



Discussion

Psychological and Evolutionary Evidence for Altruism

ALEJANDRO ROSAS

*Departamento de Filosofía
Universidad Nacional de Colombia
Santa Fé de Bogotá
Colombia
E-mail: rosas@uni-muenster.de*

Abstract. Sober and Wilson have recently claimed that evolutionary theory can do what neither philosophy nor experimental psychology have been able to, namely, “break the deadlock” in the egoism vs. altruism debate with an argument based on the reliability of altruistic motivation. I analyze both their reliability argument and the experimental evidence of social psychology in favor of altruism in terms of the folk-psychological “laws” and inference patterns underlying them, and conclude that they both rely on the same patterns. I expose the confusions that have led Sober and Wilson to defend a reliability argument while rejecting the experimental evidence of social psychology.

Key words: altruism, egoism, evolutionary psychology, folk-psychology, social psychology

Where philosophy and psychology have failed, evolutionary biology can succeed. This is a central claim of Elliot Sober and David S. Wilson’s (1998) recent book *Unto Others*,¹ regarding the resolution of the long ongoing egoism vs. altruism debate. It comes as one more in a series of recent assaults conducted by evolutionary theorizing in the fields of ethics and psychology. Their approach is convincing in many points, but not in regard to this provocative claim. They argue that the evolutionary plausibility of altruistic motivation constitutes particularly strong evidence in its favor, whereas philosophical arguments and psychological evidence have so far been unable to offer anything similar. Traditional treatments in these disciplines fail against a resistant variety of egoism; a version which claims to account for apparent altruism by appealing to sophisticated hedonistic benefits to the self. But are these claims compatible? If hedonistic egoism can successfully account for hard experimental psychological evidence apparently favoring altruism, could it not in principle be capable of accounting in the same way even for evolutionary evidence? Gilbert Harman (2000) has forcefully pursued this line of reasoning.² I here intend to develop a complementary one. Sober and

Wilson's argument to the superior reliability of altruistic motives derives its strength and plausibility from the very same theoretical grounds that lend plausibility to the experimental evidence offered by social psychologists in favor of the empathy-altruism hypothesis. Both defenses of altruism rest on the theoretical framework of folk psychology. No good reasons are given for emphasizing a difference in strength and plausibility between them. It is only, I will contend, on confused grounds that they think themselves entitled to do so.

I

Their evolutionary argument supports the conclusion that evolution favors altruistic motives for triggering parental care, and possibly other helping behaviors. This requires establishing criteria for applying considerations about adaptive value to the mechanisms or motives behind the behavior, and then showing that altruistic motives or mechanisms satisfy these criteria. They plausibly assume that selective pressures can shape adaptive behaviors in given environments and that the same applies to the mechanisms that ensure the occurrence of those behaviors. Roughly, evolution explains a behavior by explaining the mechanism that explains the behavior. Though not all behaviors can be explained in this way, the case for the evolutionary origin of parental care seems well grounded. This allows for the attempt to introduce considerations about the adaptive value of different mechanisms supporting parental care, provided that the relevant criteria for mechanism-fitness can be found.

1. *A criterion for mechanism-fitness: Reliability*

In our context, the most important criterion for mechanism-fitness concerns *reliability*: A mechanism M is reliable when it triggers a behavior B in circumstances in which B preserves or brings about a state of affairs S, and S is well correlated with the fitness of the organism possessing M. According to folk psychology, organisms with minds generate their behavior through a combination of desires and beliefs. Desires have a propositional content that represents the states of affairs that given behaviors must preserve or bring about. Desires are ultimate when this content cannot be derived from the content of other desires via beliefs. Otherwise, they are instrumental. Beliefs give information about states of affairs and about behaviors that preserve or bring about states of affairs. Organisms normally feel pleasure at the satisfaction of desire, or pain at its frustration, and can detect these internal states of pleasure and pain. In the context of deciding between hedonistic egoism

and altruism, only a comparative reliability question needs to be answered: whether an organism would control its behavior more reliably by having ultimate desires aiming at pleasure/pain or, alternatively, at external features of the environment. It could be better off with ultimate desires aiming at states of affairs other than pleasure (or avoidance of pain) and triggering adaptive behaviors in their presence (UO: 311). Since pleasure and pain are not objective states of affairs correlated with fitness, but internal indicators for objective situations that increase or diminish fitness, the question to ask will then be: How well are states of pleasure or pain correlated with those objective states of affairs that are, in turn, correlated with fitness, and about which an organism can know, and aim at, independently of pleasure and pain?

