# On the Relationship Between Evolutionary and Psychological Definitions of Altruism and Selfishness

DAVID SLOAN WILSON

*Department of Biological Sciences*
*State University of New York*
*Binghamton, New York 13902–6000, U.S.A.*

ABSTRACT: I examine the relationship between evolutionary definitions of altruism that are based on fitness effects and psychological definitions that are based on the motives of the actor. I show that evolutionary altruism can be motivated by proximate mechanisms that are psychologically either altruistic or selfish. I also show that evolutionary definitions *do* rely upon motives as a metaphor in which the outcome of natural selection is compared to the decisions of a psychologically selfish (or altruistic) individual. Ignoring the precise nature of both psychological and evolutionary definitions has obscured many important issues, including the biological roots of psychological altruism.

KEY WORDS: Altruism, evolution, group selection, selfishness.

The word "altruism" and associated words such as "selfishness", "spite" and "cooperation" are familiar to everyone as descriptions of human conduct. The same words are used routinely by evolutionists to describe the behavior of non-human species. At first glance this practice seems unobjectionable. Charges of anthropomorphism aside, if we see a baboon endangering its life to protect another baboon from a leopard, or if we see a baboon doing everything it can to put another baboon between itself and a leopard, it seems reasonable to use the same language that we would use if the baboons had been humans.

At second glance, however, the relationship between the evolutionary and everyday meanings becomes more complex. In particular, evolutionary definitions are supposed to be based solely on fitness effects. If a behavior increases the fitness of a recipient at the expense of the actor's fitness it is termed altruistic regardless of what the actor felt or thought as it performed the behavior. By contrast, everyday meanings depend largely on the motives of the actor. If we see someone benefit others at the economic, social or material expense of himself we may still regard him as selfish if we think that he derives pleasure from his action or if he regards his action as part of a broader scheme in which he, the actor, stands to gain. Definitions that are based on the actor's motives will hereafter be called psychological definitions.[1]

Since motives figure so largely in everyday meanings of altruism and associated words, it is remarkable that the words can remain intuitive when stripped of their motivational content. It is equally remarkable that, in borrowing the words from common language, evolutionists have changed the defining criterion from motives to effects largely without comment. In this note I make

*Biology and Philosophy* **7**: 61–68, 1992.

two points, one elementary and the other more subtle. First, behaviors that are altruistic in the evolutionary sense can be psychologically either selfish or altruistic. No simple relationship exists between the forces of natural selection that cause behaviors to evolve and the proximate mechanisms that elicit the behaviors. Second, evolutionary definitions remain intuitive by invoking a cryptic form of motivation, based not on what the actor thinks or feels but on what the evolutionist must account for while calculating gene frequency change. Reliance on motives as a metaphor may explain why several conflicting definitions of altruism exist among evolutionists, each with their own intuitive appeal.

## 1. FROM FITNESS EFFECT TO MOTIVES OF THE ACTOR

As mentioned above and described in detail below, evolutionists frequently disagree on whether a particular behavior is altruistic or on whether altruism in general is rare or common in nature. Nevertheless, all evolutionists agree that behaviors defined as altruistic based on fitness effects can, in principle, evolve. Thus, the fact that a behavior is *evolutionarily successful* does not imply that it is necessarily *selfish*. Let us therefore consider a behavior, defined as altruistic based on fitness effects, that is also evolutionarily successful. Any behavior that evolves must include a proximate mechanism that actually causes the organism to display the behavior. Psychological definitions of altruism and selfishness are based on these proximate mechanisms. Our question is: Are behaviors that are altruistic in the evolutionary sense necessarily caused by proximate mechanisms that are altruistic in the psychological sense?

The answer is "no". The simplest case to consider is pleasure, which can be broadly regarded as a physiological/psychological mechanism, designed in part by natural selection, that causes organisms to behave in ways that are evolutionarily successful. But "evolutionarily successful" is not the same as "selfish"! If increasing the fitness of others at the expense of one's own fitness is evolutionarily successful in the long run, it can be proximately motivated by the same feelings of pleasure that accompany successful selfishness. From the evolutionary standpoint, defining selfishness as "the pursuit of personal pleasure" would be as meaningless as defining selfishness as "everything that evolves" (see also Sober, ms).

