THE PHILOSOPHICAL FORUM
Volume XXXIII, No. 4, Winter 2002

EVOLUTION, ALTRUISM AND "INTERNAL REWARD" EXPLANATIONS

JOHN S. BRUNERO

The debate between altruists and egoists concerns the nature of our ultimate desires. Egoists argue that all of our ultimate desires are self-directed. Altruists deny this, claiming that some of our ultimate desires are other-directed and benevolent. In *Unto Others: The Evolution and Psychology of Unselfish Behavior*, Elliot Sober and David Sloan Wilson examine this debate.[1] They argue that social psychology has been unable to prove the altruist's case because an "internal reward" explanation can always be invented to explain apparently altruistic behavior in other terms. Internal rewards are the psychological benefits one receives by performing certain other-regarding actions. Internal rewards include such benefits as the avoidance of guilt, the avoidance of painful memories, and the attainment of warm, fuzzy feelings. Despite the limitations of social psychology, Sober and Wilson believe that evolutionary theory can show that it is more likely for benevolent other-regarding motivational mechanisms to have evolved, thereby supporting the altruist's claim.

Here, I will argue for two related theses. First, if internal reward explanations pose a problem for social psychology, then they also pose a problem for evolutionary theory. Second, there is no need to think that internal reward explanations pose a problem for altruists because these explanations either do not inform us about what our ultimate motives really are or they unreasonably define out of existence the possibility of altruism. The first thesis concerns the implications of internal reward explanations for scientific attempts to tackle the egoism-altruism

Special thanks to Philip Kitcher and Dana Tulodziecki for comments and advice on an earlier version of this paper and to an anonymous referee for this journal for constructive criticism.

[1] Sober, Elliot and David Sloan Wilson. *Unto Others: The Evolution and Psychology of Unselfish Behavior.* (Cambridge, MA: Harvard University Press, 1998). All parenthetical notations in this paper refer to this work.

debate, whereas the second concerns the nature of internal reward explanations themselves.

I

Sober and Wilson state that "the altruism hypothesis says that we have other-directed ultimate desires, whereas psychological egoism says that all our ultimate desires are self-directed" (229). Let us begin by noting how they define "ultimate" and "self-directed" for desires. Sober and Wilson define ultimate desires in contrast to instrumental desires. They define instrumental desires as follows:

> S wants M solely as a means to satisfying S's desire for E if and only if (a) S wants M, (b) S wants E, and (c) S wants M only because S believes that obtaining M will promote obtaining E (217).

Thus, ultimate desires are defined as follows:

> More exactly, S has U as an ultimate desire precisely when S desires U, and, for all the other desires that S has, it is false that S desires U solely because U is a means to satisfying one or more of these other desires. S may recognize that U contributes to the fulfillment of some of these other desires, but this can't be the sole reason that S desires U, if U is ultimate (350).

The "direction" of a desire is defined in terms of the propositional content of the desire (225). Every desire has some proposition that the person who has the desire "wants true" (209). For example, Sam's desiring an apple involves Sam "wanting true" the proposition "Sam eats an apple." If the propositional content of the desire makes reference to the person who has the desire, as in the case of Sam's desire for an apple, the desire is self-directed. If the propositional content of the desire makes reference to another person, the desire is other-directed.[2] We will return to the instrumentality and direction of desires later in this paper.

II

Sober and Wilson devote a chapter of their book to the social psychological experiments and observations that seem to have a bearing upon the debate

[2] However, the "direction" of an ultimate desire alone is not able to determine whether a person's motives are altruistic or egoistic. Consider that an individual may have a *malevolent* self- or other-regarding ultimate desire. An individual with the former does not seem egoistic in the usual sense of the term and an individual with the latter is clearly not an altruist. On this basis, Sober and Wilson argue that the definitions of egoism and altruism need to take into account the goodness of the person's intentions: "Egoists ultimately desire only what they think will be good for themselves; altruists have ultimate desires concerning what they think will be good for others" (230).

between egoists and altruists. They focus primarily on the experiments conducted by Daniel Batson and presented by him in *The Altruism Question*.[3] Batson's studies present data that support the hypothesis that high levels of empathy cause people to have altruistic desires, even when people can choose to act on egoistic desires. Batson considers the "empathy-altruism hypothesis," according to which, if a person has a high level of empathy for another person, she is more likely to be inclined to help that person. This hypothesis is tested along with other "egoistic" hypotheses, such as the "aversive-arousal hypothesis," according to which persons who see others in need of help have "unpleasant experiences" that they wish the avoid (261). Batson's experiments support the conclusion that individuals who have a high level of empathy for others help those people *even if* they have the opportunity to avoid the unpleasant experiences involved in the situation. This conclusion seems to support the altruist's case against the egoist.