Sober and Wilson proceed in fact to apply this reliability criterion to two rival mechanisms that could do the work of controlling parental care. ALT, the altruistic mechanism, is directly keyed to the welfare of one's children. It triggers helping behavior at the belief that one's children need help. Its reliability depends only on the accuracy with which it detects how well/badly the children are doing. HED, the egoist-hedonist mechanism, is not directly keyed to the welfare of one's children, but to internal states of pleasure and pain. There are different ways in which pleasure or pain can be connected to states of affairs that increase or diminish fitness. At one extreme, one has the system of bodily pain as indicator of bodily injury. Pain is here immediately connected with a state of affairs that diminishes fitness, that is, without the mediation of belief. But this is, perhaps, an exception. Unlike the injuries in one's body, the states of welfare of one's children do not *directly* produce parent's pleasure or pain. In humans, the usual causal channel is through beliefs about how the children are doing (UO: 313–314). This means that HED requires, as ALT does, beliefs about the welfare of one's children. The comparative reliability of these mechanisms will depend, then, on two factors that play a role in HED but not ALT: on the accuracy with which it detects its internal states of pleasure/pain, and on the degree of correlation between these states and the beliefs about the welfare of one's children. Assuming that HED detects pleasure/pain with perfect accuracy, the reliability judgement will depend on what can be said about the degree of correlation between the pleasure/pain of the parent and the state of welfare of the children. If there is reason to believe that the correlation is less than perfect, ALT would be more reliable than HED and natural selection would have favored its evolution.

2. *Reliability, the subordination thesis and folk psychology*

One of the reasons they give for this conjecture is drawn from a comparison with the reliability of pain as regulator of adaptive behavior concerning bodily preservation, which depends on the degree of correlation between pain and

bodily injury. The correlation is less than perfect, since some injuries do not produce pain, and some pains occur when bodily injury is no longer present. In the light of this they conclude: “it is improbable that psychological pain . . . will be *perfectly* correlated with the belief that one’s children are doing badly” (UO: 316). This conclusion is drawn too quickly,³ so I move on to a different ground for thinking that parental pleasure must be less than perfectly correlated with belief in the welfare of one’s children. Sober and Wilson point out that perfect correlation would require “not just that some pleasure accompany [the belief that one’s children are doing well]. The amount of pleasure that comes from seeing one’s children do well must exceed the amount that comes from [other competing sources]” (UO: 315). I think they point here in the right direction. The relevance of this observation is properly appreciated in the context of an idea that fixes the content of the egoistic hypothesis. A pretty obvious – but not trivial – idea implicit in philosophical treatments of the altruism vs. egoism debate is that egoism is committed to the thesis that the welfare of others is treated as a means to some sophisticated form of hedonistic self-benefit. Let us call this thesis about egoism the subordination thesis. Accordingly, if a parent employs HED as a mechanism to control parental care, it will be ultimately seeking a form of pleasure, to which it will subordinate the welfare of its children, as a mere means. A parent using HED may fail to feel pain and trigger parental care when the belief is present that the children are needy, simply because there are alternative sources of pleasure available and helping its children is a bore. The fact that the children’s welfare is treated by HED as a means to pleasure is the reason why HED implies a less than perfect correlation between parental pleasure or pain and parental beliefs about the welfare of its children. This, in turn, is the reason why HED is less reliable than ALT at triggering parental care when the children are needy.

