Recent Work on Human Altruism and Evolution

Neven Sesardic

Ethics, Volume 106, Issue 1 (Oct., 1995), 128-157.

SURVEY ARTICLE

Recent Work on Human Altruism and Evolution*

Neven Sesardic

Altruism and evolution do not mix well. Paraphrasing Quine (1969, p. 126) one might say that "inveterately altruistic creatures have a pathetic tendency to die before reproducing their kind." Such a view that rather simple Darwinian forces work strongly against the preservation of altruistic traits is actually the background against which various explanations of the genesis of human altruism are being defended and discussed today. Indeed, if we call "paradoxical" any situation where we have seemingly convincing evidence in favor of each of two (or more) propositions that are seemingly mutually irreconcilable, then participants in the current debate about the emergence of human altruism are also haunted by a paradox. Moreover, since the debate is so conspicuously and so persistently revolving around this basic difficulty it seems methodologically appropriate to make the "paradox of altruism" the cornerstone of my presentation of the recent work on the topic.

INTRODUCTION: THE PARADOX OF ALTRUISM

To begin with, here is a crude version of the paradox of altruism: on one hand it seems that the existence of human altruism is an undeniable psychological fact, but on the other hand it seems, on evolutionary grounds, that altruism cannot exist, because species with this trait are expected to have gone extinct through the process of natural selection. (Selfishness increases biological fitness, and only the fittest survive.)

The alleged incompatibility between these two propositions is easily resolved. A standard strategy is to remove the sting of the para-

* I would like to thank Elliott Sober and an anonymous referee for *Ethics* for their very helpful critical comments on the first draft of this article. I have also benefited from discussions after presenting the paper to audiences in Zagreb, Dubrovnik, at Notre Dame, Purdue, Rutgers, and at the University of Minnesota.

Ethics 106 (October 1995): 128–157
 0014-1704/96/0601-1001$01.00

dox by distinguishing two meanings of 'altruism', psychological and evolutionary.[1] For our purposes, psychological altruism ($altruism_p$) and evolutionary altruism ($altruism_e$) will be defined as follows:

A is behaving $altruistically_p$ $=_{df}$ A is acting with an intention to advance the interest of others at the expense of his own interests.

A is behaving $altruistically_e$ $=_{df}$ The effect of A's behavior is an increase of fitness of some other organisms at the expense of its own fitness.

Invoking this conceptual distinction we can perfectly consistently state both that $altruism_p$ is a (psychological) fact and that $altruism_e$ is an (evolutionary) impossibility. However, a deeper and subtler difficulty remains. Namely, even after avoiding a direct contradiction by separating the two senses of 'altruism' one can argue that these two senses are not completely unconnected, and that for this reason we may still be left with an epistemic tension between believing in the reality of psychological altruism and at the same time doubting the existence of evolutionary altruism.

Indeed, I shall try to show that we cannot get rid of the paradox of altruism by simply making the clashing statements speak about different things (i.e., different altruisms). With this purpose in mind I shall replace the initial crude version of the paradox of altruism with a more sophisticated form. The new version will consist of four propositions (instead of two), each of them again being seemingly very plausible despite all of them appearing to be mutually incoherent. This proposed reconstruction could do more than just help to exhibit the logical skeleton of our puzzle. Adding more structure and precision to the formulation of the paradox leads almost by itself to a novel, neat classification of alternative approaches, and (I hope) to a more fruitful comparison of these rival views. By making transparent the basic points of disagreement it could perhaps also lead to a deeper understanding of the genesis of human altruism.

Here is the "incongruous tetrad," the four conflicting propositions that I offer as a reconstruction of the paradox of altruism:

(1) $Altruism_e$ is a selectively disadvantageous trait.
(2) $Altruism_p$ tends to lead to $altruism_e$.
(3) $Altruism_p$ exists.
(4) $Altruism_p$ is a product of natural selection.

1. For good discussions of this important distinction see Kavka 1986; Sober 1988, 1993a; and Wilson 1992.

Our predicament is in short this: how is it possible (3) that altruism$_p$ exists and (4) that it is a product of natural selection if (2) it tends to lead to altruism$_e$ which itself (1) is a selectively disadvantageous trait? Proposition (1) is a statement about evolutionary altruism, (3) and (4) are statements about psychological altruism, and (2) supplies a link between them that creates a logical strain in the tetrad.

I have to explain why I have introduced (3) as a separate claim about psychological altruism, although its truth is obviously presupposed by (4). (Altruism$_p$ cannot be a product of natural selection unless it exists.) The reason is that (4) has two components: it presupposes that altruism$_p$ exists, and it states that altruism$_p$ is a product of natural selection. For the sake of clarity these two components ought to be considered and evaluated separately. The supposition—expressed by (3)—can be attacked by insisting that the behavior satisfying the definition of psychological altruism simply does not exist; or, alternatively, conceding the existence of psychological altruism, one can deny (4) by asserting that this kind of behavior is actually not a product of natural selection. Both of these arguments have been defended in the literature, and it seemed to me that it would only invite unnecessary confusion to fuse both controversial points into one sentence.

Note also that on both definitions of altruism the so-called reciprocal altruism is a misnomer: it is not altruism at all. Evolutionarily altruistic acts imply the net loss of fitness of the actor, and psychologically altruistic acts imply the intention of the actor to genuinely sacrifice his own interests. This terminological decision to exclude reciprocal altruism from the scope of altruism proper accords well with a widespread biological and philosophical usage. For instance, Peter Singer says that "reciprocal altruism is not really altruism at all; it could more accurately be described as enlightened self-interest" (Singer 1981, p. 42). Rawls prefers not to talk about reciprocal altruism but to call it simply reciprocity (Rawls 1971, p. 503), and Robert Trivers in his classical paper "The Evolution of Reciprocal Altruism" claims that "models that attempt to explain altruistic behavior in terms of natural selection are models designed to take the altruism out of altruism" (Trivers 1978, p. 213). The feeling that reciprocity is not altruism *sensu stricto* is best expressed in one of La Rochefoucauld's maxims: "When we help others in order to commit them to help us under similar circumstances, [the] services we render them are, properly speaking, services we render to ourselves in advance."[2]

There is, however, an additional, substantive reason for shutting the door on debating reciprocal altruism in the present context. Speak-

2. Sober is also arguing against classifying reciprocal altruism as altruism (1988, p. 84).

ing in evolutionary terms, reciprocal altruism has no puzzling features. Its being easily explainable as a selectively advantageous trait (to individual organisms) robs it of any biological "queerness." Hence the crucial difficulty reflected in (1) does not apply to it. On the psychological side, similarly, reciprocal altruism is unproblematic: it can arise through natural selection simply by riding on its biological counterpart (evolutionary reciprocal altruism), whose evolutionary credentials are impeccable, as we have just seen. Therefore, one who wants to focus on the paradox of altruism as here formulated is well advised to get reciprocal altruism out of the way as a red herring.

Our paradox can be resolved by choosing between the following two strategies. Either one can reject (at least) one of the propositions of the incongruous tetrad or else one can attempt to prove that contrary to the appearances the whole set is in reality perfectly coherent. The first (eliminativist) strategy comes in four possible variants (i.e., each one of the four assumptions can be dropped); interestingly enough, all these variants had their advocates in the continuing debate about altruism. On the other hand, the second (reconciliationist) strategy splits upon analysis into three possible versions that, again happily, represent the currently most important theoretical standpoints. Moreover all these enumerated options exhaust the logical space of possible solutions. (Of course, I do not want to say that there will be no novel approaches to the problem, only that any of them will have to fall into some place in my scheme.) Let us therefore follow this emerging order and discuss in turn the two strategies in all their subvariants.

FIRST STRATEGY: ELIMINATION

None of the four propositions creating the paradox carries its truth on its sleeve. Each one of them has been occasionally regarded by some as dubious and by others as outright false. Our task in this section is to see whether there are good reasons for rejecting any of these claims in particular. To anticipate a little, my conclusion will be that the examination largely bears out the truth of all four propositions, and that, consequently, the road to solution is better sought in the reconciliationist approach.

Claim (1)

Starting with claim (1)—"Altruism$_e$ is a selectively disadvantageous trait"—there are two phenomena that prima facie speak against it: (*a*) kin selection and (*b*) group selection.

a) The originator of the idea of kin selection was J. B. S. Haldane, although he did not use the term. As early as 1932 he wrote, "Insofar as it makes for the survival of one's descendants and near relations, altruistic behavior is a kind of Darwinian fitness, and may be expected to spread as a result of natural selection" (Haldane 1932, p. 131). So,

if we take account of the effects of behavior on the agent's relatives (or, more precisely, on the carriers of the same genes) it may happen, contrary to (1), that altruistic behavior becomes selectively advantageous, despite being harmful or even lethal for the individual in question.