However, Sober and Wilson argue that an egoistic explanation can be invented that is *consistent* with this experimental result. For this specific experiment, the egoist will claim that even if high empathy subjects help others when they have the opportunity to avoid "unpleasant experiences," their helping could be explained by other egoistic mechanisms such as the desire to avoid the pain of guilty feelings afterwards (264). Because the question remains open, Batson's experiments do not resolve the dispute between the altruist and the egoist. More generally, Sober and Wilson concede that psychological experiments prove that certain simple or crude forms of egoism are false. But for many of these experiments, Sober and Wilson point out that it "is not difficult to invent an egoistic explanation" for the outcomes (268). These egoistic explanations are the basis for more complex and sophisticated accounts of egoism that are neither confirmed nor disconfirmed by the experimental outcomes. Thus, "[t]he failures of simple forms of egoism don't prove that more complex formulations also must fail" (271).

Sober and Wilson note that "the reason it is difficult to obtain experimental evidence that discriminates between egoism and motivational pluralism [a view that entails that people sometimes have altruistic ultimate motives] is that we have allowed egoism to appeal to *internal* rewards" (273). An egoist can appeal to the individual's desire to feel good about herself, avoid guilt or other unpleasant experiences, or get a warm, fuzzy feeling from helping others. These experimental difficulties only emerge with internal rewards because it is quite easy to test whether one is motivated by an altruistic motive or an *external* reward, such as

[3] Batson, C. D. *The Altruism Question: Toward a Social-Psychological Answer*. (Hillsdale, NJ: Lawrence Erlbaum Associates, 1991).

money, power, or reputation. There are clear cases in which individuals sacrifice external rewards in order to help others. If we define altruism in contrast to external rewards, the altruist can make her case quite easily.

Sober and Wilson argue that there is no good reason to define altruism in contrast to external, but not internal, rewards. If altruism is defined in contrast to the pursuit of external rewards, then "individuals are altruistic when they help just because they think that helping will make them feel good" (230). They appeal to the hypothetical case of a heroin addict whose ultimate desire is for the pleasure of heroin and whose actions are all directed toward this ultimate desire. Imagine that this addict is now placed in an environment where he can get heroin only if he helps people. Intuitively, the addict is not an altruist. They argue that this case is no different from the case of those who help in order to get "internal rewards":

> An addict who helps others only because the effect of helping is a drug-induced euphoria is not thereby an altruist. The same point applies to people in the real world who do not take heroin; if they are "hooked" on helping people because of the pleasure that helping affords and the pain that it allows them to avoid, their actions do not make them altruists. We must not muddy the waters by treating altruism as a form of hedonism (230).

Therefore, there is no philosophical basis for preventing the egoist from referring to internal rewards. Because an internal reward explanation can be formulated for experiments of the kind currently employed by social psychologists, the methods of social psychology are not adequate for the task of settling the dispute between egoists and altruists. We will return to the heroin addict in Section V of this paper.

III

Sober and Wilson argue that evolutionary theory can make a useful contribution to the dispute between egoists and altruists. Specifically, they argue that evolutionary theory supports the conclusion that "no version of egoism is plausible for organisms such as ourselves" (297). Because the psychological evidence, which focuses on behavioral effects, is unable to settle the debate for the reasons presented above, Sober and Wilson "shift the focus from behavioral effects to evolutionary causes . . . Even if two motivational mechanisms both are capable of generating a certain type of behavior, it remains possible that one of them was more likely than the other to have evolved" (298).

