation that his model is intended 'to take the altruism out of altruism.'

Notice that applying the contrast between altruism and selfishness to a natural population can be quite difficult. When you go out in the woods and see all the beavers in a group cooperating to build a dam, you have no idea whether the trait in question should be called altruistic. You first have to ask yourself what the other traits were against which the one you observe was competing. This may take some imagination, because you have to envisage what variation was found in the ancestral population for natural selection to act upon. Unfortunately, selection frequently destroys the kind of evidence that is needed to reconstruct its history; selection requires variation to proceed, but typically it destroys the preconditions for its own existence.

III The tragedy of the commons

The Darwinian treatment of evolutionary altruism subverts the idea that natural selection must improve fitness. It is interesting to note that precisely the same phenomenon can arise in a very different domain. Rather than think of the natural selection of organisms, let us consider rational agents who deliberate about actions with a clear view of the consequences of what they do. When agents are fully informed and rationally deliberate, shouldn't they end up better off than they would be if they were irrational? The tragedy of the commons (also known as the prisoners' dilemma) in game theory provides a negative answer to this question, for reasons isomorphic with the Darwinian analysis of evolutionary altruism.

Let's imagine that you are deciding whether to put an emission control device on your car. We suppose that this is not a matter of law, but of individual choice. The cost to you is modest – $20. But what are the benefits? That depends on what other people do. If no one buys the device, it won't be worthwhile for you to buy one. Though the atmosphere would improve infinitesimally, the gain is so trivial that you'd rather save the $20. On the other hand, if everybody else buys the device, the atmosphere will be very good. But

here again, the improvement in the atmosphere that would be added if you also bought the device would be trivial. Again, you'd rather save the $20.

Your preferences, with 4 indicating best and 1 indicating worst, are shown in the following table:

		States of the World	
		Everybody Else Buys One	Nobody Else Buys One
Acts	You Buy	3	1
	You Don't Buy	4	2

The rational act in this game is to not buy the device. That action 'dominates' the alternative; whatever everybody else does, you're better off not buying (4>3 and 2>1).

But here is the rub: Everybody else has the same preferences, so each other agent rationally decides not to buy the device. What is the result? The group ends up with no one getting the device, which means that everybody receives two units of value. Notice that everybody is now worse off than they would have been if they had all decided to buy; in that case, the pay-off for each would have been three units.

This problem has the following paradoxical property. The rational action for each individual to choose is known in advance to make all the players worse off than they would have been if they had all chosen the irrational action.

In the above two-by-two table, I represented only two extreme states of the world – everybody else buys a device and nobody else buys a device. But there are intermediate frequency ranges – 90%

buys, 80% buys, and so on. The full game is not specified by a two-by-two table, but by a two-by-infinite table, so to speak. However, there is a simpler representation: merely use the fitness function for selfishness and altruism. Buying a device is altruistic; not buying is selfish. The payoffs from the table are inscribed as entries on the graph shown in Figure 1c. The problem of evolutionary altruism is an instance of the general game theoretic problem. Instead of rational deliberation, we have natural selection. And instead of preferences concerning dollar outlay and pollution, we have benefits computed in the currency of survival and reproductive success.

The fact that the problem posed by this decision problem has a rather depressing solution is not necessarily cause for despair. It is not carved in stone that human beings must play the game I have just described. For example, it is an assumption of this game that actions and states of the world are independent. Your buying an emission control device is independent of whether anybody else does. But suppose we pass a law that says that everybody has to do the same thing. Then we have a new game, with the only possible outcomes being the ones on the main diagonal of the previous table. The result is that we all choose to buy the device, which is a much cheerier prospect than the one obtained initially.[6]

There is an important truth behind the misleading idea that there are various ways of 'solving' the prisoners' dilemma problem. In the game as initially described, there is exactly one rational solution, which leads to a deleterious universal selfishness. The rational kernel, though, is that it is within the power of rational agents to restructure the games they play. The important thing to remember is that the solution to a game is contingent on the assumptions that went into defining the problem. If the assumptions can be changed, so too may the solution. What is inevitable within the framework of one game may not be within the framework of another.

Although human beings can consciously restructure the games they play, organisms in general do not have this ability when it

---

6 Another reformulation of the problem is provided by the iterated prisoners' dilemma, which is explored in Axelrod (1984).

comes to the problems posed by natural selection. Still, there is nothing absolute about the negative verdicts we have reached so far about evolutionary altruism. I have said that evolutionary altruism cannot evolve, if the game being played is Darwinian selection. But there has been a tradition of thinking in biology – one which has waxed and waned in the course of the development of evolutionary theory from Darwin to the present – that says that altruism is a reality, which means that Darwinian selection is not the game that organisms always play. We now need to examine this nonDarwinian idea. For our grasp of the concept of evolutionary altruism will be incomplete unless we see clearly how it is connected to the idea of group selection.