Here all turns again on the definition of evolutionary altruism. Recall that it was defined as the behavior of an organism which decreases its own fitness while increasing the fitness of some other organisms. In evolutionary theory, at least since William D. Hamilton's (1964) important theoretical contribution, 'fitness' is often taken to mean inclusive fitness, and for the purposes of our discussion we shall stick to this usage. But adopting the inclusive fitness approach entails that our measurement of the fitness of a behavioral disposition is not limited solely to the consequences of that behavior on its emitter; "inclusive fitness" incorporates, *ex vi termini,* the effects of this behavior on close relatives. To take a concrete example, someone who sacrifices his own life and thereby, say, saves the lives of more than two of his full siblings is thereby actually increasing his (inclusive) fitness, and by so acting he is not behaving altruistically (according to this definition of altruism$_e$).[3] Therefore, when (1) is properly interpreted, it is in no way threatened by the existence of kin selection. (To evade the objection from kin selection the following concise formulation of [1] suggests itself: "The behavior that systematically decreases the inclusive fitness of its emitter is selectively disadvantageous.")

b) A more serious challenge to (1) is group selection. The groups consisting of altruists can fare better than the groups of selfish organisms, and consequently it seems that these groups can even be favored by selection despite the fact that such altruistic behavior, at the individual level, continues to decrease the inclusive fitness of any particular altruist organism. The range of group selection is a matter of great controversy in biology. In the earliest stage, group selection was being routinely invoked without much awareness of formidable problems concerning its way of operation. The central difficulty is that, although it is undoubtedly in the evolutionary interest of any individual to be a member of a group of altruists, he always gains a selective advantage by being an egoist himself. This fact that group selection is open to subversion from within (i.e., from the level of individual selection) is transparent in figure 1.[4]

3. "From a genetic perspective, you are helping part of your *self* (i.e., replicas of your genes) when you help your brothers and sisters. Faced with a decision between saving yourself or three full siblings, you save more of your (genetic) self by saving your siblings" (Krebs 1982, p. 453).

4. This kind of diagram is a standard way of presenting the basic structure of the group selection problem. See Sober 1984, p. 186; 1988, p. 80; 1993a, p. 206; 1993b, p. 98; Elster 1989, p. 127; Peressini 1993, p. 572).

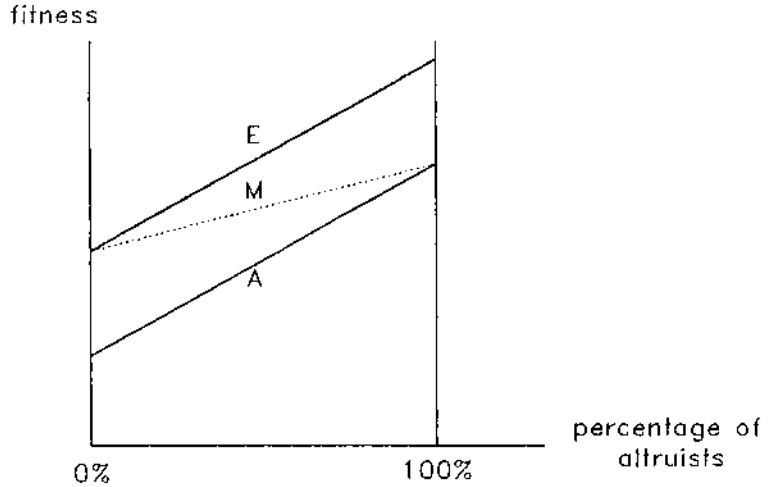

FIG. 1

Three lines, *E*, *M*, and *A*, represent, respectively, the fitness of an egoist, the mean fitness of the group, and the fitness of an altruist. Each of them is plotted as a function of the percentage of altruists in the group. Obviously, from the point of view of the group it is better to have as many altruists as possible (because line *M* rises if we move from 0 to 100 percent of altruists), but from the individual perspective it always pays to be selfish (because line *E* is always higher than *A*).

In order to explain altruism by group selection it is emphatically not enough to show that universal altruism is a collectively preferred state. There is an additional obligation (and not a trivial one at that) to specify a causal mechanism by which this state can be reached. (The property of being a collectively preferred state doesn't by itself cause anything.) The argument for group selection was elaborately worked out by Wynne-Edwards (1962), and this whole approach was then forcefully attacked by George C. Williams in his classic *Adaptation and Natural Selection* (Williams 1966). In overreaction to the once too facile use of group selection explanations there was a tendency later to dismiss them all as being pseudoexplanations. Today, however, it has become a matter of consensus that group selection is a possible evolutionary process, but still many biologists believe that it can work only under very special and relatively rarely satisfied conditions (e.g., if groups are small, if their extinction rate is high, if there is little intergroup migration, etc.). Assuming the fragility of group selection (i.e., its operating only under rather exceptional circumstances), our belief in (1) would be partly vindicated. That is, barring the unlikely concurrence of all the prerequisites for group selection it would remain true that altruism$_e$ is a selectively disadvantageous trait.

It should be noted, however, that there is a growing opinion in evolutionary circles that the skepticism toward group selection was

mainly motivated by bad arguments and elementary conceptual confusions (involving the distinction between "replicators" and "vehicles"), and that group selection should be recognized as an extremely important and strong causal force both in biology and in human behavioral sciences (see particularly Wilson and Sober 1994). Whatever the outcome of this discussion about the general status of group selection, it is by itself not likely to impinge so much on the issue of altruism for the simple reason that, as Wilson and Sober themselves say, even extreme individualists (the die-hard opponents of group selection) "acknowledge group-level adaptations *when they are easily exploited within groups*" (p. 599; emphasis added). In other words, the main thrust of Wilson and Sober's criticism is that group-level adaptations fail to be recognized in the absence of altruism (Wilson and Sober 1994), and it seems therefore that they would themselves agree that the operation of group selection has been fairly transparent in those situations (imagined or real) where it worked against the forces of individual selection (i.e., in cases of altruism).

To recapitulate our discussion of (1), the objection from kin selection was met by defining altruism$_e$ in terms of inclusive fitness, and by thus showing that the alleged counterexample of selectively advantageous altruism toward relatives does not count as altruism at all. The objection from group selection was answered by pointing to its rare occurrence and by concluding therefrom that it is improbable that a massive presence of human altruism could be adequately accounted for by such a delicate and extremely fine-tuned causal mechanism.[5]

Claim (2)

Claim (2)—that altruism$_p$ tends to lead to altruism$_e$—would be trivially true if there were no difference in meaning between the two altruisms. This would in fact amount to sliding back to the crude version of the paradox of altruism, dismissed at the beginning but not wholly without adherents. For example, even Edward O. Wilson did not hesitate to state that "altruism is defined in biology, *as in everyday life,* as self-destructive behavior for the benefit of others" (Wilson et al. 1973, p. 953; emphasis added; repeated verbatim in Wilson et al. 1977, pp. 458–59).

Most authors, however, are most of the time aware that a connection between evolutionary and psychological altruism is not so immediate and purely semantic. The misfit arises on two counts: evolutionary altruism is defined in terms of actual effects with respect to fitness,

5. Later (see section titled Version [4.3]) I shall consider another approach which I have classified as a reconciliationist view but which can with no less justification be regarded as a way of denying (1).

whereas psychological altruism is defined in terms of intended effects with respect to personal interests. Obviously, then, a conjectured link between two kinds of altruism can break either because intended effects of human action do not in general correspond to actual effects or because the effects expressed in the currency of interests do not correspond to the effects as measured by fitness. Therefore, in order to make good claim (2)—that psychological altruism tends to lead to evolutionary altruism—we have to show (*a*) that human goals tend to be realized as intended, and (*b*) that there is some correlation between interests and fitness.

a) What matters evolutionarily is how intentions are actualized, not what they are inside the mind. If it were literally and massively true that the road to hell is paved with good intentions or, to put it less metaphorically, if human actions happened in general to have effects contrary to those envisaged by the subject, psychological altruism would cease to create any evolutionary puzzle. In that case, altruistic$_p$ acts would be favored by selection, for they would deviantly and perversely promote the self-interest of the agent against his own will. But as a matter of fact intentions are not so wholly impotent and disconnected from reality. More often than not they are realized according to the plans of the agent. Having an intention to ϕ leads to ϕing, other things being equal. And, frequently, other things are equal. So, although altruistic$_p$ acts are defined in terms of altruistic intentions rather than in terms of actual effects, altruistic intentions tend in reality to produce genuinely altruistic consequences.[6] For this reason, the gap between intentions and their realization offers a poor basis for attacking (2).[7]

b) An alternative and more promising route for disputing (2) is by driving deeper a wedge between interests and fitness. A thin end of the wedge is already there: the advancement of one's interests does not always coincide with the increase of one's fitness. For example, a couple's decision not to have children may serve their interests (say, because of the high probability of their life being destroyed by the birth of a seriously handicapped child), although this decision may significantly decrease their fitness (by their forgoing a nonzero chance of having perfectly healthy offspring). Conversely, too, one's interests may occasionally be harmed by acts which happen to increase one's fitness. All this, however, is not sufficient to undermine (2). It merely

6. I am here indebted to Kavka 1986.

7. There is a possibility, however, that despite intentions being typically realized according to the agent's plans, psychological altruism may still not lead to evolutionary altruism because, completely unbeknownst to the subject, these genuinely altruistic$_p$ acts just regularly happen to have fitness-increasing consequences as an altogether unintended by-product.