A less obvious and therefore more interesting case involves cognition, which can be broadly regarded as a system of mental operations, designed in part by natural selection, that causes organisms to behave in ways that are evolutionarily successful. Consider an imaginary experiment in which a human subject is asked to choose between two behaviors, A and B. If he elects A he must donate 1 dollar and 5 dollars will be given to a member chosen at random from his group. If he elects B no money will be taken from him. In both cases he stands to receive 5 dollars from other A-types in his group. We assume that the subject is psychologically selfish and quickly decides to be a B-type. Now we inform the subject that 2 groups exist, one with 20% A-types and one with 80% A-types. If he elects to be an A-type he will be placed in groups 1 and 2 with probability 0.2 and 0.8 respectively, and visa versa if he elects to be a B-type. Our subject now embarks upon the following mental calculation: "In group 1, I can expect to receive 0.8(5) = 4 dollars from other A-types while in group 2 I can expect to receive only 0.2(5) = 1 dollar. It is true that as an A-type I will lose a dollar, but I also have an 80% chance of being in group 1, yielding an expected gain of 0.8(4) + 0.2(1) - 1 = 2.4 dollars. As a B-type I will keep my dollar but I also will probably end up in group 2, yielding an expected gain of 0.2(4) + 0.8(1) = 1.6 dollars". Our subject now elects to be an A-type despite the fact that he is psychologically selfish and did not include the welfare of others in his mental calculation.

Readers familiar with the evolutionary literature will recognize that, if we substitute the word "offspring" for "dollars", the two-group version of the experiment is equivalent to a structured population model for the evolution of altruism in which A-types increase the fitness of others at the expense of themselves but nevertheless evolve because of a clustering process that causes altruists to interact primarily with other altruists (Wilson 1983, 1989).[2] Nevertheless, it is possible for a utility-maximizing organism to elect to be an A-type without directly considering the welfare of others by combining both the effect of its own behavior and the population structure in a single measure of fitness averaged across all contexts. In effect, the self becomes a *representative* A-type or a *representative* B-type, existing in the capacity of both actor and other, which makes the inclusion of others as a separate category redundant. In this fashion, a thinking organism whose mental operations are psychologically selfish can, in principle, adopt any evolutionarily successful behavior.[3]

The fact that evolutionarily successful behaviors are not necessarily selfish, and that proximate mechanisms are designed to elicit evolutionarily successful behaviors regardless of whether they are selfish or altruistic, destroys any hope for a simple relationship between definitions based on fitness effects and definitions based on motives. Not only can altruistic behaviors (in the evolutionary sense) be selfishly motivated (in the psychological sense), but the reverse is also true; individuals that care truly for others can be selfish in the evolutionary sense (Sober, ms). These observations are elementary but they are not sufficiently appreciated by evolutionists or philosophers interested in the concept of altruism.

## 2. MOTIVES AS A METAPHOR IN EVOLUTIONARY DEFINITIONS

Given the problems outlined above, it is a wonder that evolutionary discussions of altruism and associated words appear as natural as they do. Consider the following passage from Nunney (1985, p. 226).

> Suppose that you are offered two financial options. Under the selfish option you receive 10 dollars and keep it all. Under the benevolent option you receive 10 million dollars but 6 million must be given to a neighbor. Given that neighborhoods are random samples of a large population, the choice is clearly the benevolent option, a choice based purely on individual greed and not on the general benefit to the neighborhood. Replacing money by fitness, it can be seen that benevolence spreads by individual selection because a net gain of 4 million units of fitness is superior to a net gain of 10 units of fitness. The 6-million gain of the neighbor is irrelevant.

In the first half of this passage, a person chooses to receive 4 million dollars rather than ten dollars despite the fact that a neighbor, randomly chosen from a large population, will receive even more than himself. This is because the person is motivated entirely by self-interest and ignores his effects on others (positive or negative) in his decision. The second half of the passage concerns an evolutionary model in which A-types have an effect d on themselves and an effect r on members of their group. If groups are formed at random from a large population then the probability of the recipients being A will equal the frequency of A in the large population. Any effect on recipients, positive or negative, on average will not alter the global frequency of types. Thus, only effects on self (represented by $d$) produce evolutionary change and effects on others (represented by $r$) appear irrelevant. If groups are not formed at random then the probability of the recipients being A can exceed the frequency of A in the large population and positive effects on others ($r > 0$) can be selected despite negative effects on self ($d < 0$). Such behaviors are altruistic according to Nunney.

Notice that the proximate mechanisms that motivate the A-types are never an issue. The similarity between the first and second half of the passage is between a *human actor* on the one hand, who cares only about himself, and an *evolutionary process* on the other, in which only absolute effects on self are relevant to gene frequency change. Thus, despite the fact that Nunney's definitions of altruism and selfishness are based on fitness effects, they owe their intuitive appeal to a metaphor based on motives: When groups are formed at random, the products of natural selection are *like* the decisions of a psychologically selfish individual that cares only about its own absolute fitness. When groups are formed nonrandomly, the products of natural selection are *like* the decisions of a psychologically altruistic individual that values the welfare of others.