Sober and Wilson try to determine whether human parental care for children is based upon an egoistic ultimate motive or an altruistic ultimate motive (or a combination of the two, which would be sufficient to refute the egoist's claim

that all ultimate motives are self-directed). Although only parental care is considered in this context, Sober and Wilson believe that their arguments can be generalized to encompass other kinds of care (314). Sober and Wilson consider two possible motivational mechanisms providing instructions for the provision of parental care. A hedonistic motivational mechanism, labeled "HED," provides the following instructions: "Perform an action if and only if you believe that it will maximize pleasure and minimize pain" (312). An altruistic motivational mechanism, labeled "ALT," provides the following instructions: "Perform an action if and only if you believe that it will do the best job of improving the welfare of your children" (312). They explain their reasons for considering HED and ALT as follows:

> In representing hedonism and altruism in this way, we are using the idea that desires provide *instructions*; they tell you what to do, given the beliefs you have. This is a useful way to represent the functional role that desires play in the regulation of behavior; an organism's corpus of desires constitutes a device that takes beliefs as inputs and yields behavior as outputs. Individuals who follow ALT have an altruistic ultimate motive (312).

The ultimate motive of altruism in ALT is connected to a specific behavioral manifestation; having the motive is necessary and sufficient for the performance of a specific action.

The motivational mechanisms of ALT and HED are assessed according to certain considerations that are relevant to predicting which proximate mechanisms will evolve, the most important of which, in this case, is the reliability of the mechanism. Sober and Wilson argue that ALT is a more reliable mechanism than HED for producing parental care. A parent acting according to HED will have to believe that caring for children yields pleasure and that failing to care for children yields pain. If the act of taking care of children did not yield pleasure, or did not yield as much pleasure as some alternative option, the parent would cease to care for the child (316). HED requires a correlation between the parent's pleasure and pain and the child's well-being. However, because pain, more generally, is not an entirely reliable guide to well-being, HED faces a difficulty in that "it is quite improbable that the psychological pain that hedonism postulates will be *perfectly* correlated with believing that one's children are doing badly. One virtue of ALT is that its reliability does not depend on the strength of such correlations" (316). ALT is a more "direct" mechanism in that it involves a direct concern for the well-being of the child; it is not mediated by a imperfectly reliable calculation of pleasure or pain. On this basis, ALT is a more reliable mechanism for the production of parental care and therefore more likely to have evolved.

IV

I do not believe that the evolutionary strategy outlined above can successfully resolve the debate between egoists and altruists. The evolutionary strategy fails for the same reason that the experiments of social psychologists fail, namely, because an egoistic internal reward explanation can be formulated that is consistent with observations of apparently altruistic behavior.

The mechanism ALT instructs an individual to "perform an action if and only if you believe that it will do the best job of improving the welfare of your children" (312). However, why does this "instruction" need to involve an *ultimate* desire? Consider another motivational mechanism, EGO-ALT, which provides the same instructions as ALT, but exists only because of the internal rewards yielded by the provision of parental care. The individual acting according to EGO-ALT does not have an altruistic ultimate motive. However, natural selection only selects based upon the behavior exhibited by organisms. Because EGO-ALT and ALT do not differ in the behavioral instructions they provide, they are equally as likely to have evolved. Because EGO-ALT remains as an evolutionary possibility, Sober and Wilson's argument cannot use the claim that ALT is more likely to have evolved than HED in order to refute the egoism hypothesis.

The hypothetical EGO-ALT proposal does not deny that there exists a desire (ALT) that instructs a person's behavior. Rather, EGO-ALT postulates a background ultimate desire or set of desires that explains the existence of this instructing desire. The individual who performs an action if and only if he believes it will do the best job of improving the welfare of his children follows this instruction only because doing so allows him to avoid guilt, gain warm, fuzzy feelings, and so forth. The ultimate egoistic desire need not impact the instructing altruistic desire.

The EGO-ALT proposal makes use of the basic line of criticism presented by Sober and Wilson against the conclusiveness of social psychological experiments. Sober and Wilson did not deny that the experimental subjects were "instructed" by their other-regarding desires to perform certain other-regarding and benevolent acts. Rather, they argued that this desire-behavior connection could be explained by a further appeal to an ultimate egoistic desire or set of desires (i.e., Ann wants Bill to do well only because of the internal rewards involved). EGO-ALT makes the same point (although the point is made in the context of considering the evolution of motivational mechanisms rather than the behavior of experimental subjects).