shows that there is no perfect match between interests and fitness, and this is fully compatible with (2) which was deliberately formulated as the statement of tendency. To vindicate (2), we do not even need to start by defining the concept of interest—which is very fortunate in the light of the notorious murkiness of this notion in the social science literature. Indeed, we can avoid the general issue of what interests are by simply confining our whole effort to showing that some important particular interests (counting as such on any definition of interest) have a systematic connection with fitness. Paradigmatic cases of this kind of interest are protection of one's health, avoidance of danger, keeping one's possessions, and so forth. So, again with a ceteris paribus clause, it is obviously against one's interest to act so as to destroy one's health, to bring oneself into a dangerous situation, to lose one's possessions, and so forth. In addition, it is easily recognized that such acts by themselves lead naturally to the loss of inclusive fitness (provided, of course, that they do not benefit close relatives of the agent). Therefore, we gain support for our belief that, at least with respect to some basic interests, the psychological propensity to sacrifice one's own interests tends to produce an evolutionarily self-defeating disposition to lower one's fitness.[8]

Claim (3)

Statement (3), the affirmation of the existence of psychological altruism, can be denied in two ways: by claiming (*a*) that it is necessarily false or (*b*) that it is contingently false.

a) The idea that there cannot be altruistic$_p$ acts is at the core of a philosophical thesis known as "psychological egoism." (The name derives from the need to distinguish it from another kind of egoism, "normative" or "ethical" egoism.) The main support for psychological egoism comes from a hedonistic interpretation of human motivation. Stripped to essentials, the argument is as follows: we are moved to action solely by the expectation of pleasure; attaining pleasure is a purely selfish aim; ergo, our behavior is never altruistic. In the opinion of many philosophers this view has been conclusively refuted already by Bishop Butler in his sermons (first published in 1726).[9] There are two basic difficulties for psychological egoism. On one hand, there are

8. It should be stressed here that the connection between interests and fitness is not only probabilistic but that it is also context dependent, and that it can easily break with changes in the environment. To borrow an example from Herbert Simon (1993, p. 158), although wealth was perhaps a major contributor to fitness in earlier centuries, in the contemporary Western societies there is a negative correlation between income level and reproduction rate; the statistics show that, as a matter of fact, the poor are fitter than the rich.

9. See Butler 1983, pp. 47–49.

many actions that we can only with great strain and implausibility reinterpret as being a search for pleasure and self-gratification. Who would not agree with Chesterton (quoted in Feinberg 1971, p. 498) that a philosopher is misusing the word 'self-indulgent' if he says that a man is self-indulgent when he wants to be burned at the stake? On the other hand, even with respect to some actions that do result in obtaining pleasure it seems demonstrably not true that they are undertaken in order to obtain pleasure. If I give $100 to charity this may give me a pleasant feeling of being a generous person. But I cannot be pleased with this act of generosity if I know that my contribution was exclusively stimulated by the anticipation of this pleasure. For in that case I am not being generous at all, and consequently I have nothing to be pleased with. In Jon Elster's terminology (Elster 1983, pp. 43–108), pleasure is often "a state that is essentially a by-product," and on that account hedonism cannot be the whole story about human motivation.

But taking this line is perhaps trying to prove too much. In order to put psychological egoism into doubt we are actually not obliged to make a positive step and produce a philosophical argument establishing the existence of at least some nonegoistically motivated desires. Rather, it is entirely sufficient to show, purely negatively, that the aprioristic argument for general egoism is unconvincing.[10] But this is obviously not such a formidable task anymore. Even those philosophers who are skeptical about the positive achievements of the Butlerian argument (see, e.g., Sober 1992) would certainly not be willing to argue that human altruism can be excluded on analytical grounds, by merely investigating the nature of practical reason.[11]

Psychological egoism is now too often serving just as the last resort to some ardent sociobiologists, when they find themselves confronted with an ostensibly altruistic act, a living counterexample to their simplistic theory of human behavior. Acutely challenged by this phenomenon, but unable to reduce it either to reciprocity ("soft-core altruism") or to helping one's kin ("hard-core altruism"), they simply fall back on the entirely aprioristic thesis of general egoism, and by appeal to this defunct philosophy they hope to explain away all these remaining

10. If the issue is to be decided empirically, purely conceptual or philosophical arguments (like psychological egoism) are *eo ipso* run out of court.

11. Bernard Williams suggests the following empirical test of altruism: "a man might be faced, by some manipulator, with the choice between the following: on the one hand, that *p* should be the case later but that he (the subject) should after a few minutes believe that not-*p*; on the other hand, that not-*p* should be the case later, but that he, after a few minutes, should believe that *p*. No conceptual manoeuvres could possibly persuade a man who wanted that *p* that he had to choose the latter alternative. If *p* involved someone else's welfare, this set-up could constitute something of a test for altruism" (1973, p. 262).

recalcitrant cases as well. For a typical move in this vein see Wilson (1978, p. 165), and for the criticism that is exactly on target see Kitcher (1985, pp. 402–3).

b) An alternative attack on (3), the attempt to show that it is contingently false, relies on empirical argument. In the case of any prima facie altruistic behavior the task is here to uncover the presence of some concealed motivation that changes completely the initial picture, and that, when taken into account, turns the behavior in question into a purely egoistic action. For instance, some acts of blood donation may on closer inspection prove to be motivated less by a wish to help others, and more by trying to improve one's social image. Surely, all that glitters is not altruism. Yet it is another thing, entirely, to claim that every appearance of altruism is deceptive and that some strong and dominating selfish motives are always there to be found to favor an egoistic interpretation. True, the route to a belief in universal egoism is facilitated by the fact that apparently altruistic acts regularly result in the reduction of negative arousal, the avoidance of external and internal punishments, or in a mood enhancement, and it is then surely legitimate to suspect that all the seemingly shining examples of human selflessness may in reality be motivated by a narrow-minded and purely egoistic anticipation of such likely consequences of our acts. But, surprisingly enough, the egoistic hypothesis is here meeting formidable empirical difficulties. In various ingeniously devised laboratory experiments (see Batson 1991, 1992; Batson and Oleson 1991; and in particular Batson and Shaw 1991, and the discussion of their target article in the same issue of *Psychological Inquiry*), Daniel C. Batson has created the situations where, exceptionally, each of the standard egoistical accounts gives opposite predictions from the altruism hypothesis, and he has also shown, most important, that it is the altruism hypothesis that consistently comes out as the winner in these decisive confrontations. Batson's research program is today the most serious challenge to psychological egoism, and quite possibly its burier too.

Claim (4)

If claim (4), that psychological altruism is a product of natural selection, is dropped psychological altruism ceases to have any puzzling evolutionary features. For like other traits that lower the fitness of their possessors, it presents no problem for evolution unless it is believed to have arisen by natural selection. Are there good reasons, though, to believe this?

Let us right away admit that at the present stage of the debate there is no orthodox or commonly accepted account of how psychological altruism was maintained by the forces of selection. (If there were such an account the paradox of altruism would immediately dissolve; for we would then be in the position to know how altruism$_p$ was

selected despite its connection with the selectively inferior altruism$_e$.) What we do have at the moment are just various speculations about the evolutionary origins of psychological altruism, the hypothesized Darwinian histories with different degrees of plausibility. Most important, our belief in general thesis (4) does not draw its strength from any of these particular selective accounts being very convincing or imposing its truth on us. On the contrary, it seems that each of these selective accounts was first and foremost prompted precisely by the hunch that (4) must be true, that is, that there has to be some evolutionary explanation for the genesis of psychological altruism.

To put the matter differently, although we lack direct evidence in favor of (4)—because there is no generally accepted causal story about how altruism$_p$ was produced by natural selection—there is nevertheless weighty indirect evidence supporting it. The very nature of the altruistic$_p$ predisposition in humans—the fact that it is so widespread, that it extends at least as far back in time as to the period of hunter-gatherers, that it is sustained by powerful emotions, that it is already present in very early childhood (Schwartz 1993, p. 322), and that it occupies such a manifestly central place in human mentality—all this taken together strongly suggests that altruism has its roots in our evolutionary past. This is not a knockdown proof, but in the opinion of many scholars such considerations carry enough weight to regard (4) as much more than just a fruitful working hypothesis. By way of illustration, John Rawls confidently states that "the theory of evolution would suggest that [human nature] is the outcome of natural selection," and that his postulated source of altruistic behavior, "the capacity for a sense of justice and the moral feelings is an *adaptation* of mankind to its place in nature" (Rawls 1972, pp. 502–3; emphasis added). Alan Gibbard is just one among contemporary empiricist philosophers who view the fact that beings with a sense of justice seem to fare worse than pure egoists as crying out for an evolutionary explanation: "What kinds of psychological propensities are involved [in a sense of justice], and how might evolutionary theory explain humans' having those propensities?" (Gibbard 1982, p. 33). Alexander Rosenberg expresses a widely shared view when he writes that "one is tempted to . . . say that *the only likely* explanation of why Homo sapiens cooperate, despite the temptations of costless free riding, must be evolutionary" (Rosenberg 1988, p. 832; emphasis added). These and similar quotations (see also Darwin 1874, pp. 149–50; Mackie 1977, pp. 113, 192) are not meant as an appeal to authority, but as an illustration of a growing consensus that even if proposition (4) is going eventually to be rejected the possibility of its truth should at present be taken very seriously indeed.