IV Simpson's paradox

It is a basic rule about natural selection, both in the simple format we have considered so far and in the context of the more complicated models we will consider now, that a trait must have a higher fitness if it is to increase in frequency. This is as true for altruism as it is for speed in the deer example. But we have already seen that within any group, altruism is less fit than selfishness. This is a matter of definition. How then can altruism evolve by natural selection?

To see that this is possible, one must grasp a paradox. Let us now consider not one group, but an ensemble of many groups. Within each group, altruists do worse on average than selfish individuals. But this fact does not guarantee that altruism is less fit when you average over the ensemble of groups. What is true within each group need not be true overall.

This is a concept that is very hard to grasp; we are so used to thinking that what happens in the part must translate directly into what happens in the whole. How can an organism get bigger if each of its parts gets smaller? That, I grant, does sound impossible. Suppose I told you that in every state of the USA, Democrats were declining in frequency and Republicans increasing. Would it follow that Democrats are becoming rarer in the US taken as a whole? The kneejerk reaction here is to say that what happens in the part must happen in the whole. We now must see that this need not be so.

Let's start with some very simple examples of how this decoupling of part and whole can occur. Imagine an audience in which men are on average taller than women. Is it possible to divide this audience into two groups, so that within each group, women are taller than men? Here's an example of how this can happen:

Group 1	Group 2	Global Average
10(F): 10	90(F): 5	100(F): 5.5
90(M): 9	10(M): 4	100(M): 8.5

There are a hundred females and a hundred males in total. The female average is 5.5 units of height; the male average is 8.5. We then split the total population of two hundred individuals into two groups. The first contains ten females and ninety males; the ten women are 10 units tall and the men are 9. The second group contains ninety females and ten males, with average heights of 5 and 4, respectively. The heights of the women and men within each group are given. Notice that males are taller on average, though women are taller within each group.

Another example of this phenomenon I owe to Nancy Cartwright (1979). She reports that the University of California at Berkeley was once investigated for discriminating against women in admission to graduate school. The reason for the suspicion was that women were turned down far more frequently than men. However, when departments were investigated one at a time, it emerged that the rejection rates of women and the rejection rates of men within each department were the same. Women were not turned down more often than men in Biology, in Philosophy, in Physics, or in any other department. But in the whole university of which these departments are parts, they were.

Let's construct a hypothetical example to see how this is possible:

	Department 1	Department 2	
applicants	10(M)	90(M)	= 100(M) total
	90(F)	10(F)	= 100(F) total
rejection rate	90%	10%	
number rejected	9(M)	9(M)	= 18(M) total
	81(F)	1(F)	= 82(F) total

We imagine that a hundred men and a hundred women apply to the two departments. Notice that in each department, a woman has the same chance of admission as a man. Yet women are turned down more often overall, because they disproportionately apply to a department with a very high rejection rate.

The phenomenon I have been discussing is sometimes called Simpson's paradox, in tribute to a statistician who wrote about it in the 1950s (Simpson 1951). However, the phenomenon has been noticed by statisticians for a long time.[7]

Let us review the two examples. In the first one concerning height, women were taller on average than men within each group, but men are taller than women overall. In the second, each academic department rejects women no more often than it rejects men, yet women are rejected more often overall. In both cases, we make two comparisons. First, we compare male and female averages within each group. Then, we compare the overall male average with the overall

7 Skyrms (1980, 107) cites Edgeworth, Pearson, Bravais, and Yule as having noted the phenomenon.

female average. The inequality within groups need not be maintained when we average across groups.

I hope these examples give you a feel for the pattern involved in Simpson's paradox. Now let's ask a separate question: what causes Simpson's paradox to arise? What allows inequalities within groups to reverse when we take the overall averages in these examples? The answer is *correlation*. In the first case, tall women tend to be found in the taller group. In the second, women tend to apply to departments with high rejection rates. If the male average and the female average were the same across groups, Simpson's paradox would disappear.

We now can show why Simpson's paradox is at the heart of the idea that group selection can allow altruism to evolve. Let's imagine that we have not one group, but an ensemble of them. In each there is some mixture or other of selfish individuals and altruists. We need to consider two questions. First, are selfish individuals fitter than altruists within each group? Second, are selfish individuals fitter than altruists, when we average over the ensemble of groups?