In reply to the EGO-ALT proposal, Sober and Wilson could deny that EGO-ALT and ALT would result in the same behavioral outputs. They could argue that an ultimate desire to help one's children would yield behavior different from a merely instrumental desire to help one's children and do so in a way significant

to natural selection. However, if one can draw such a connection between motivational mechanisms and behavioral outputs, it is unclear why social psychology is incapable of settling the dispute between altruists and egoists. If we say that there is a significant difference in how reliably ALTers and EGO-ALTers care for their children, then, presumably, we could observe, perhaps through some experiment, the behavioral difference and then infer from this the underlying motivations. This task could be accomplished by social psychology.

V

Sober and Wilson argue that because egoists can appeal to internal rewards, the methods of social psychology are not adequately equipped to settle the dispute between egoists and altruists. I have argued that if internal rewards pose a problem for social psychology, they also pose a problem for the evolutionary argument. However, why should internal reward explanations pose a problem at all? Is it really the case that one is not an altruist if one has a desire to perform benevolent actions only because having this desire allows one to feel less guilty, avoid painful memories, or get a warm, fuzzy feeling? Sober and Wilson think so. They argue that the internal rewards case is, in all relevant respects, the same as the case of a heroin addict performing benevolent actions in order to obtain heroin. I believe that there is a significant difference between internal rewards and the rewards sought by the hypothetical heroin addict. This difference is based on two distinct kinds of instrumental relations between desires.

Let us begin by noting that we can distinguish between two ways of speaking of an instrumental relation between desires. We can distinguish between: (1) my desiring of *X as a means* to some Y, and (2) *my desiring of X* being a means to some Y.

In the first case, something is desired as a means to something else. For example, Alice may desire that her father be healthy as a means to her mother's happiness. Here, obviously, there is no instrumental relation between Alice's *desiring* of her father's health and her mother's happiness; the desire does not serve as a means to her mother's happiness. Rather, her father's being healthy serves as a means to her mother's happiness and is desired on this basis.

In the second case, one desires some X, and it is this desiring of X, and not X itself, that serves as a means. For example, Barry may desire his father's health and his desiring of his father's health may serve as a means to a warm, fuzzy feeling. It is clear that it is not his father's health, but his desiring of his father's health, that is an instrument to the warm, fuzzy feeling. If Barry did not desire his father's health, his father's health could not be believed to be an instrument to the warm, fuzzy feeling. This is not so with Alice. If Alice did not desire her father's health, her father's health could still be believed to

be an instrument to her mother's happiness. This marks a central difference between the two kinds of instrumentality. As for preliminary terminology, let us call the first kind of instrumentality "external instrumentality" and the second "internal instrumentality."

The two kinds of instrumentality function quite differently in explanations. For external instrumentality, we explain a person's behavior by pointing to beliefs he has about causal relations in the world and by pointing to his desires. Alice desires that her mother be happy and believes that her mother will be happy if her father remains healthy. For internal instrumentality, in contrast, we do not need to make use of the agent's beliefs about causal relations. Rather, we only need to specify a causal relation between desires. Barry does not have beliefs about the causal relation between his desire and the prospect of a warm, fuzzy feeling—or, at least, the internal reward explanation need not require the existence of such beliefs.

Let us now consider Sober and Wilson's hypothetical heroin addict. They argue that a person receiving internal rewards from his benevolent actions is no different from a heroin addict performing good acts to receive a drug-induced high. However, there is a difference. In the case of the heroin addict, the good acts are externally instrumental to the pleasures of heroin. The heroin addict has the belief that *helping others* is a means to obtain heroin and he desires to help people on this basis. It is not simply that his desire to help others stands in a causal relation to the rewards of heroin. Rather, his motivation is guided by the belief in the causal relation. This is unlike the typical case of internal rewards. For internal rewards, the *desire to help others* stands in an (unrecognized) causal relation to some internal reward. Internal reward explanations track a relation between desires and, therefore, track internal, not external, instrumentality. The heroin addict case, in contrast, is a case of external instrumentality.