As with all metaphors, this one has the advantage of endowing the unfamiliar (a model) with properties of the familiar (human decision making). Unfortunately, at least two other methods exist to calculate gene frequency change that are equally amenable to the same metaphor.

a) Hamilton (1964) invented a new measure called inclusive fitness, that includes the effects of a gene not only on itself but on all copies that are identical by descent, which by definition are present only in genetic relatives. Thus, an individual can increase "its" inclusive fitness by aiding relatives, even at the expense of its personal fitness. Hamilton's measure is widely regarded as the utility that well adapted organisms strive to maximize. But by replacing the utility of personal fitness with the utility of inclusive fitness, the boundary of selfishness can be pushed outward to include aid-giving to relatives. The products of natural selection are *like* the decisions of a psychologically selfish individual who cares only about maximizing copies of its genes. "True" altruism now appears to require aid-giving to nonrelatives at the expense of the actor.[4]

Many authors have shown that inclusive fitness theory can be reformulated as a structured population model in which kin groups are random samples of the parents gametes and therefore nonrandom samples of the population at large (e.g., Michod 1982). Thus, what is "selfish" according to inclusive fitness theory is "altruistic" according to Nunney.

b) Group selection models examine gene frequency change within groups and then consider the differential productivity of groups to calculate global gene frequency change (Wilson 1983, 1989, 1990; Wilson and Sober 1989). In single groups, evolutionary success depends critically on fitness relative to one's neighbors. To paraphrase Nunney, it is like choosing between 10 dollars for yourself or 10 million dollars, 6 million of which must be given to your archenemy. Despite the fact that 4 million is greater than 10, you might well prefer to be 10 dollars richer than 2 million dollars poorer. Stated in terms of effects on self and others, A-types spread within groups when $d > r$ and all traits for which $r > d$ require between-group selection, even when absolute effects on self are positive (Wilson 1980, 1990). The products of natural selection within single groups are *like* the decisions of a psychologically selfish individual who cares only about surpassing his rivals. What is "motivated by individual greed" according to Nunney (traits for which $d > 0$) can appear altruistic in a group selection model.[5]

To summarize, evolution in structured populations is a complicated process that can be modelled in several ways. Each method can be made intuitive by employing the metaphor of human decision making, but the distinction between selfish vs. altruistic that the metaphor makes "natural" is different for each method. The price of employing the metaphor as a heuristic tool is an insidious form of pluralism in which different frameworks use the same words to refer to different things (see also Wilson and Sober 1989 and Wilson and Dugatkin 1991).

## 3. DISCUSSION

Definitions of altruism and associated words are so entrenched, and so useful in their own contexts, that it is impossible to resolve the semantic issue by abolishing all but one set of definitions, even within the evolutionary literature. Neither is it practical to reserve the common words for one set and invent new words for the others. That leaves the awkward necessity of realizing that single words have many meanings, of being explicit about our own usage, and making the dynamics of multiple meanings a subject of inquiry in its own right. In this spirit, I conclude with a number of questions that deserve further study.

### *a. Multiple Meanings in Common Language*

The pluralism referred to above is not confined to the evolutionary literature. In common language, some people use the word "selfish" as a pejorative and others use the same word as a principle to explain all rational behavior. Some define "altruism" to include all prosocial behaviors while others make it a synonym for "stupid". These observations are so mundane that they evidently have not been deemed worthy of serious study. One philosopher with whom I raised the subject merely shrugged his shoulders and replied that "people have agreed to disagree". But such radical variation in the meanings of words so intimately connected to human conduct requires an explanation.

### *b. Why Are Motives so Important in Common Language Definitions?*

Since people are only affected by the external actions of others, why do they care about the inner machinery that motivates behaviors? I suggest that the importance of understanding motives is to predict how a person is likely to behave in the future. Putting it another way, altruists (defined psychologically) are more altruistic (defined in terms of effects) than selfish types (defined psychologically) *on average* even though both types might behave identically at a given moment. If this interpretation is correct, then definitions based on motives owe their relevance to definitions based on effects.