In conclusion, it is worth stressing that our basic problem does not go away even if we assume that human altruistic predispositions

have nothing to do with biology, and that they are merely a product of "culture." For in the process of cultural selection it is again the case that fitter cultural variants will tend to be preserved, so that it would still remain pretty mysterious how a selectively unfavorable trait like altruism was not stamped out a long time ago by the forces of cultural selection. Appealing to the claim that the cultural inheritance of altruistic traits fulfills a useful function at the social level (that it is "socially functional") is definitely insufficient, even when true; as already said, without pointing to possible mechanisms by which the allegedly functional state could be attained, such appeals easily degenerate into pseudoexplanations.[12] Expressed more concisely, the perseverance of an altruist "meme" in a population is no less puzzling than the perseverance of an "altruist" gene.

SECOND STRATEGY: RECONCILIATION

The consideration of the four propositions making the incongruous tetrad has shown that each of these claims is a serious candidate for truth. Moreover, many people are inclined to think that the evidence for them is collectively compelling, and that all these propositions ought to be jointly embraced. This leads to the increased pressure for finding a workable reconciliationist solution of the paradox. Perhaps the four propositions are in reality cotenable?

Pursuing this path, our attention is immediately directed to (4), the claim that altruism$_p$ is a product of natural selection. Referring to Sober's (1984) well-known and important distinction between "selection of" and "selection for," the truth of (4) obviously entails that there was selection *of* altruistic$_p$ organisms but not necessarily that altruism$_p$ was also selected *for*. To put it differently, a trait can be selected although it does not confer any selective advantage on its bearers, or, what comes to the same thing, without its having any adaptive value. This can happen in two crucially different ways: either a trait may have ceased to be adaptive, albeit it was itself one selected for in an earlier, different environment, or else it may never have been an adaptation but it was nevertheless selected through being tightly connected to some other adaptive trait that was the true target of selection. The first possibility is that a selected feature be only a relic of a past adaptation while the second possibility amounts to its being only a by-product of a still-adaptive trait. In both cases we see that a selectively not advantageous (or even disadvantageous) trait can be a product of selection.

12. For an interesting dual-inheritance scenario positing one such mechanism see section titled Version (4.2*b*), and for a purely cultural account of the emergence of altruistic norms of behavior see Allison (1992).

This is actually what makes room for a reconciliationist move. That is, even after granting (1) that $altruism_e$ is a fitness-decreasing trait, (3) that $altruism_p$ exists, and (2) that $altruism_p$ leads to $altruism_e$, it is still possible to claim without inconsistency (4) that $altruism_p$ is a product of selection. For, $altruism_p$ may have been selected in a different environment in the past when it did have fitness-increasing consequences, or it may have been selected, despite its selective disadvantage, by being inseparably tied to another, strongly adaptive trait (which more than compensated for its own harmful effects).

Here are these two versions of (4), concisely formulated:

(4.1) *Vestige theory:* $Altruism_p$ is a product of natural selection, *but it was adaptive only long ago.*

(4.2) *By-product theory:* $Altruism_p$ is a product of natural selection, *but it was never adaptive.*

There is also a third, strongest version of (4), according to which $altruism_p$ not only was adaptive but has until now preserved its adaptive qualities:

(4.3) *Continuing adaptation theory:* $Altruism_p$ is a product of natural selection, *and it is adaptive.*

On the face of it, the set containing (1), (2), (3), and (4.3) may appear to be formally inconsistent. (How could $altruism_p$ possibly be selected for, if it leads to $altruism_e$ which is selectively disadvantageous?) Therefore, there is a strong temptation to see (4.3) as having no place in discussing the reconciliationist strategy. In a way this is exactly right. Yet in another sense (see n. 21 for clarification) it is just here that we encounter a most ingenious reconciliationist argument for resolving the paradox of altruism. The three different versions of (4) are given in figure 2. Let us consider them in turn.

(4.1) Vestige Theory

We can believe that a given feature was produced by natural selection and yet be completely in the dark about how it was produced. This point was made by Mark Ridley and Richard Dawkins: "Civilized human behavior has about as much connection with natural selection as does the behavior of a circus bear on a unicycle. . . . Similarly, there probably is a connection to be found between civilized human behavior and natural selection, but it is unlikely to be obvious on the surface" (Ridley and Dawkins 1981, p. 32). There is little doubt that the capacity which enables the bear to keep his balance on a unicycle was shaped by natural selection, but it can only be a joke to suggest that the very skill of riding a unicycle made bears better adapted to their environment.

Applying this to our case, one can argue that for the question whether $altruism_p$ is a product of selection it is irrelevant to consider

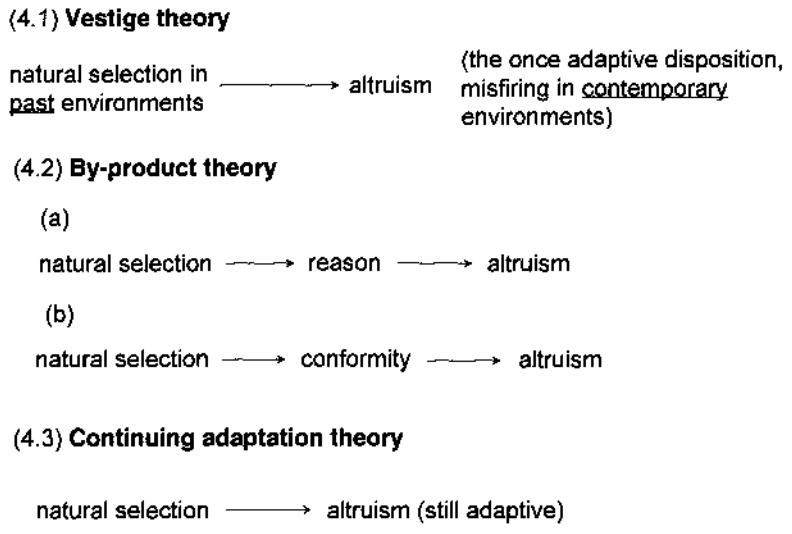

FIG. 2

its consequences for fitness in the contemporary environment.[13] If this trait was selected at all, it must have been so tens of thousands years ago under greatly different circumstances. Hence the fact that altruism_p is now indeed disadvantageous (via leading to altruism_e which decreases fitness) does not tell decisively against the hypothesis of selection because it is possible that altruism_p was selected by having been evolutionarily advantageous under very dissimilar conditions obtaining in the distant past.

The only story I know of that tries to flesh out this explanation sketch is the kin-selection account. The idea is simple: if humans once lived in small groups consisting mainly of close relatives, kin selection would favor indiscriminate altruism. For, under these circumstances the beneficiaries of altruistic acts would be almost exclusively the close relatives of the "altruist," and it is easily understood how such a trait could be brought to fixation by selection. Moreover, it can be shown that, against such a background, unrestricted altruism, might be *on informational grounds* (see Sober 1981, p. 104) fitter than discriminating between relatives and nonrelatives, and then helping only the former: in a group where any interaction is most likely to be with a close relative, a mechanism for distinguishing relatives from nonrelatives

13. "Our genes gave us the propensities we had at conception—propensities to have certain characteristics in various hunting-gathering environments. That tells us nothing directly about what we are like in fact, in our own environments" (Gibbard 1990, p. 27).

would be all but useless; besides, it could be acquired only at some cost (in evolution, no less than in economics, there's no such thing as a free lunch). With a sufficiently low probability of encountering a nonrelative the usefulness of being able to recognize relatives would become so small that the costs of acquiring such an ability would have to be greater than the gain. It would then always pay to be an indiscriminate altruist: better to make an occasional, very rare error (to aid a nonrelative) than to carry the costs of an expensive error-avoiding mechanism which practically does no useful work.

In the situation as described indiscriminate altruism is actually serving as a cheap ersatz for kin selection. Under altered circumstances, however, with humans living in large communities and with a lot of migration, altruism loses its evolutionary justification. It becomes a deleterious trait, a mere vestige of an adaptation. In this way, there is no more puzzle about how altruism was selected despite its presently being selectively disadvantageous. This explanation is gestured at by Tooby and Cosmides (1989).

It is a nice story, but it is doubtful whether it is anything more than that. The skepticism originates from two sources. First, and most important, the starting assumption that humans once lived in groups consisting mostly of relatives is, to use an understatement, very far from being generally accepted; various lines of evidence suggest that our Pleistocene ancestors lived in fairly large groups consisting of approximately 150 members (Dunbar 1994, p. 770). With the key assumption so empirically compromised the whole approach can hardly move off from the ground.

