Biology and Philosophy (2006)
DOI 10.1007/s10539-006-9030-1

Evolution, altruism and cognitive architecture: a critique of Sober and Wilson's argument for psychological altruism

STEPHEN STICH
Department of Philosophy & Center for Cognitive Science, Rutgers University, 26 Nichol Avenue, New Brunswick, NJ 08901-2882, USA
(e-mail: stich@ruccs.rutgers.edu; phone: +732-932-9861; fax: +732-932-8617)

Received 3 August 2005; accepted in revised form 22 March 2006

Key words: Altruism, Cognitive architecture, Egoism, Evolution, Intrinsic and instrumental desire, Natural selection, Sub-doxastic states

Abstract. Sober and Wilson have propose a cluster of arguments for the conclusion that "natural selection is unlikely to have given us purely egoistic motives" and thus that psychological altruism is true. I maintain that none of these arguments is convincing. However, the most powerful of their arguments raises deep issues about what egoists and altruists are claiming and about the assumptions they make concerning the cognitive architecture underlying human motivation.

In their important book, *Unto Others*: *The Evolution and Psychology of Unselfish Behavior*, Elliott Sober and David Sloan Wilson offer a new and interesting evolutionary argument aimed at showing that in the venerable dispute between psychological altruism and psychological egoism, altruism is the likely winner. In this paper, I'll argue that Sober and Wilson's argument relies on an implicit assumption about the cognitive architecture subserving human action, that much recent work in cognitive science suggests the assumption may be mistaken, and that without the assumption, their argument is no longer persuasive. Before getting to any of that, however, we'll need to fill in a fair amount of background.

Preliminaries

Far too many discussions of evolution and altruism founder because they fail to draw a clear distinction between two very different notions of altruism which, following Sober and Wilson, I'll call *evolutionary altruism* and *psychological altruism*. One of the many virtues of Sober and Wilson's book is that they draw this distinction with exemplary clarity, and never lose sight of it.

A behavior is *evolutionarily altruistic* if and only if it decreases the inclusive fitness of the organism exhibiting the behavior and increases the inclusive fitness of some other organism. Roughly speaking, inclusive fitness is a measure of

how many copies of an organism's genes will exist in subsequent generations.[1] Since an organism's close kin share many of its genes, an organism can increase its inclusive fitness either by reproducing or by helping close kin to reproduce. Thus many behaviors that help kin to reproduce are *not* evolutionarily altruistic, even if they are quite costly to the organism doing the helping.[2]

Evolutionary altruism poses a major puzzle for evolutionary theorists, since if an organism's evolutionarily altruistic behavior is heritable, we might expect that natural selection would replace the genes that influence the behavior with genes that did not foster altruistic behavior, and thus the altruistic behavior would disappear. In recent years, there has been a great deal of discussion of this problem. Some theorists, Sober and Wilson prominent among them, have offered sophisticated models purporting to show how, under appropriate circumstances, evolutionary altruism could indeed evolve, while others have maintained that the evolution of altruism is extremely unlikely, and that under closer examination all putative examples of altruistic behavior will turn out not to be altruistic at all. In the memorable words of biologist Michael Ghiselin (1974, 247) "Scratch an 'altruist' and watch a 'hypocrite' bleed." Since my focus, in this paper, is on psychological altruism, I'll take no stand in the controversy over the existence of evolutionary altruism.

A behavior is *psychologically altruistic* if and only if it is motivated by an ultimate desire for the well-being of some other organism, and as a first pass, we can say that a desire is *ultimate* if its object is desired for its own sake, rather than because the agent thinks that satisfying the desire will lead to the satisfaction of some other desire. Though I'll need to say more about ultimate desires and psychological altruism, what's already been said is enough to make the point that evolutionary altruism and psychological altruism are logically independent notions – neither one entails the other. It is logically possible for an organism to be evolutionarily altruistic even though it has no mind at all and thus can't have any ultimate desires. Indeed, since biologists interested in evolutionary altruism use the term 'behavior' very broadly, it is possible for paramecia, or even plants, to exhibit evolutionarily altruistic behavior. It is also logically possible for an organism to be a psychological altruist without being an evolutionary altruist. For example, an organism might have an ultimate desire for the welfare of its own offspring. Behaviors resulting from that desire will be psychologically altruistic though not evolutionarily altruistic, since typically such behaviors will increase the inclusive fitness of the parent.

---

[1] Giving a more precise account would raise some of the deepest issues in the philosophy of biology. (See, for example, Beatty 1992). Fortunately, for our purposes no more precise account will be needed.

[2] Some writers, including Sober & Wilson, define evolutionary altruism in terms of *individual* fitness rather than inclusive fitness. I prefer the inclusive fitness account since, as we'll soon see, it makes it easier to understand how Sober & Wilson's wise decision to focus on parental care sidesteps the debate over the existence of evolutionary altruism.