This line of reasoning relies on two premises to prove the inferior reliability of HED. The first characterizes the egoistic-hedonistic mechanism with the subordination thesis: this mechanism treats the welfare of needy children as a means to pleasure and uses, therefore, an instrumental desire. The second premise formulates a plausible psychological principle concerning the comparative reliability of instrumental vs. ultimate desires. In general, desires with a propositional content *S* generate behavior *B* when the belief is present that *B* preserves or brings about *S*. Call this belief B_S . The principle then says that an instrumental desire with propositional content *S* is *less* reliable in triggering *B* in the presence of B_S than an ultimate desire with content *S*. The plausibility of this principle is not difficult to see. It follows from the difference between instrumental and ultimate desires. Of an ultimate desire it is true that the generation of *B* follows, in general, upon the presence

of B_S . But this is not true of an instrumental desire. Since instrumental desires are conditional upon ultimate desires, the generation of B will require an additional belief: the belief that S preserves or brings about a state of affairs P, which would be the propositional content of the ultimate desire supporting the instrumental one (in our case, roughly, that I feel pleasure). This belief may be absent when other competing sources of pleasure are available, in which case the behavior in question will not be triggered. It follows that a mechanism that aims at S through behavior B, will more reliably trigger B in the presence of B_S by using an ultimate desire than by using an instrumental desire.

So it turns out that the inferior reliability of HED compared to ALT can be proved from the subordination thesis in the context of philosophical psychology. The philosophical psychology presupposed is of the folk type, because it uses principles assuming nomological or quasi-nomological relations between beliefs, desires and behavior. The reliability argument assumes a psychological principle stating that a behavior B promoting a state of affairs S is more reliably produced by an ultimate desire with propositional content S than by an instrumental desire with the same propositional content. It assumes, therefore, that inferences can be drawn from desires to behavior and vice versa, and this argument is only intelligible in the light of such folk-psychological assumptions. Given this assumption, one cannot consistently have any principled doubts about the accessibility of psychological motives, desires and/or beliefs on grounds of behavioral evidence. One should note, besides, that the reliability of altruistic motives does not require their dissociation from pleasure. ALT treats the children's welfare as an ultimate goal, and this is not incompatible with the fact that satisfying ALT gives pleasure to the organism. The presence of this pleasure does not turn ALT automatically into a hedonistic mechanism. This would ignore the true empirical content of the egoism/hedonism hypothesis. This content is given by the subordination thesis, which in our case implies that parental care or concern for the welfare of one's children is subordinated to attaining parental pleasure and consequently, that the decision whether to care for one's children is taken only on the basis of comparative judgements about the amount of pleasure that such a behavior produces.

II

3. *Egoism and altruism as empirical hypotheses*

It is important not to lose out of sight the empirical import of the egoism vs. altruism debate. Egoism has been defended with confused semantic

manoeuvres that try to show the impossibility of altruistic desires. Some already belong to those commonplace confusions that have been exposed countless times, some are more rare and sophisticated.⁴ It would be a mistake, however, to give the exposition of these confusions the status of an argument proving the falsity of egoism/hedonism.⁵ Though this “refutation” strategy is popular among philosophers and has become part of the “folklore of our tribe”,⁶ it ignores the empirical content of the debate. For the following discussion, I will adopt the following conventions: The terms “egoistic” and “altruistic” are applicable to desires and possibly to other propositional attitudes. A motive will be said to be altruistic or egoistic depending on whether the desire that individuates it is altruistic or egoistic. I will say that a desire is egoistic when it is ultimate and its content is given by a proposition that describes a benefit for the subject of the desire, or when it is instrumental and can be ultimately derived from other desire with that content. I will use “E” to refer to this type of content, and to the desires it defines. A desire will be altruistic when it is ultimate and its content is given by a proposition that describes a benefit for someone other than the subject of the desire, or when it is instrumental and can be ultimately derived from a desire with such content. I will use “A” to refer to this type of content, and to the desires it defines. I will further say that a desire is apparently altruistic, when its content is given by a type-A proposition, while it is still open whether the desire depends on some other desire, which could turn out to be egoistic. Egoism is the hypothesis that holds that all our desires are egoistic. Altruism claims that some of our desires are altruistic. Since there is no doubt, even for partisans of egoism, that some of our desires are apparently altruistic, that is, that we sometimes have a desire – or some other sort of propositional attitude (e.g. hope) – whose propositional content describes some benefit to some other person, the debated question can be given a more precise formulation. It concerns whether our apparently altruistic desires are in fact all derivable from desires that are egoistic. Egoism gives an affirmative answer to this question. It is committed to what I above called the subordination thesis: it claims, in the less formal terms I used above, that when we aim at a benefit for others, it is only as a means to obtaining some benefit for ourselves. The empirical consequence of this, according to folk psychology, is simply that we will observe the apparently altruistic desires and the corresponding behaviors disappear, when the subjects (which are presumably egoistic according to this hypothesis) believe that the egoistic benefit will not obtain. This predictive claim and its negation I will consider to be the truly empirical and testable contents of the egoistic and altruistic hypotheses respectively.⁷