Second, it is not clear that indiscriminate altruism would be an evolutionarily best strategy, even if the postulated conditions obtained. True, under the conditions it would not be worth the effort to try to recognize relatives from nonrelatives (because ex hypothesi there would almost be no nonrelatives), but even then it would certainly make much evolutionary sense to distinguish between relatives with different degrees of relatedness (and, consequently, to help the closer kin more). However, once the behavior is guided by the variable degree of relatedness nonrelatives are automatically "perceived" and treated as having coefficient r of zero, and we are back to square one: altruism toward strangers remains inexplicable. The basic difficulty here is that in the situation where one interacts mainly with relatives one is well advised on evolutionary grounds to be an altruist, but not a nondiscriminating altruist. Unfortunately, the theory under consideration depends crucially on the presence of nondiscriminating altruism. For, it says that in the past it was fitness increasing to help anyone, whomever you met (without bothering to establish who the individual was), because, anthropomorphically speaking, you could have been practically sure in advance that he would turn out to be your relative.

Altruism is indeed explained by the fact that such a behavioral disposition misfires in a changed environment when your group expands and when it comes to consist predominantly of biological strangers. What is thereby not explained at all is why there was such a blanket altruism in the first place, that is, why the readiness to help others was not adjusted and apportioned according to the degree of genetic proximity. No doubt, this could be accounted for, too, by complicating the story and by introducing additional hypotheses. But this modified and more demanding explanation has yet to be satisfactorily elaborated in detail.[14]

(4.2) By-Product Theory

According to (4.1) the conflict between altruism$_p$ appearing to be counteradaptive and its also appearing to be a product of natural selection was resolved by the claim that, due to the change in the environment, altruism$_p$ ceased to be adaptive. In contrast, (4.2) claims that it was at no time an adaptation, but that it was nevertheless selected by having been inseparably connected with some other trait, which was adaptive. This amounts to the idea that altruism$_p$ is a by-product, or spin-off (or spandrel), of selection.

To defend this kind of approach one is under a threefold obligation: (i) to identify some other trait, (ii) to show that it was selected for, and (iii) to demonstrate that it is inextricably tied to altruism$_p$. I shall consider here two instances of such an approach.

a) According to an influential argument, the emergence of rationality is easy to incorporate into an evolutionary scenario. Those who were more rational (i.e., those who made fewer systematic errors when thinking or solving problems) were likely to cope better with their environment, and to leave more descendants than others. Hence, according to some philosophers (Daniel Dennett, for example) there is a strong case for regarding the faculty of rationality as an adaptation.[15] With rationality thus evolutionarily fortified one can then attempt to derive altruism as a consequence, by making use of a classical Kantian argument in practical philosophy. In a nutshell, the argument purports to show that being rational entails being concerned for the inter-

14. One possible move in that direction (suggested both by Elliott Sober and by the anonymous referee for *Ethics*) is to argue that the informational costs of distinguishing various degrees of relatedness would be much higher than making the "relative/nonrelative" distinction, and that perhaps the costs would be too high to make it worth doing in an all-fairly-close-relative society. But then again, it seems that the theory would be shipwrecked on the hard empirical fact that, after all, primates actually happen to have the capacity to discriminate between relatives and nonrelatives (Dunbar 1994, p. 773).

15. For doubts about this line of reasoning see Stich (1990, pp. 55–74).

ests of all rational beings, or, by contraposition, that the lack of such a concern (being completely self-interested) is a sign of irrationality. On this view, a rational person cannot be a complete egoist.

The ablest contemporary defender of this Kantian line is Thomas Nagel (1970). At one time he even went so far as to argue that one who in full awareness did not care for the interests of others had to be a solipsist. Put differently, he thought that by merely recognizing the existence of other people a rational person is necessarily committed to have at least some minimal concern for their interests. (Later, under criticism, Nagel was forced to weaken his claim significantly; cf. Nagel 1986, p. 159.) Nagel was not concerned with an evolutionary account of altruism; he only wanted to show that out-and-out egoism is incompatible with rationality. It was Colin McGinn who first openly combined this view that altruism is a consequence of rationality with the thesis that human rationality is a biological adaptation, carrying thereby an explicit suggestion that altruism is a by-product of natural selection. There is also an obvious (and acknowledged) debt to McGinn's ideas in Peter Singer's (1981) book *The Expanding Circle,* where this hypothesis is worked out in more detail. The gist of McGinn's argument stands out in the following sentences:

> If morality is founded upon naturally bestowed appetites in accordance with the principles of natural selection, and if these appetites simply cannot, consistently with the laws of evolutionary biology, extend as far as moralists have insisted, why then surely the idea of pure, disinterested altruism is a chimera which it is pointless to pursue. . . . If we want to secure morality against the forces of natural selection, *we need to associate it with possession of some characteristic whose evolutionary credentials are undisputed:* I suggest that *the cognitivist's associating it with reason meets this condition,* while the noncognitivist's appetitive theory does not. (McGinn 1979, pp. 85, 93; emphasis added)

The argument has two steps: the first biological (that rationality is an outcome of an evolutionary process), and the second philosophical (that rationality leads to altruism). It is the second step that does main explanatory work, and that raises most questions. Since, for obvious reasons, this philosophical claim cannot here be given the full consideration it deserves, the following brief comment must suffice. Judged by the present state of metaethical discussions, the strong Kantian version of cognitivism that ascribes concern for others to any rational being qua rational being is highly controversial among philosophers. The basic difficulty with it is well described by Peter Railton: "Although rationalism in ethics has retained adherents long after other rationalisms have been abandoned, the powerful philosophical currents that have worn away at the idea that unaided reason might afford a standpoint from which to derive substantive conclusions

show no signs of slackening" (Railton 1986, p. 163). But if this kind of cognitivism is so problematic even in its own philosophical province, it is then surely all the more unsuitable as an instrument for reaching not strictly philosophical conclusions.[16]

In addition, its standing is not exactly improved by the fact that much social science literature (rational choice theory, decision theory, microeconomics, public choice theory) takes as its starting point the assumption that reason is motivationally inert. According to this widely shared view, rational considerations cannot move us to action by themselves, the main impulse always coming from some of our basic preferences that stand completely outside the jurisdiction of reason. In David Hume's memorable words (1888, p. 416): "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger." Many contemporary cognitivists recognize the force of noncognitivist arguments, and they try to save cognitivism by weakening it significantly. Consequently, they are now often ready to concede that our moral attitudes cannot derive from rationality alone, and that what is needed besides is some kind of moral sensitivity, minimal concern for others, and the like (cf. Williams 1972, p. 26; Singer 1981; Lindley 1988, p. 528). After this concession, however, our fundamental question immediately reappears: How did these initial and rudimentary nonegoist dispositions (that are a prerequisite for full altruism) arise by the process of natural selection?

A possible answer might be that a budding altruistic disposition (in the psychological sense) could have evolved through the process of kin selection without being altruistic in the evolutionary sense.[17] According to our two definitions of altruism the tendency to help one's own relatives counts as altruism$_p$, but not as altruism$_e$. So, it seems after all that kin selection could provide a mechanism for injecting the first, minimal dose of altruism$_p$ and that by conferring a selective advantage on the carriers of this trait it could then set the stage for the circle of altruism expanding further afterward (see Singer 1981, p. 91).

The point that there is this cleavage between the two concepts of altruism (i.e., that altruism$_e$ is about genes whereas altruism$_p$ is about individuals) is both cogent and important. It shows that psychological

16. It is interesting to note here that a long time ago Charles Sanders Peirce tried to reach the same conclusion about the inherent irrationality of out-and-out egoism, by taking a completely different route. He developed a highly idiosyncratic argument from the nature of probability purporting to prove that "to be logical men should not be selfish" (see Peirce [1878] 1992, pp. 149–50). I am indebted to Michael Kremer for drawing my attention to this article.

17. This idea is defended in Kitcher (1993, pp. 508–9) and Sober (1994, pp. 18–19).

altruism toward relatives can be a purely Darwinian product. What remains highly dubious, however, is whether this kind of incipient nonegoism could serve as a foothold for our reason in its ascent to the completely generalized altruism. If we want to justify rationally the normative standpoint of universal altruism this surely cannot be achieved by relying solely on the purely factual premise that humans already display a kind of selective altruism. Those philosophers who want to claim that it is our reason that helps us to cross the border between the narrow-scope altruism and principled altruism are under obligation not only to show (*a*) that the border between the two altruisms is arbitrary, and (*b*) that we already happen to be narrow-scope altruists. They have also to establish (*c*) that this initial, minimal altruism is itself rationally justified. This is necessary simply because reason cannot generalize something on the ground that it exists, but only on the ground that it is reasonable in the first place. Therefore the cognitivist argument, labeled here (4.2*a*), cannot bootstrap itself by appeal to the factual premise that evolutionary forces have produced one kind of altruism$_p$, in the hope that it has then only to proceed further and broaden the scope of this other-benefiting behavioral tendency. No, reason has to take the uphill path and develop the rational defense of altruism all the way from the very beginning. A further difficulty for connecting altruism so closely with rationality is that, according to recent research in child psychology, the concern for others is in some forms already present at the developmental stage when children are definitely incapable of moral reasoning. (See Wilson 1993, p. 130, and references given therein.)

b) In a diametrically opposed approach, Robert Boyd and Peter J. Richerson (1985, pp. 204–40; 1990) suggested that it is actually a tendency not to use one's reason that links altruism with natural selection. They argued, first, that under certain conditions there is a selective advantage in uncritical conformity, that is, in simply copying the most common behavior in a subpopulation, without checking beforehand whether that behavior is also the most appropriate to the circumstances. It can be shown, particularly in heterogeneous environments, that due to selection the most fit behavioral variant tends to become the most common one in a subpopulation as well; therefore, by being a conformist (by merely imitating the behavior most frequently encountered in the environment) an individual would increase the chance of acquiring the best behavior without costly individual learning. In this way it is explained how a psychological disposition to conform (to behave according to the rule *A Roma alla romana*) is selected for.