The answer to the first question is *yes*, given the fitness functions shown in Figure 1c. No matter what the frequency of altruism is in a group, altruists do less well than selfish individuals in the same group. But how are we to answer the second question? How are we to calculate and compare the overall fitnesses of altruists and selfish individuals?

Just to illustrate how Simpson's paradox applies here, let's imagine that our ensemble consists of two groups made of a hundred individuals each. The first is 1% selfish; the second is 99% selfish. Below, I've written the within group fitnesses and the overall fitnesses (rounded off, for simplicity) given by Figure 1c:

Group 1	Group 2	Global Average
1(S): 4	99(S): 2	100(S): 2
99(A): 3	1(A): 1	100(A): 3

Altruism is less fit within each group, but more fit when one averages over the ensemble of groups.

I mentioned before that models of natural selection of the sort we are considering imply that fitter traits increase in frequency. The present example is no exception. I stipulated that the two population ensemble begins with 50% altruists and 50% selfish individuals. What will happen to the frequencies of the traits in the next generation? Within each group, altruism will decline in frequency because it is less fit. But across the ensemble of groups, altruism will increase in frequency because it is on average fitter. So if we census the two population ensemble after one generation has passed, altruism will have increased in frequency.

What will happen if we follow the system over many generations? If the two groups remain intact – that is, if there is no extinction or splitting of groups to found colonies – then the two groups will grow larger and larger (assuming that the fitness values shown in Figure 1c represent reproduction above replacement level). Within each group, altruism will decline. So sooner or later, altruism must disappear from the two population ensemble. The increase in frequency in the first generation was momentary; starting with 50% altruists and 50% selfish individuals who are distributed into groups in the way described, altruism will initially increase. But sooner or later, the pattern that Dawkins (1976) once called 'subversion from within' must take its toll.

So we still have not seen how altruism can evolve and be maintained. But we are on the right track. One condition is before us: altruism must be fitter overall than selfishness, if it is to increase in frequency. How can this be achieved? As in the other examples of Simpson's paradox, the key idea is correlation. What is essential is that like live with like. Altruists must associate with each other more frequently than would be expected if association were at random. This could be achieved by having relatives live together; or it could happen if similar individuals preferred each other's company, regardless of whether they are relatives.

But like living with like is not enough. Even if the two groups just described are subgroups, subversion from within will drive altru-

ism to extinction as the kin reproduce. What is essential is that the groups fragment and found colonies.[8]

To see how this might happen, let's imagine that a group goes extinct if there are more than 50% selfish individuals in it. Imagine further that when a population reaches a certain census size, it fragments into many small subgroups, which then start growing. Notice that colonies are always founded by individuals from the same parent population. And founded colonies may not have exactly the same frequency of altrusm as their parents. Imagine that a parent population reaches the fission size of 1000 and then splits into 50 colonies of 20 individuals each. The parent population, we may suppose, is 75% altruistic. What will the 50 offspring colonies be like? Probably, some will be less than 75% altruistic, whereas others will contain more than 75% altruists.

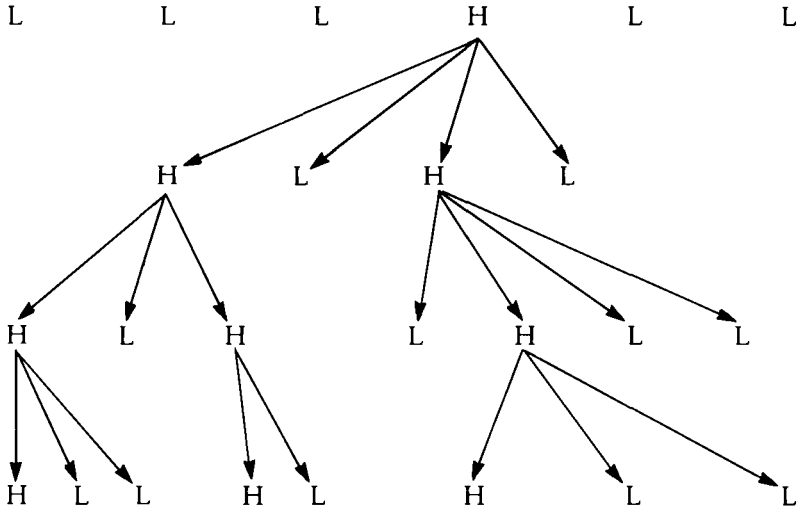
One more ingredient is needed, if altruism is to evolve and be maintained by group selection. It is the factor of timing. Suppose that selfishness is sufficiently fitter than altruism that if a group holds together for fifty generations, selfishness will go to 100% in it, no matter what the initial frequency was in the group. In this case, altruism will disappear if a parent population fragments and founds colonies less often than once every fifty generations. The fragmentation will come too late; by then, altruism will have disappeared. So groups must found colonies sufficiently often, how often being determined by how fast selfishness is displacing altruism within each group.