### *c. The Evolution of Psychological Altruism*

When are organisms with feelings of pleasure based on empathy and with cognitive processes in which others are perceived as valuable in their own right more fit than organisms for whom others are psychologically merely tools for manipulation? As we have seen, the answer to this question is not as simple as determining when benefitting others at the expense of self is evolutionarily successful. The evolutionary literature is therefore curiously silent when it comes to explaining the biological roots of psychological altruism. Several possibilities suggest themselves. Perhaps psychological altruists are so popular as associates that they succeed despite frequent exploitation. Perhaps clustering mechanisms that segregate altruists from nonaltruists are so consistent that the ability to calculate when to be altruistic (in terms of effects) is unnecessary. Or perhaps the "bounded rationality" of the human brain (Simon 1983) makes simple empathy an efficient rule of thumb, compared to Machiavellian thought (Ruse 1986).

These and other questions become interesting when we realize that the evolution of altruism, defined in terms of fitness effects, does not by itself determine the proximate mechanisms that elicit the altruistic behaviors.

## ACKNOWLEDGEMENTS

I thank Lee Dugatkin, Elliott Sober and George Williams for helpful discussions. Supported by grant no.: DE-FG02-89ER60884 awarded by the Ecological research division, Office of Health and Environmental research, U.S. Department of Energy.

## NOTES

[1] Bertram (1982), Kitcher (1985) and Sober (1988) make a similar distinction between evolutionary and psychological definitions but their development of the theme is somewhat different from mine, as discussed below.

[2] Another major difference between evolutionary and psychological definitions is that the currency of altruism in evolutionary models is always fitness whereas in common language it can be any desirable commodity (Sober 1988). As Sober (1988) points out, however, evolutionary and economic models based on prisoner's dilemma and tragedy of the commons scenarios are otherwise parallel.

[3] Kitcher (1985) and Sober (1988) also discuss the tenuous connection between evolutionary and psychological altruism. Both assert that psychological altruism can exist even if evolutionary altruism never evolves, either because the proximate mechanisms that motivate evolutionarily selfish behavior are psychologically altruistic or because there is more to human behavior than fitness maximization. My main point, that evolutionarily altruistic behavior can be psychologically selfish, is complementary to theirs.

[4] Even within inclusive fitness theory, evolutionists can't agree on definitions of altruism. Some regard helping relatives at the expense of self as a form of altruism that is explained by inclusive fitness theory, while others regard the same behaviors as selfish because they maximize the individual's inclusive fitness.

[5] I disagree with Sober's (1988) point that evolutionary definitions of altruism are comparative. Despite their differences, all definitions of altruism reviewed above are based on effects on self and others (weighted by coefficients of relatedness in the case of inclusive fitness theory) without reference to other behaviors that exist in the population.

## LITERATURE CITED

Bertram, B.C.R.: 1982, 'Problems with Altruism', in P. Bateson (ed.), *Current Problems in Sociobiology*. Cambridge University Press, Cambridge, pp. 251–268.

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

Kitcher, P.: 1985, *Vaulting Ambition*, MIT Press, Cambridge, Mass.

Nunney, L.: 1985, 'Group Selection, Altruism and Structured-Deme Models,' *American Naturalist* **126**, 212–230.

Ruse, M.: 1986, *Taking Darwin Seriously*, Basil Blackwell, Oxford.

Simon, H.: 1983, *Reason and Human Affairs*, Stanford University Press, Palo Alto.

Sober, E.: 1988, 'What Is Evolutionary Altruism?', in B. Linsky and M. Matthen (eds.), *New Essays on Philosophy and Biology*, Canadian Journal of Philosophy supplementary volume 14, 75–99.

Sober, E.: 1989, 'What Is Psychological Egoism?', *Behaviorism* **17**, 89–102.

Sober, E.: (in press), 'Evolutionary Altruism, Psychological Egoism and Morality:

Disentangling the Phenotypes', in M. Nitecki (ed.), *Evolutionary Ethics*.

Wilson, D.S.: 1980, *The Natural Selection of Populations and Communities*, Benjamin Cummins, Menlo Park CA.

Wilson, D.S.: 1983, 'The Group Selection Controversy: History and Current Status', *Annual Review of Ecology and Systematics* **14,** 159–189.

Wilson, D.S.: 1989, 'Levels of Selection: An Alternative to Individualism in Biology and the Human Sciences', *Social Networks* **11,** 257–272.

Wilson, D.S.: 1990 'Weak Altruism, Strong Group Selection', *Oikos* **59**, 135–140.

Wilson, D.S. and L.A. Dugatkin: 1991, 'Altruism', in E.F. Keller and L. Lloyd (eds.), *Keywords in Evolutionary Biology*, Harvard University Press, Cambridge, Mass.

Wilson, D.S. and E. Sober: 1989, 'Reviving the Superorganism', *Journal of Theoretical Biology* **136,** 337–356.