The second step is to derive altruism as a consequence of conformity. The mechanism proposed for this purpose is cultural group selection. Assuming that egoism and altruism are behavioral traits that

are culturally transmitted, Boyd and Richerson are in the position to demonstrate that the aforementioned genetic disposition to conform is a key factor which makes groups mainly consisting of altruists evolutionarily stable and resistant to change. Namely, if individuals have the built-in propensity to adopt whatever behavior is the most common in their environment, then groups with predominantly altruist members will in all likelihood, through the process of cultural conformist transmission, give rise to new generations consisting mainly of altruists. Conformity plays here a crucial role, making it possible for groups of altruists to persist through time long enough for the process of (cultural) group selection to take effect. As we saw earlier (pp. 132–33), one of the main obstacles to the genetic group selection of altruism was that altruist groups were open to subversion from within (they were easily invadable by the small number of genetic egoists). In the case of cultural group selection, as described by Boyd and Richerson, the obstacle disappears. True, a few occasional egoists would still fare consistently better than the more numerous altruists in the group, but they would not threaten the predominance of altruism: owing to the special character of cultural inheritance (conformity) even the descendants of egoists would tend to resemble the average type; that is, they would tend to be altruists. In short, egoism continues to be individually advantageous, but the main reason why it cannot spread in groups of altruists is that it tends not to be transmitted (because of the conformist bias).

A major virtue of the Boyd-Richerson approach is that the emergence of altruism is not seen as being the result of a single causal influence (biology or culture). Trying to do justice to the complexity of the issue, they proposed that the genesis of altruism should be explained by the combined operation of natural and culural selection. In their model, these two causal factors are combined in an original, elegant, and initially plausible way. It is an open question, however, whether all the empirical assumptions underpinning the model will be borne out by future scientific research.[18]

(4.3) Continuing Adaptation Theory

This theory says that altruism_p is a product of natural selection, and that it is adaptive. As noted earlier, adding (4.3) to (1), (2), and (3) gets us on the brink of contradiction. Assuming (1) that altruism_e is a selectively disadvantageous trait, (2) that altruism_p leads to altruism_e, and (3) that altruism_p exists, it is hard to see how it is possible (4.3) that altruism_p is adaptive (i.e., selected for). Or, to put the question in general terms and in a more pointed way, can we coherently entertain the idea that it may be selectively advantageous to possess a selec-

18. A similar account is proposed by Herbert Simon (1990).

tively disadvantageous trait? With the caveat already mentioned (and spelled out in n. 21) this idea in fact proves to be explanatorily very fruitful and promising.

Consider first the simple Prisoner's Dilemma game between two players, A and B, exhibited in figure 3 in the so-called extensive form. Each player has a choice between cooperating (c) and defecting (d). There are four possible outcomes with different payoffs for A and B (with A's payoffs always being given first). Suppose that A acts first and that B makes his choice after gaining knowledge about whether A has cooperated or defected. This slightly modified version of the Prisoner's Dilemma doesn't make much difference. Obviously, defecting is still the dominant option for both, and hence the solution of the game is (0, 0). Because of the structure of this decision problem the outcome (1, 1), although preferred by both players, is simply unattainable.

Let us now change the situation in two respects. First, suppose that B can acquire a disposition to cooperate if A cooperates. And second, suppose that A has a relatively reliable (though not necessarily infallible) method for recognizing the presence of such a conditional behavioral disposition in B. That is, B cannot hope to deceive A that he is disposed to respond with cooperation to A's cooperation when he is not so disposed. Or, still differently but equivalently, if we call the disposition in question N, we are supposing that B's best way to persuade A that he has N is to really acquire N.

The conditions of choice are thereby essentially changed. Now A expects that B with a built-in N disposition will respond with cooperation to cooperation, and with defection to defection. So, confronted with such a "tit-for-tat" opponent, A actually faces a dilemma between cooperating, with the certain final outcome (1, 1), and defecting, with the certain final outcome (0, 0); clearly, if he is rational he chooses the former, and the collectively preferred state becomes attainable. But it is B's perspective that deserves our attention.

Assuming that payoffs are measuring fitness consequences of different outcomes for A and B, it is easily seen that manifestations of disposition N decrease B's fitness. Namely, after A has cooperated it is patently in B's evolutionary interest to defect (payoff of defection = 2) and not to cooperate (payoff of cooperation = 1). If, however, by possessing disposition N, B responds with cooperation this is an act of altruism$_e$ pure and simple, selectively disadvantageous and lacking any evolutionary justification. Nevertheless, despite the fact that disposition N is in a sense a systematically fitness-decreasing trait (i.e., every manifestation of it leads to the net loss of fitness) the possession of disposition N can still be selected for.

To see this, note that it is in B's interest to induce A to cooperate, because he is then sure of getting at least 1 whereas if A defects, B can get no more than 0. Indeed, let us assume that B can induce A to

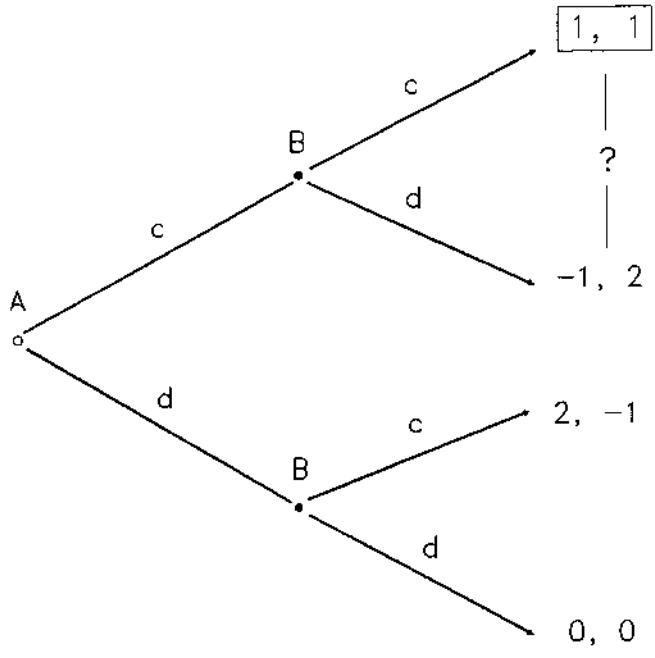

FIG. 3

cooperate only if he comes to persuade A that he (B) has an "altruistic" disposition N to respond in kind to A's cooperation. If, furthermore, the situation is as previously described, and B's best way of persuading A that he has disposition N happens to be the actual possession of N, B's fitness will in fact be enhanced by acquiring N. Namely, if B does not have N, the decision problem reduces to the standard Prisoner's Dilemma, and the equilibrium is (0, 0). If B does have N, however, he thereby makes A notice this and by assuring A that he is not a "cheater" B induces A to cooperate; finally, in accordance with N, B in turn also cooperates, and the outcome is (1, 1).

Disposition N is altruistic$_e$ (by our definition) because any manifestation of it (responding with cooperation to cooperation) results in the net loss of one point of fitness. By cooperating after A's cooperation B gets only 1 point whereas by defecting he could have acquired 2 points. So in a sense, N is evolutionarily self-defeating. But although any manifestation of N is fitness-decreasing, the possession of N can be fitness increasing. This can come about because good effects of having N can offset bad effects of manifesting N. In our situation, the optimal outcome for B is $(-1, 2)$ but it is inaccessible to him. If he lacks N (if he is not disposed to cooperate after A's cooperation) this will induce A to defect, producing the result (0, 0). If he has N,

however, this will motivate A to cooperate, but this will then also make B cooperate, leading to the suboptimal outcome from B's perspective (1, 1), instead of the optimal one (−1, 2).