Figure 2 shows an example of group selection of the sort just described, whereby altruism can evolve and be maintained. Groups found colonies at a good clip and groups with low frequencies of altruists go extinct. If the numbers are right, one will find that in every generation of this process, altrusm is represented.

---

8 Although many biologists believe that Hamilton's (1964) concept of inclusive fitness allows self-sacrifice among relatives to be treated as a form of individual selection, I believe that this is a mistake. A single kin group that holds together for many generations will experience subversion from within just as much as a group of unrelated individuals. Hamilton (1975) recognizes this very point: an inclusive fitness treatment is *not* an argument against group selection. See Wilson and Sober (forthcoming) for further discussion.

Figure 2



V Evolution and the genealogy of morals

What connection is there between vernacular altruism and evolutionary altruism? I noted early on that it is possible for an individual to be an evolutionary altruist without being a vernacular one. Traits that are group beneficial but individually deleterious, like the example of the plant that leaches an insecticide, need not be psychological. I also pointed out that an individual can be a vernacular altruist without being an evolutionary one; this is what I am when I give you the piano sonatas out of the goodness of my heart, thereby distracting you from the business of reproduction.

Besides these simple distinctions, however, there is the question of what connection human morality has with natural selection. If systematically altruistic behavior (not just the occasional transfer of a volume of music) is a reality, what does this imply about our evolutionary past?

A strict Darwinian, in the current sense of that term, will deny the existence of evolutionary altruism. The reason is that the trait

implies the existence of a selection process that the Darwinian rejects. But even the strictest of Darwinians may sometimes lapse from the Darwinian straight and narrow. This is what Darwin himself did when he considered the evolutionary consequences of vernacular altruism. In *The Descent of Man*, Darwin formulated the issue in terms of his characteristic calculus of individual advantage:

> It is extremely doubtful whether the offspring of the more sympathetic and benevolent parents, or of those which were the most faithful to their comrades, would be reared in greater number than the children of selfish and treacherous parents of the same tribe. He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature. The bravest men, who were always willing to come to the front in war, and who freely risked their lives for others would on average perish in larger numbers than other men. (Darwin 1871, 163)

But rather than concluding that vernacular altruism does not exist, Darwin argued that what is bad for the individual may be good for the group:

> It must not be forgotten that although a high standard of morality gives but a slight or no advantage to each individual man and his children over the other men of the same tribe, yet that an advancement of well-endowed men will certainly give an immense advantage to one tribe over another. (Darwin 1871, 166)

Darwin's assumption here seems to be that vernacular altruism was under the direct control of natural selection. The trait is present now because, historically, there was selection for it. Darwin went the route of group selection because he did not doubt the trait's reality; some of his latter-day followers, on the other hand, have accepted the assumption, but have concluded that vernacular altruism cannot exist on the ground that individual selection is the name of the game.[9]

---

9 For example, Dawkins (1976, 3) asserts that human beings are 'born selfish' and Barash (1979, 135; 167) says that 'real, honest-to-God altruism simply doesn't occur in nature' and that 'evolutionary biology is quite clear that "What's in it for me?" is an ancient refrain for all life, and there is no reason to exclude *Homo sapiens*.'

However, there is another possibility that needs to be considered, which rejects the idea that a trait, if it exists now, must have been under direct selective control. It is the idea of evolutionary spin-off. Human beings now have the ability to do trigonometry; yet no one supposes that there must have been selection for that ability in ourancestral past. Rather, it is far more plausible to think that there was selection for some other suite of mental characteristics. Perhaps there was selection for increased intelligence and language use. Once these traits evolved and human beings subsequently found themselves in environments rather unlike the ancestral ones, various spin-off properties became visible.[10]

This is the scenario that Peter Singer (1981) explores in his book *The Expanding Circle*. Perhaps the ability to reason abstractly evolved because of its individual advantageousness. But once in place, this intelligence led human beings to see that rational considerations oblige them to take the interests of others as seriously as they take their own. If something like this is right, then vernacular altruism may find its pedigree not in evolutionary altruism, but in the sophisticated thoughts and feelings that a mind produced by individual selection was first able to formulate.