4. *Psychological evidence: The empathy-altruism hypothesis*

The social psychologist Daniel Batson and his colleagues have undertaken an attempt to decide the issue by means of a series of experiments designed to test the “empathy-altruism hypothesis”.⁸ This is the hypothesis that a social emotion, usually called “empathy”, evokes altruistic motivation. Uncontroversially, this feeling evokes *apparently* altruistic desires, since it is usually linked to behaviors showing concern for some needy person, e.g. in the form of helping behavior. What is controversial and in need of testing is whether the apparently altruistic desires evoked by empathy are dependent on ultimate egoistic desires or not. Partisans of egoism claim that they are so dependent. There is, they claim, not just one, but a series of possible egoistic desires that are empathy-specific, that is, that are especially evoked when subjects feel empathy, and that can therefore adequately explain the generally observed link between empathy and the desire to help.

Given that one needs evidence to decide whether or not the apparently altruistic desires accompanying empathic feelings depend on ultimately egoistic desires, Batson and his colleagues have proceeded to design experiments that claim to have access to the content of the desires operative in subjects that are manipulated to feel empathy for a needy person and who are confronted with the decision whether to help or not. Observing their behavioral response in experimental situations where control over independent variables allows knowledge of their relevant psychological states, they claim to be able to infer the nature of the underlying motives. This procedure shows that the experiments are embedded in a folk psychological framework, which assumes that inferences to unknown psychological states can be drawn from a conjunction of observed behavior and other known psychological states. Certainly, they do not argue in support of this framework, they just assume it. But they do claim that their experiments are so designed that, given this framework, they can provide reliable knowledge of the nature of some psychological states of the experimental subjects, like their beliefs and their feelings, and that on this basis they can access the content of their desires. They claim, for example, that their experimental settings allow them to know that some of their subjects feel empathy, while others do not.

I will not discuss any of their techniques for manipulating empathy or its absence, since Sober and Wilson’s objections do not turn on this point. A further claim is that the design of their experiments will place experimental subjects in different belief-states about the causal connection between their behavior (in most cases, helping or not helping a needy subject) and a state of affairs that consists in a benefit either to themselves or to some needy subject or to both or to neither, and that they can know which subjects are in which states. One of Sober and Wilson’s lines of objection seems to turn on this

point, so it will be addressed again below. But we need, first of all and before discussing the objections, to have a clear grasp on how the experiments, in ideal conditions, claim to provide the evidence required to solve the question of whether empathy evokes altruistic motives or not. For the sake of clarifying the logic of these experiments, let us assume that techniques exist to successfully manipulate empathy level, and that experimental situations can be carefully designed to reveal the beliefs of the experimental subjects. Then these experiments place empathically predisposed subjects in one or the other of the following conjunctions of beliefs about the casual connection between their helping behavior and states of affairs of type A or E:

1. helping causes states A and E; not-helping causes not A nor E (E and A *only* by helping);
2. helping causes states A and E; not-helping causes not A but E (E *also* by not-helping);
3. helping causes state A but not E; not-helping causes not A but E (E *only* by not-helping); where “not-helping” means either abstention from helping or doing something other than helping.