B's predicament consists in the fact that by possessing N he is better off before A cooperates (because ex hypothesi it is only B's possession of N that can make A cooperate), but after A cooperates B would be better off if he lacked N (because, then, acting in accordance with disposition N he deprives himself of the optimal result). B's interest would be best served if he had disposition N only until the moment A cooperates, and if he lost it or at least did not manifest it afterward. But this hardly seems possible. For, as John Mackie remarked, "dispositions cannot be switched on and off in deference to the calculation of likely consequences on particular occasions" (1977, p. 192). By their very nature dispositions have some kind of persistence over time, and it may therefore well be that in our case, too, there is only a choice between taking or leaving the whole package (disposition N which has good effects first, but bad manifestations later). This reveals that even from a purely egoistic perspective it may sometimes be advisable to be altruistic. One's own self-interest may be best promoted by one's readiness to sacrifice it. The argument for this paradoxically sounding claim was first offered in David Gauthier's (1975) article "Reason and Maximization," and it was later more rigorously elaborated in his book *Morals by Agreement* (Gauthier 1986).[19] Finally, it was Robert Frank (1988) who in his *Passions within Reason* fleshed out this approach with rich empirical detail and showed that the whole idea of altruism being ultimately founded on egoism is not a mere abstract possibility but that it can have a very wide and surprisingly fruitful application in explaining human behavior.[20]

It is very important, however, not to confuse this approach with a completely different way of giving an egoistic rationale for altruism. That is, acting altruistically on a particular occasion can be egoistically justified by the fact that an agent who so acts gains thereby a reputation of an altruist, and this in turn may have good effects for him in making other agents more ready to interact cooperatively with him in the

19. "It would, after all, be paradoxical if the only way to justify a nonegoistic enterprise like morality were by the use of an egoistic argument" (Frankena 1980, p. 87).

20. Therefore, when Kenneth Binmore (1993, p. 138) says that "people cannot see inside each other's heads and [that] it is idle to examine models in which they can," he is simply wrong. Namely, beside many and various indications (usefully collected and described in Robert Frank's book) that the human mind systematically and unintentionally leaks information about its content to the outside it has recently been even experimentally demonstrated (Frank, Gilovich, and Regan 1993) that people can "see" inside each other's heads, i.e., that in playing the Prisoner's Dilemma game people somehow manage to recognize the presence of a cooperative disposition in others if allowed to interact with them even for as brief a period as half an hour!

future. Taking into account all the benefits that he could himself reap from these later cooperations just by first depriving himself of a much smaller immediate gain it is quite clear that only a very myopic egoist would refuse to cooperate under the circumstances. Seen from a wider perspective, such a conduct should not really be classified as altruism at all. Rather different but also purely egoistically inspired acts of cooperation in the iterated Prisoner's Dilemma (or in the so-called centipede game) would consist in cooperating with the sole purpose of appearing naive, stupid, or irrational "in the hope of tempting the opponent into an unwise attempt at exploitation" (Binmore 1988, p. 11).

In contrast, the behavior that interests us and that falls under description (4.3) is genuine altruism, for here we are assuming that the agent stands to gain nothing later by temporarily sacrificing his interests. For all that matters, there may simply be no interactions after the one we are considering. In that case, altruistic behavior is actually justified not by its subsequent effects, but instead by earlier beneficial effects of having the altruistic behavioral disposition. At the moment when the disposition manifests itself, however, the act is a genuinely altruistic one because by being "nice" the agent suffers a loss, never compensated afterward. To repeat, he would be best off if he could manage somehow both to acquire the altruistic disposition and to not let it ever be actualized. But this is not a feasible project. What is feasible and, indeed, best for the subject is to acquire the altruistic disposition although it is evident in advance that his interests will be harmed by this later. In this way a path is cleared for an evolutionary explanation of the genesis of altruism. The *nervus explanandi* is the claim that, under specified conditions, the possession of altruistic behavioral dispositions may maximize the fitness of its bearers.[21]

21. One might here object that this kind of "altruistic" disposition is no more selectively disadvantageous because organisms possessing such a disposition would definitely have a higher inclusive fitness than those lacking it. Consequently, it could be argued that the view (4.3) should actually be interpreted as denying the proposition (1) of the incongruous tetrad, and hence that it is more properly subsumed under the rubric of eliminativism than reconciliationism. This is a good point (made by an anonymous referee for *Ethics*, and by Gordon Belot in a discussion). Yet I have decided to retain my nomenclature for the following reason. Usually, the (dis)advantageousness of a behavioral disposition depends alone on whether its manifestations are advantageous or not. With the disposition in question, however, any organism possessing it would be better off never to manifest it (i.e., never to act in accordance with it): the good effects are in this case coming, not from the acts, but from the side effects of having the disposition. Here it is fitness increasing to have the tendency to produce behaviors that are all individually fitness decreasing. So, not in the least disputing the legitimacy of the proposed eliminativist interpretation of the view under consideration I want simply to point out there is also a secondary sense in which the crucial disposition

Someone could perhaps be tempted to argue here that there is a structurally similar decision-theoretical account of the emergence of altruism which is got by simply replacing 'behavioral disposition' with 'conditional intention', and 'fitness' with 'interest'. But what worked with dispositions does not work with conditional intentions.

Referring again to figure 3, let us assume now that A has a reliable way of recognizing not B's behavioral dispositions, but B's true intentions. Also, to cut verbiage, let us say that the condition C is fulfilled if A cooperates. Then, by the same argument as before, B is well advised as a rational agent to form a conditional intention to cooperate-if-C. Namely, if B forms that intention, A will recognize the presence of such an intention in B and this will motivate him to cooperate. B would thus gain at least 1 point, whereas otherwise (if B did not form the intention) A would defect, and B would be left with 0 points. But by forming the conditional intention to cooperate-if-C, B *eo ipso* insures that, after C is eventually fulfilled, he will then have the unconditional intention to cooperate. This is derived from the following intuitively plausible principle: (i) if at t_1 B forms a conditional intention to ϕ at t_3 in the event that condition C obtains at t_2, (ii) if C is realized at t_2, and (iii) if nothing intervenes, then B will have at t_3 an unconditional intention to ϕ at t_3.

It is precisely here that the basic difficulty comes. If we assume (as we did) that B is a completely self-interested and fully rational agent, then it is not clear how he can bring himself to intend to cooperate at t_3 when he definitely knows that at t_3 he can only lose by doing so. True, he knows that he can gain much by side effects of his forming the conditional intention to cooperate, but this is of no avail to him in the process of forming the intention: for, at t_3, good side effects (of A's cooperation) already belong to the past, and for B as a fully rational and self-interested chooser there is at that time simply no reason whatever to cooperate. But, of course, the fact that B knows all this in advance makes it impossible for him, even at t_1, to form the conditional intention to cooperate. For it is hard to see how B, who is driven only by his self-interest, could at t_1 form the intention to ϕ at t_3, when he is fully aware that when the time comes, at t_3, his interests would only be harmed by ϕing. (Illuminating discussions of this kind of decision-theoretic predicament are to be found in Kavka 1983; 1987, pp. 15–32; Bratman 1987, pp. 101–6.)

---

does not falsify (1), i.e., a sense in which it is selectively disadvantageous: namely, all its realizations are systematically and without exception selectively disadvantageous. I have let myself be guided by this secondary sense in classifying (4.3) because for expository purposes this solution to the paradox of altruism falls neatly into place at the end of the sequence of ever more astringent reconciliationist answers.

In the situation as described, B is, judging strictly by his interests, best off by forming the conditional intention in question, but the catch is that insofar as he is fully rational he cannot form that intention. Therefore, it is not only that reason doesn't pave the way for altruism; under the circumstances reason is a positive obstacle. To put it differently, although the indubitably best option for the agent, egoistically speaking, is to form a conditional intention to cooperate, a rational and self-interested person just cannot plan in full consciousness to form such an intention. If he remains both rational and self-interested when the time comes to act, his preferences, being as they are, will simply compel him to defect. But since he knows all this from the start there is something incoherent in the idea that he could (even conditionally) intend to cooperate.

In contrast to those who have hoped that only reason could bridge the gap between pure egoism and the moral point of view,[22] it is revealed here that, in some contexts at least, narrow selfishness can be transcended in no other way than by modifying the "nonrational" parts of the mind, and by natural selection working on the mental dispositions, habits, and emotions. This opens up an interesting possibility that, despite the notorious selfishness of its units of selection and its "blind" way of operation, biological evolution can still give rise to certain forms of altruism that are inherently unattainable even to infinitely intelligent selfish deliberators, as long as they remain fully rational. That is, a purely rational agent may happen to be stuck in the trough of myopic egoism, with his only chance of "tunnelling through" to the position of the enlightened self-interest (which here paradoxically coincides with genuine altruism) by Darwinian forces shaping his behavioral dispositions behind the back of his reasoning self.[23]

CONCLUSION

The main intention of this article was to propose a novel, "natural" classification of different approaches to the paradox of altruism, in the hope that by imposing the overarching structure on this continuing controversy the issue could be joined in a more fruitful way. I have argued that on the whole the so-called reconciliationist strategy holds more promise than the eliminativist one, and more specifically that among reconciliationist answers those designated as (*4.2b*) and (4.3) in my scheme are particularly well grounded and deserving further elaboration. I did not want, however, to exclude completely the possi-

22. Compare (*4.2a*).

23. These remarks about altruism and rationality are very sketchy, and they need a lot more spelling out. I hope to develop this line of thought in more detail on another occasion.

bility that some other of the discussed moves could account for the emergence of certain forms of human altruisms, although (very probably) only under very special or rarely satisfied conditions. We should also heed a warning (coming from Christopher Jencks) that "while it is analytically useful to label many different forms of behavior as ["altruistic"], the use of a single label encourages the illusion that there is a single underlying trait ["altruism"] that determines whether an individual engages in all these different forms of behavior" (Jencks 1990, p. 66). Although the opinion today prevails that at the core there is indeed one underlying, deep-seated behavioral disposition (perhaps shaped by evolutionary forces) that accounts for various manifestations of human altruism, at the present stage of our knowledge at least some room should be left for the possibility that the story will turn out to be more complicated and less orderly. For in the case (which cannot be ruled out a priori) that Jencks happens to be right in his hunch that different forms of altruism are only loosely connected to one another (and that there is simply no unifying trait or nucleus, below the surface), then searching for the explanation of the origins of human altruism would inevitably lead to Procrustean accounts of this multifaceted phenomenon.