In cases of external instrumentality, it is clear that the person *treats* X *as a means* to Y. A selfish banker who gives to others only because he wants tax relief treats other people (or the act of helping other people) as a means to the end of tax relief. The banker believes that helping others is a means to attaining his desired end of tax relief and desires to help people on this basis. However, it is not clear that internal instrumentality involves the agent "treating" her desire for X as a means. In many cases of internal rewards, it would seem odd to formulate the internal instrumental relation as involving the agent treating the altruistic desire as a means. For example, we would not say that Barry, in our example above, *treats* the desire that his father be healthy as a means to avoidance of guilt. For internal rewards, it is not the case that the agent is guided by a belief about some casual connection between his altruistic desire and the resulting internal rewards. Some very odd people might have such beliefs, but internal reward explanations are thought to apply also to those individuals who do not

have such beliefs. The point of internal reward explanations is that the altruistic desire itself simply stands in a causal relation to some other ultimate egoistic desire.

For external instrumentality, an explanation reveals *what* a person's ultimate motives really are. If P treats some X as a means to some Y, this reveals that P's desire for X is not ultimate. But explanations using internal instrumentality do not tell us what our motives really are—they only note causal connection; they do say anything about whether or not the causal connection guides the agent's motivations. First of all, note that some X can be an instrument to some Y without being desired as such. A father's health can be an instrument to a mother's happiness or an instrument to an evil daughter's unhappiness, as she waits for his death and her inheritance. Many facts of the world are causally linked to other facts of the world. What is relevant to the dispute between egoists and altruists, which is a dispute about what our ultimate motives really are, is which of these causal connections are perceived by the agent and guide the agent's motivation; we want to know what the agent *treats* as a means. The same is true of our desires. Many of our desires might stand in some causal relation to one another. The desire to help one's father get well might cause one pleasure or relieve one's guilt. However, what is relevant to the dispute between egoism and altruism is how the desires are *treated*. However, because internal instrumentality does not involve a notion of "treating," it does not select out which causal connections guide the agent. Therefore, it does not tell us what a person's motives are.

The egoist has a reply to this challenge. The egoist will introduce the notion of a second-order desire—a desire about our desires. If we consider second-order desires, we can provide internal instrumentality with a notion of "treating." The egoist will claim that Barry has a second-order desire that his desire to help his father be a means to satisfying his desire to avoid unpleasant feelings. Thus, the desire to help his father is treated as a means.

However, this reply is problematic. Recall that whether a desire is self- or other-directed is determined by the propositional content of the desires. Second-order desires are desires about first-order desires. Therefore, the propositional content of these desires will refer to first-order desires. Because the propositional content refers to one's own psychology, the second-order desire is self-directed. This is true of *all* second-order desires, which are desires *about* desires.

If one treats some X as an ultimate end, there is an open question about whether the X is self-directed or not. If one treats (through a second-order desire) some desire for X as an ultimate end, there is not an open question about whether the desire for X is self-directed or not because the desire for X is a feature of one's own psychology. Given that we are defining altruism in terms of the propositional content of desires, we rule out, *by definition*, the possibility of altruism for second-

order desires. (We even rule out the possibility of altruism for those very admirable individuals who value their altruism and have a second-order desire that their first-order altruistic desires be ultimate!) Thus, we are no longer dealing with a question to which scientific reasoning (in either social psychology or evolutionary theory) can provide an illuminating answer.

In the argument above, I have presented a dilemma for the egoist's appeal to internal reward explanations. Either the explanations are unable to select which causal connections are relevant to an agent's motivation or the explanation will posit a second-order desire, in which case what the agent treats as ultimate will necessarily be self-directed. This dilemma can be illustrated by an example borrowed from C. D. Broad and discussed by Sober and Wilson (261).[4] Consider a physician who travels to some foreign land in order to help the sick and does so knowing that his life over there will be less pleasurable than his life at home. Clearly, helping the sick is not a means to pleasure, whereas staying at home is a means to pleasure. The egoist will claim that even though the physician is not treating the act of helping others as a means to pleasure, he has a desire to help others only because this desire provides him with internal rewards. In making this claim, the egoist is doing one of two things. First, he could be stating that the desire to help others has some causal influence on, or is (internally) instrumental to, some self-directed internal reward. However, simply noting this causal connection says nothing about what the agent's motivations are. Second, he could be stating that the agent has a second-order desire to treat his desire to help others as a means to an internal reward, which is treated as ultimate. However, in treating his desire as ultimate, the propositional content of the second-order desire is such that it must be self-directed. So long as propositional content determines whether an agent's motivations are self- or other-directed, second-order desires rule out the possibility of altruism.