Given that one has this sort of knowledge about the beliefs held by the experimental subjects, the claim is then that one can infer the content of their desires on the basis of their observed behavior. The pattern of inference that allows this is, basically, that if empathically predisposed subjects display a high proportion of helping behavior when they have beliefs corresponding to cases 2. – A *only* by helping and E *also* by not helping – and 3. – A *only* by helping and E *only* by not-helping –, then egoism is disconfirmed, since folk psychology allows one to assume that subjects with these beliefs help only if they are aiming at a state of affairs of type A. More formally, the conflicting patterns of inference that would permit to resolve the nature of motivation would be the following:

- E1: If x feels empathy for y, and x believes that *only* helping y causes A, and x believes either that *also* not-helping y causes E or that *only* not-helping y causes E, and x does not help y, then x has an E-desire.
- A1: If x feels empathy for y, and x believes that *only* helping y causes A, and believes either that *also* not-helping y causes E or that *only* not-helping y causes E, and x helps y, then x has an A-desire.

An example may be useful. One of the experiments is designed to record the behavior of empathically predisposed subjects when they are confronted with a person whom they believe is about to suffer a series of ten electric shocks. After watching two shocks, the subjects are given the opportunity to help by volunteering to take the place of the suffering subject. But subjects receive different experimental treatments: some of them are offered the possibility of walking out of the experiment, that is, of ceasing to watch if they

don't want to help (they are given the easy-escape treatment), while others have antecedently agreed to watch all ten shocks (the difficult-escape treatment). This is an independent variable, designed to place the subjects in predictable states of belief under experimental control. All subjects, namely, will believe that only helping will bring about a state of affairs of type A (relief for the suffering subject). But subjects in the easy-escape treatment will believe that they can avoid the aversive feelings arising from the sight of someone suffer (bring about a type E state of affairs) also by not-helping – walking out – while subjects in the difficult-escape situation will believe that only helping will free them from those aversive feelings. If it is true that empathically predisposed subjects usually help to avoid aversive arousal (a desire of type E) then one should observe a higher proportion of helping behavior when they believe that only helping will bring about E, than when they believe that they can also bring about E by declining to help. Alternatively, if they help because of a desire of type A, then the proportion of helping will be high in both cases, that is, it will not be relevantly affected by whether subjects can avoid watching the suffering subject or not. The patterns of inference A1 and E1 above are intended to express this. They are exactly the patterns of inference that Batson and his colleagues have employed to conclude, on the basis of the behavioral evidence experimentally obtained, that empathic subjects have altruistic desires. Based on these inferential patterns, if one knows the behavior and the beliefs of the subjects, one can infer the content of their desires. The interesting point for our purposes is that there is a straightforward relationship between these inferential equations and the psychological principle behind Sober and Wilson's argument to the higher reliability of ALT. These equations explicate the difference in the frequency of helping behavior as a function of the difference in the propositional contents of the desires motivating the behavior. If the desire motivating the helping behavior is of type E, the behavior will have a low frequency across different contexts; if it is of type A, the behavior will have a high frequency across different contexts. Given the soundness of those inferential equations, it is then possible to state a principle stating that a desire of type A will more reliably produce, across different contexts, the helping behavior specific to parental care. This is the principle used by Sober and Wilson in their reliability argument, and it rests on the soundness of these equations.

5. *Confusions*

Sober and Wilson see in the psychological evidence provided by the experiments described above a weakness that results from a limitation in their method. They grant that new methods in experimental psychology could succeed, but they believe that this one has failed (UO: 272). What accounts

for its failure? They do not unequivocally explain this. But there is one line of thought that Sober and Wilson would probably be satisfied to see it highlighted here as their considered objection. They point out that egoism, being essentially an abstract hypothesis defining a research program, has a flexibility that allows its partisans to conceive of increasingly subtler versions and present them as replacing those that have been experimentally refuted (271; 288ff). The experimental treatment of the debate by the social psychologists can only, however, refute the versions of egoism one by one. This line of thought can be given a formal expression using the pattern of inference A1 above. This pattern was, we recall:

A1: If x feels empathy for y , and x believes that *only* helping y causes A , and believes either that *also* not-helping y causes E , or that *only* not-helping y causes E , and x helps y , then x has an A -desire.