REFERENCES

Allison, Paul D. 1992. The Cultural Evolution of Beneficient Norms. *Social Forces* 71:279–301.

Batson, Daniel C. 1991. *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, N.J.: Lawrence Erlbaum.

Batson, Daniel C. 1992. Experimental Tests for the Existence of Altruism. *PSA* 2:69–78.

Batson, Daniel C., and Oleson, Kathryn C. 1991. Current Status of the Empathy-Altruism Hypothesis. Pages 62–85 in *Prosocial Behavior*, ed. M. S. Clark. Newbury Park, Calif.: Sage Publications.

Batson, Daniel C., and Shaw, Laura L. 1991. Evidence for Altruism. *Psychological Inquiry* 2:107–22.

Binmore, Kenneth. 1988. Modeling Rational Players: Part II. *Economics and Philosophy* 4:9–55.

Binmore, Kenneth. 1993. Bargaining and Morality. Pages 131–56 in *Rationality, Justice and the Social Contract*, ed. D. Gauthier and R. Sugden. Ann Arbor: University of Michigan Press.

Boyd, Robert, and Richerson, Peter J. 1985. *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Boyd, Robert, and Richerson, Peter J. 1990. Culture and Cooperation. Pages 113–32 in *Beyond Self-Interest*, ed. Jane J. Mansbridge. Chicago: University of Chicago Press.

Bratman, Michael E. 1987. *Intentions, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

Butler, Joseph. 1983. *Five Sermons*. Indianapolis: Hackett.

Darwin, Charles. 1874. *The Descent of Man and Selection in Relation to Sex*. Chicago: Rand McNally.

Dunbar, Robin I. M. 1994. Sociality among Humans and Non-Human Animals. Pages 756–82 in *Companion Encyclopedia of Anthropology*, ed. T. Ingold. London: Routledge.

Elster, Jon. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.

Elster, Jon. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Feinberg, Joel. 1971. Psychological Egoism. Pages 489–500 in *Reason and Responsibility*, ed. J. Feinberg. Encino, Calif.: Dickenson.

Frank, Robert H. 1988. *Passions within Reason*. New York: Norton.

Frank, Robert H., Gilovich, Thomas, and Regan, Dennis T. 1993. The Evolution of One-Shot Cooperation. *Ethology and Sociobiology* 14:247–56.

Frankena, William K. 1980. *Thinking about Morality*. Ann Arbor: University of Michigan Press.

Gauthier, David. 1975. Reason and Maximization. *Canadian Journal of Philosophy* 4:424–33.

Gauthier, David. 1986. *Morals by Agreement*. Oxford: Clarendon.

Gibbard, Allan. 1982. Human Evolution and the Sense of Justice. *Midwest Studies in Philosophy* 7:31–46.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings*. Oxford: Clarendon.

Haldane, John B. S. 1932. *The Causes of Evolution*. London: Longman.

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

Hume, David. 1888. *A Treatise of Human Nature*, Selby-Bigge ed. Oxford: Clarendon.

Jencks, Christopher. 1990. Varieties of Altruism. Pages 53–67 in *Beyond Self-Interest*, ed. Jane J. Mansbridge. Chicago: University of Chicago Press.

Kavka, Gregory S. 1983. The Toxin Puzzle. *Analysis* 43:33–36.

Kavka, Gregory S. 1986. *Hobbesian Moral and Political Theory*. Princeton, N.J.: Princeton University Press.

Kavka, Gregory S. 1987. *Moral Paradoxes of Nuclear Deterrence*. Cambridge: Cambridge University Press.

Kitcher, Philip. 1985. *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge, Mass.: MIT Press.

Kitcher, Philip. 1993. The Evolution of Human Altruism. *Journal of Philosophy* 90:497–516.

Krebs, Dennis. 1982. Psychological Approaches to Altruism: An Evaluation. *Ethics* 92:447–58.

Lindley, Richard. 1988. The Nature of Moral Philosophy. Pages 517–40 in *An Encyclopedia of Philosophy*, ed. G. H. R. Parkinson. London: Routledge.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.

McGinn, Colin. 1979. Evolution, Animals, and the Basis of Morality. *Inquiry* 22:81–89.

Nagel, Thomas. 1970. *The Possibility of Altruism*. Princeton, N.J.: Princeton University Press.

Nagel, Thomas. 1986. *The View from Nowhere*. New York: Oxford University Press.

Peirce, Charles S. (1878) 1992. The Doctrine of Chances. Pages 142–54 in *The Essential Peirce*, ed. N. Houser and C. Kloesel. Bloomington: Indiana University Press.

Peressini, Anthony. 1993. Generalizing Evolutionary Altruism. *Philosophy of Science* 60:568–86.

Quine, Willard V. O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Railton, Peter. 1986. Moral Realism. *Philosophical Review* 95:163–207.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Ridley, Mark, and Dawkins, Richard. 1981. The Natural Selection of Altruism. Pages 19–37 in *Altruism and Helping Behavior*, ed. J. P. Rushton and R. M. Sorrentino. Hillsdale, N.J.: Lawrence Erlbaum.

Rosenberg, Alexander. 1988. Grievous Faults in *Vaulting Ambition*? *Ethics* 98:827–37.

Schwartz, Barry. 1993. Why Altruism Is Impossible . . . and Ubiquitous. *Social Service Review* 67:314–43.

Simon, Herbert A. 1990. A Mechanism for Social Selection and Successful Altruism. *Science* 250:1665–68.

Simon, Herbert A. 1993. The Economics of Altruism. *American Economic Review* 83:156–61.

Singer, Peter. 1981. *The Expanding Circle: Ethics and Sociobiology*. Oxford: Oxford University Press.

Sober, Elliott. 1981. The Evolution of Rationality. *Synthese* 46:95–120.

Sober, Elliott. 1984. *The Nature of Selection*. Cambridge, Mass.: MIT Press.

Sober, Elliott. 1988. What Is Evolutionary Altruism? Pages 75–99 in *Philosophy and Biology*, ed. M. Mathen and B. Linsky. Calgary: University of Calgary Press.

Sober, Elliott. 1992. Hedonism and Butler's Stone. *Ethics* 103:97–103.

Sober, Elliott. 1993a. Evolutionary Altruism, Psychological Egoism, and Morality: Disentangling the Phenotypes. Pages 199–216 in *Evolutionary Ethics*, ed. M. H. Nitecki and D. V. Nitecki. Albany: SUNY Press.

Sober, Elliott. 1993b. *Philosophy of Biology*. Boulder, Colo.: Westview.

Sober, Elliott. 1994. Did Evolution Make Us Psychological Egoists. Pages 8–27 in his *From a Biological Point of View*. Cambridge: Cambridge University Press.

Stich, Stephen. 1990. *The Fragmentation of Reason*. Cambridge, Mass.: MIT Press.

Tooby, John, and Cosmides, Leda. 1989. Evolutionary Psychologists Need to Distinguish between the Evolutionary Process, Ancestral Selection Pressures, and Psychological Mechanisms. *Behavioral and Brain Sciences* 12:724–25.

Trivers, Robert. 1978. The Evolution of Reciprocal Altruism. Pages 213–26 in *The Sociobiology Debate*, ed. A. L. Caplan. New York: Harper & Row.

Williams, Bernard. 1972. *Morality: An Introduction to Ethics*. Cambridge: Cambridge University Press.

Williams, Bernard. 1973. *Problems of the Self*. Cambridge: Cambridge University Press.

Williams, George C. 1966. *Adaptation and Natural Selection*. Princeton, N.J.: Princeton University Press.

Wilson, David S. 1992. On the Relationship between Evolutionary and Psychological Definitions of Altruism and Selfishness. *Biology and Philosophy* 7:61–68.

Wilson, David S., and Sober, Elliott. 1994. Reintroducing Group Selection to the Human Behavioral Sciences. *Behavioral and Brain Sciences* 17:585–608.

Wilson, Edward O. 1978. *On Human Nature*. Cambridge, Mass.: Harvard University Press.

Wilson, Edward O., et al. 1973. *Life on Earth*. Stamford, Conn.: Sinauer.

Wilson, Edward O., et al. 1977. *Life: Cells, Organisms, Populations*. Sunderland, Mass.: Sinauer.

Wilson, James Q. 1993. *The Moral Sense*. New York: Free Press.

Wynne-Edwards, V. C. 1962. *Animal Dispersion in Relation to Social Behavior*. Edinburgh: Oliver & Boyd.