At first glance, the problem of internal reward explanations is an epistemological problem in that social psychology cannot rule out certain egoistic explanations that can be invented to explain apparently altruistic behavior. In the argument above, I have tried to show that egoistic internal reward explanations are themselves problematic. The explanations either do not explain what an agent's motivations are or they present an explanation that must be egoistic so long as we think, as Sober and Wilson do, that the propositional content of desires determines whether or not the desire is self- or other-regarding.

---

[4] See Broad, C. D. *Ethics and the History of Philosophy*. (London: Routledge and Kegan Paul, 1952), pp. 218–231.

VI

The argument presented above can be viewed as a response to Sober and Wilson's objections to Bishop Butler's arguments against hedonism.[5] (Hedonism is a specific version of egoism in that hedonists claim not only that all our motives are self-directed, but that they are also directed at pleasure. However, because Butler's critique of hedonism undermines egoism as well, we will examine it here.) Butler argues that in order to receive any pleasure in performing some act, there must be a "prior suitableness" between the act and one's desires.[6] For example, in order for me to get pleasure from eating food, I must already have a desire for food. Because what I desire is some external thing (food) and not the resulting pleasure, hedonism is false. Sober and Wilson see this line of argument echoed in the work of contemporary philosophers, including Thomas Nagel, who argues that the avoidance of guilt cannot be the basic reason someone performs an action, because the presence of guilt already supposes the prior recognition of some other altruistic reason.[7] Basically, in order for me to get any pleasure or internal reward in helping others, it must be the case that there is already a more basic altruistic desire to help others.

Sober and Wilson argue that this line of argument misrepresents the hedonist position. Even if it is true that the reward comes about because something else is desired, such as food, it still remains open what caused this desire for food. Hedonists argue that the desire for food is caused by the desire for pleasure (278). "Hedonism attempts to *explain* why people want external things; it does not *deny* that they do so" (279). Likewise, for Nagel's analysis of guilt, the crucial question is left open of whether or not the belief that one should help others is caused by the desire to avoid guilty feelings.[8] Egoists claim that people want others to do well only because of a desire for internal rewards.

My argument aims to show that there are two ways of understanding the "only because" part of the egoist thesis as construed by Sober and Wilson. This is where the distinction between internal and external instrumentality is crucial. An external instrumental relation involves a belief about causal relations which, in an explanation, can reveal certain motives as ultimate or not. Internal instrumental relations do not have this feature. (This is why the heroin addict analogy fails.)

---

[5] This objection was originally presented by Sober in "Hedonism and Butler's Stone," *Ethics* 103 (October 1992), pp. 97–103. The discussion of it in *Unto Others* is not substantively different.

[6] Butler's argument on this point is found in *Fifteen Sermons on Human Nature*. Reprinted in L. A. Selby-Bigge (ed.), *The British Moralists: Being Selections from Writers Principally of the Eighteenth Century*, vol. 1 (New York: Dover Books, 1965), pp. 180–241.

[7] Nagel, Thomas. *The Possibility of Altruism*. (Oxford: Oxford University Press, 1970), p. 80.

[8] Sober, "Hedonism and Butler's Stone," *ibid.*, p. 103.

Sober and Wilson fail to note that the central question for the dispute between egoists and altruists is whether these internal rewards guide the motivations of individuals. To answer this question, we need to be concerned with external, not internal, instrumentality. So, in the end, Sober and Wilson's response to Butler does not succeed, and we can conclude that internal reward explanations need not pose a problem for altruists.

Columbia University, New York, NY