The precision needed here is that the conclusion can only state that the A -desire is only apparently altruistic, or altruistic relative to a specific E -desire, in so far the experimental results show that the A -content is not dependent on a given particular E -content. It would be altruistic if this particular E -content were the only egoistic content upon which the A -content could depend. But there could be, for all this inference shows, a desire with propositional content E_i such that x would not help y if he believes either that *also* not-helping causes E_i or that *only* not-helping causes E_i , where x believes in both cases that *only* helping causes A . Concretely, if empathically predisposed subjects volunteer to help a person whom they believe is about to suffer a series of electric shocks even when they can walk out and cease watching after two shocks, then the egoistic desire to avoid the aversive reaction produced by the *sight* of the suffering person does not motivate them. This does not automatically mean that they are altruistically motivated. They could be motivated by the desire to avoid the aversive feelings produced by the *knowledge* that the subject will continue to receive electric shocks even if they walk out, or by the desire to avoid the aversive feelings arising from self-censure for failing to help or by any other desire aiming at a plausible self-benefit in this context.

There is nothing to say against this line of argument, except that it does not identify any particular methodological limitation. It follows from the fact that egoism is not a hypothesis about a particular motive behind helping behavior, but about a type of explanation. It is, as Sober and Wilson say, a research program (UO: 289), and it is kept alive as long as plausible new particular hypotheses can be formulated to replace those old particular hypotheses that have undergone empirical or experimental disconfirmation. No particular experiment refuting a particular hypothesis falling under a research program can invalidate the whole program. Scientists can, however, convince themselves of abandoning the program if they encounter repeated experimental

disconfirmation of the successive particular hypotheses offered. I do not see how else a research program could be discarded on the basis of empirical evidence and I believe that Sober and Wilson are not fair if they are expecting an experimental methodology to be able to do more than this.

I believe, therefore, that Sober and Wilson have not made a sound case for the thesis that the experimental methodology used by social psychologists suffers under any special weakness, given that they accept, as they must do, the folk-psychological framework supporting it. But even if they should drop the thesis that there is a weakness in their methodology, and even if they did acknowledge that this methodology is the only empirical avenue open for the refutation of egoism, they can still object that social psychologists have overlooked some possible egoistic motives behind helping behavior. This, note, would not be any objection against the methodology, but rather an invitation to carry out experiments that have not yet been carried out. Several reasons could account for this omission. Psychologists may simply have not thought of some plausible egoistic benefit, or they may have false beliefs about the real nature of some egoistic benefit, which in turn would have led them to use an inappropriate fixed variable in relation to it in their experiments, one that does not in fact place the subjects in any belief-state about it. This invalidates their results, because in their experimental strategy a false belief attribution precludes access to the content of the desires of the subjects under consideration. Sober and Wilson have ventured two such new hypotheses about egoistic benefits that could be envisaged by people who help others (UO: 268, 271).⁹ One of them is particularly interesting in this context, and I want to comment on it before finishing. I want to argue that, though it looks as if Sober and Wilson were calling attention to an egoistic hypothesis not yet tested, or tested with an inappropriate fixed variable, the nature of the egoistic benefit they are considering is suspicious. Moreover, I will venture the thesis that it is not an egoistic benefit at all.

One of the several egoistic hypotheses offered (the negative-state relief hypothesis) says that when individuals feel empathy for someone they witness suffering, they fall into a negative affective state (sadness) and they help to find relief from this state. Batson conducted experiments that have shown that when subjects have been led to believe that they will be compensated by a mood-enhancing experience even if they don't help (*E also* by not-helping), they still help a needy person. But this gives no evidence against egoism, say Sober and Wilson, because the hypothesis under consideration does not allow that just any state of affairs *E* of a mood-enhancing type will lift one out of the specific sadness one is experiencing. Insofar as this is true, subjects in those experiments will simply have no reason to believe that the promised mood-enhancing experience will give them the egoistic benefit that

the hypothesis postulates is driving them to help. So it seems that the social psychologists have used an inappropriate fixed variable, probably because they have not correctly identified the nature of the benefit to the self that this egoistic hypothesis is envisaging. Given their mistake, their experiments cannot show that the desire producing the helping behavior of the subjects is not a desire for an egoistic benefit, correctly interpreted. So the question turns, necessarily, on the precise nature of the egoistic benefit that presumably drives our helping behavior according to this hypothesis.