I will not evaluate the plausibility of this spin-off explanation of vernacular altruism. My point here is a conceptual one. Even if we suppose that group selection never happened – that selection is always selection for traits that are individually advantageous, it does not follow that vernacular altruism could not have evolved. It is one thing to hold that all selection is individual selection, quite another to maintain that all characters are under direct selective control.

10 The difference between direct selective control and spin-off is explained in Sober (1984) in terms of the distinction between 'selection of' and 'selection for.' Gould and Lewontin (1979) use the term 'spandrel' to mark the concept of evolutionary spin-off.

VI Concluding remarks

Evolutionary altruism is a kind of trait. In our plant example, it involves leaching an insecticide into the soil; in a species of crow, it might involve issuing warning cries. Traits are altruistic, so altruism is a trait of a trait. The evolutionary problem is to see whether the physiological, behavioral, and morphological traits found in nature are examples of evolutionary altruism.

The trait in question must not be confused with the trait of vernacular altruism. Although it is possible to propose a causal connection between evolutionary altruism and vernacular altruism, as Darwin did in one direction and some contemporary sociobiologists have done in the other, this is not inevitable. Evolutionary altruism does not imply vernacular altruism, nor does vernacular altruism imply evolutionary altruism.

Group selection can lead to the evolution and maintenance of evolutionary altruism. Darwinian selection cannot. Although altruists are by definition less fit than selfish individuals within the same group, this does not settle the question of their comparative fitnesses when we average over the ensemble of groups. To see why this is so, one must grasp the meaning of Simpson's paradox. Once this is achieved, one can understand how altruism can evolve, given the right assumptions about like living with like and appropriate rates of extinction and colonization.

All this is not to say one word about whether evolutionary altruism is found in nature. My concern here has been to say what altruism is, not whether it exists. I have mentioned that evolutionary opinion has swung back and forth on this question. At the moment, Darwinism is the dominant mode of thought; although group selection and altruism are not treated with total scorn by all biologists, it is a small minority of biologists that takes the idea seriously.

The reasons for this opinion bear examining. Sometimes opposition to group selection is based on spurious arguments. It is sometimes suggested that altruism cannot evolve simply because, by definition, altruists are less fit than selfish individuals within each group. An understanding of Simpson's paradox should make us immune to the attractions of this *non sequitur*. However, even if Darwinism is right in rejecting group selection, it is important that it

do so for the right reasons. There are substantive questions here about natural selection that need to be resolved; removing confused arguments may help biologists see these questions for what they are.

References

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

Barash, D. (1979) *The Whisperings Within* (Harmondsworth: Penguin).

Cartwright, N. (1979) 'Causal Laws and Effective Strategies,' *Nous* **13**, 419-37.

Darwin, C. (1872) *The Descent of Man, and Selection in Relation to Sex* (London: J. Murray [First edition]; Princeton: Princeton University Press 1981).

Dawkins, R. (1976) *The Selfish Gene* (Oxford: Oxford University Press).

Falconer, D. (1981) *Introduction to Quantitative Genetics* (London: Longman).

Gould, S. and R. Lewontin. (1978) 'The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme,' *Proceedings of the Royal Society of London* **205**, 581-98. Reprinted in E. Sober, ed., *Conceptual Issues in Evolutionary Biology* (Cambridge: MIT Press 1984), 252-70.

Hamilton, W. (1964) 'The Genetic Evolution of Social Behavior,' *Journal of Theoretical Biology* **7**, 1-52.

Hamilton, W. (1975) 'Innate Social Aptitudes of Man: An Approach from Evolutionary Genetics.' In R. Fox, ed., *Biosocial Anthropology* (London: Malaby Press), 37-67.

Kitcher, P. (1985) *Vaulting Ambition: Sociobiology and the Quest for Human Nature* (Cambridge: MIT Press).

Simpson, E.H. (1951): 'The Interpretation of Interaction in Contingency Tables,' *Journal of the Royal Statistical Society B* **13**, 238-41.

Singer, P. (1981) *The Expanding Circle* (New York: Farrar, Straus, and Giroux).

Skyrms, B. (1980) *Causal Necessity* (New Haven: Yale University Press).

Sober, E. (1984) *The Nature of Selection* (Cambridge: Bradford/MIT Press).

Sober, E. (1985) 'Methodological Behaviorism, Evolution, and Game Theory.' In J. Fetzer, ed., *Sociobiology and Epistemology* (Dordrecht: D. Reidel).

Sober, E. (forthcoming) 'The Anatomy of Egoism.'

Trivers, R. (1971) 'The Evolution of Reciprocal Altruism,' *Quarterly Review of Biology* **46**, 35-57.

Wilson, D. and Sober, E. (forthcoming) 'Reviving the Superorganism.'