If I am right in saying that this objection is not an objection of principle, directed against the methodology itself, but against an inappropriate application of it, then it is also correct to say that it is an invitation to design an appropriate experiment, using appropriate fixed variables. This may be difficult or even impossible. It is not inconceivable that an egoistic benefit exists, and drives helping behavior, yet it is impossible, for contingent reasons, to make subjects believe that they can only obtain it by not helping someone or can also obtain it by not helping her. It will be contingently impossible for subjects to have this belief if they believe themselves to live in a world that is so arranged that the crucial egoistic benefit producing helping behavior is contingently but inextricably bound to a benefit for others. In such a world, you will believe that you will receive this benefit only if you help others. Given this belief, no experiment of the type described will be able to access egoistic and altruistic desires in isolation from each other on the basis of helping behavior and belief attribution.

But in the case of the egoistic hypothesis proposed by Sober and Wilson, the isolation of those motives from each other is, I contend, not just contingently, but logically impossible. They stress the fact that sadness is usually sadness about something in particular. Agreed, and I would add that feelings like sadness can be viewed as having propositional objects. This makes it easier to inquire into the desires triggered by or associated to those feelings. For example, if one really feels sad *that* someone else is suffering, presumably one would have the desire *that* this person be better off. One here really feels sad for the other person. But if this desire depends on some other, ultimate desire, such as the desire to avoid an unpleasant sight, then one does not really feel sad *that* someone else is suffering. Rather, one feels disturbed *that* one has to endure the unpleasant stimulus of someone suffering, or disturbed on one's behalf. In the first case we have a proposition describing a benefit to a person other than the subject of the feeling or desire, and provided it does not depend on further desires, nor ultimately on an egoistic one, we have a case of genuine altruism according to our definition. In this case, it would not be surprising, as Sober and Wilson note, that the pain we experience "is not completely assuaged by any old pleasant experience" (UO: 271). But note

that it is not the mere fact that sadness is about something in particular, or can be viewed as having a propositional object, that makes for the unsurprising results of the experiments. What makes them unsurprising is that, under Sober and Wilson's "egoistic" hypothesis, the specific propositional object of the sadness in question is the same state of affairs that would be the object of an ultimate altruistic concern, namely the welfare, or lack of it, of some other person.

According to this interpretation, what Sober and Wilson are postulating is an egoistic desire for relief from the specific sadness *that* someone else is suffering, where this relief requires, *necessarily*, an improvement in the situation of the person who is the object of empathy.¹⁰ This is equivalent to defending psychological egoism by claiming that our concern for somebody's welfare is best expressed by saying that this person's welfare gives us pleasure, and that it is therefore our own pleasure that we want. But when we say that somebody's welfare gives us pleasure, we do not mean that the welfare and the pleasure are two things in the way in which good deeds and the good reputation they bring about are two things, such that we would probably like to get the second without investing in the first – if I may modify an example by Ryle.¹¹ Those are rather cases where feeling pleasure at something is an expression of valuing it *for its own sake*. If this state of mind, call it "valuing", "pleasure" or whatever, turns altruism into hedonism, then it will certainly be impossible to find any evidence for altruism based on the inferences A1 and E1, including, of course, evidence based on reliability arguments.

Acknowledgements

For useful discussion and suggestions I thank Dan Batson, Elliot Sober and Kim Sterelny. Thanks are also due to the *Alexander von Humboldt Foundation* for financing my research stay in Germany and to Ludwig Siep for supervising and supporting my project.

Notes

¹ Esp. pp. 295, 297f. Referred to as UO.

² Harman has pointed out that some hedonistic explanations suggested by Sober and Wilson presuppose an understanding of hedonism against which it would be impossible to find empirical evidence. I would add, however, that such an understanding erases the distinction between hedonism and altruism. Since Sober and Wilson claim that the difference between them is real and empirical, I do not see these suggestions as their considered opinion, but rather as a *lapsus*.

³ It is here important to remember that bodily pain indicates a situation diminishing fitness (bodily injury) in an exceptionally direct form. To judge the mechanism of bodily pain, the relevant fact is that our most reliable beliefs in bodily injury are based on pain sensations, and that we do not feel pain on the basis of a previous belief that we are injured. The syndrome known as *congenital absence of pain* shows how unreliable our knowledge of bodily injury is without pain. In the case of HED as a mechanism for producing parental care, the situation is significantly different. Pleasure and pain would not be direct indicators of the welfare of one's children, but would indicate it through beliefs about how they are doing. Their reliability as indicators would heavily depend on the accuracy of our beliefs, which is not the case of pain as indicator of the injuries of one's body. It is therefore possible to say that the parent's pleasure is less than perfectly correlated with the children's welfare, precisely because beliefs about the children's welfare are less than perfectly accurate, and not because of a less than perfect correlation between parent's pleasure and beliefs in children's welfare. No clear conclusion about this second correlation can be drawn from a comparison with the reliability of the mechanism of bodily pain, and therefore no support to the contention that ALT will be more reliable than HED.

⁴ Against two such sophisticated manoeuvres cf. e.g. Bernard Williams (1973), pp. 250–265, esp. 261–262.

⁵ A strategy used by Butler and by some of his commentators. Cf. Butler (1729), p. 228; Sigdwick (1907), pp. 43–44; Broad (1930), pp. 186–191; Duncan-Jones (1952), pp. 95–97, 99–100.

⁶ Cf. C. Morillo, esp. p. 169f.

⁷ Carolyn Morillo has also argued that this is the empirical content of the issue between altruism and egoism: if a given behavior has a hedonistic motivation, it must decrease or be absent when not leading to the hedonistic reward. Its performance in absence of the reward would disconfirm hedonism. Morillo speculates about a hypothetical neurological test for an egoistic/hedonistic theory of motivation. It would be based on the possibility of identifying the hedonistic reward as a “brain event”, and of controlling – suppressing or inducing – it directly in experiments conducted to observe the effects of such control on behavior. This takes us beyond a pure folk-psychological framework in the altruism vs. egoism debate, but as far as I know no such experiments have yet been conducted. Cf. C. Morillo (1990), *passim*, esp. p. 184.

⁸ Cf. D. Batson and Laura Shaw (1991), 107–122.

⁹ Cf. the commentary by Batson (2000).

¹⁰ They do not put it exactly in this way in *Unto Others* (cf. p. 271), but that is how Sober has expressed it to me in a personal communication. Note that such a desire would imply a perfect correlation between an E-content and an A-content and, *simultaneously, the impossibility of a reliability argument for altruism*, which depends, as explained above, on an imperfect correlation between them.

¹¹ Cf. Gilbert Ryle (1949), p. 132.

References

- Batson, D.: 2000, ‘Commentary’, *Journal of Consciousness Studies* 7(1–2), 207–210.
 Batson, D. and Shaw, L.: 1991, ‘Evidence for Altruism: Towards a Pluralism of Prosocial Motives’, *Psychological Inquiry* 2(2), 107–122.
 Broad, C.D.: 1930, *Five Types of Ethical Theory*, Kegan Paul, London.
 Butler, J.: 1729, ‘Sermons’, in L.A. Selby-Bigge (ed.), *British Moralists*, Dover, New York.

- Duncan-Jones, A.: 1952, *Butler's Moral Philosophy*, Penguin Books, Harmondsworth.
- Harman G.: 2000, 'Can Evolutionary Theory Provide Evidence against Psychological Hedonism', *Journal of Consciousness Studies* 7(1–2), 219–221.
- Morillo, C.: 1990, 'The Reward Event and Motivation', *The Journal of Philosophy* 87(4), 169–186.
- Ryle, Gilbert: 1949, *The Concept of Mind*, Hutchinson, London.
- Sidgwick, H.: 1907, *the Methods of Ethics*, 7th edn., Hackett, Indianapolis/Cambridge.
- Sober, E. and Wilson, D.S.: 1998, *Unto Others. The Evolution and Psychology of Unselfish Behavior*, Harvard University Press, Cambridge, MA.
- Williams, Bernard: 1973, 'Egoism and Altruism', in *Problems of the Self*, C.U.P., Cambridge, pp. 250–265.