I've said that to be psychologically altruistic, a behavior must be motivated by an ultimate desire for the well-being of others. That formulation invites questions about what it is for a behavior to be *motivated* by an ultimate desire and about which desires are *for the well-being of others*. The second question, though it certainly needs to be considered in any full dress discussion of psychological altruism, can be put-off to the side here, since a rough and ready intuitive understanding of the notion is all I'll need to explain Sober and Wilson's argument and my concerns about it.[3] One interpretation of the traditional notion of *practical reasoning* provides a useful tool for explaining the relevant sense of a behavior being motivated by an ultimate desire. On this account, practical reasoning is a causal process via which a desire and a belief give rise to or sustain another desire. That second desire can then join forces with another belief to generate a third desire. And so on. Sometimes this process will lead to a desire to perform what Goldman calls a "basic" action, and that, in turn, will cause the agent to perform the basic action without the intervention of any further desires.[4] Desires produced by this process of practical reasoning are *instrumental* desires – the agent has them because she thinks that satisfying them will lead to something else that she desires. But not all desires can be instrumental desires. If we are to avoid circularity or an infinite regress there must be some desires that are *not* produced because the agent thinks that satisfying them will facilitate satisfying some other desire. These desires that are not produced or sustained by practical reasoning are the agent's ultimate desires. A behavior is *motivated* by a specific ultimate desire when that desire is part of the practical reasoning process that leads to the behavior. Figure 1 depicts some of these ideas in a format that will come in handy later on.

If a behavior is produced by a process of practical reasoning that includes an ultimate desire for the well-being of others, then that behavior is psychologically altruistic. Psychological egoism denies that there are any ultimate desires of this sort; it maintains that all ultimate desires are self-interested. According to one influential version of egoism, often called *psychological hedonism*, there are only two sorts of ultimate desires: the desire for pleasure and the desire to avoid pain. Another, less restrictive, version of egoism allows that people may have a much wider range of ultimate self-interested desires, including desires for their own survival, for wealth, for power and for prestige. Egoism acknowledges that people sometimes have desires for the well-being of others, but it insists that all these desires are instrumental. Psychological altruism, by contrast, concedes that many ultimate desires are self-interested but insists that there are also some ultimate desires for the well-being of others. Since psychological altruism maintains that people have both self-interested ultimate desires and ultimate desires for the well-being of others, Sober and Wilson sometimes refer to the view as *motivational pluralism*.

[3] For some substantive discussion of the question see Stich et al. (in preparation).

[4] For a classic statement of this account of practical reasoning, see Goldman (1970).

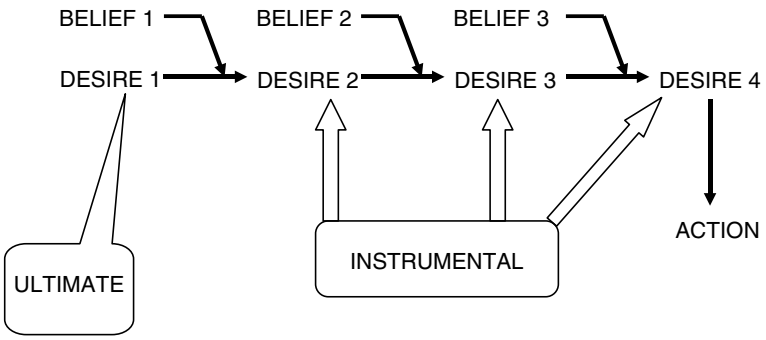

Figure 1. Practical reasoning, a causal process via which a desire and a belief give rise to or sustain another desire. An ultimate desire is one which is not produced by practical reasoning.

Sober and Wilson's evolutionary argument for psychological altruism

Sober and Wilson believe that there is an evolutionary argument for the existence of the sort of motivational structures required for psychological altruism. "Natural selection," they maintain, "is unlikely to have given us purely egoistic motives."[5] While granting that their case is "provisional" (8), they believe that their "analysis...provides evidence for the existence of psychological altruism" (12).

In setting out their argument, Sober and Wilson adopt the wise strategy of focusing on the case of parental care. Since the behaviors that organisms exhibit in taking care of their offspring are typically *not* altruistic in the evolutionary sense, we can simply put-off to the side whatever worries there may be about the existence of evolutionary altruism. Given the importance of parental care in many species, it is all but certain that natural selection played a significant role in shaping that behavior. And while different species no doubt utilize very different processes to generate and regulate parental care behavior, it is plausible to suppose that in humans *desires* play an important role in that process. Sober and Wilson believe that evolutionary considerations can help us determine the nature of these desires. Here is how they make the point:

> Although organisms take care of their young in many species, human parents provide a great deal of help, for a very long time, to their children. We expect that when parental care evolves in a lineage, natural selection is relevant to explaining why this transition occurs. Assuming that human parents take care of their children because of the desires they have, we also expect that evolutionary considerations will help illuminate what the desires are that play this motivational role." (301)

[5] Sober and Wilson (1998), p. 12. Hereafter, all quotes from Sober and Wilson (1998) will be identified by page numbers in parentheses.

Of course, as Sober and Wilson note, we hardly need evolutionary arguments to tell us about the content of some of the desires that motivate parental care. But it is much harder to determine whether these desires are instrumental or ultimate, and it is here, they think, that evolutionary considerations can be of help.

> We conjecture that human parents typically *want* their children to do well – to live rather than die, to be healthy rather than sick, and so on. The question we will address is whether this desire is merely an instrumental desire in the service of some egoistic ultimate goal, or part of a pluralistic motivational system in which there is an ultimate altruistic concern for the child's welfare. We will argue that there are evolutionary reasons to expect motivational pluralism to be the proximate mechanism for producing parental care in our species. (302)

Since parental care is essential in our species, and since providing it requires that parents have the appropriate set of desires, the processes driving evolution must have solved the problem of how to assure that parents would have the requisite desires. There are, Sober and Wilson maintain, three kinds of solutions to this evolutionary problem.

> A relatively direct solution to the design problem would be for parents to be psychological altruists – let them care about the well-being of their children as an end in itself. A more indirect solution would be for parents to be psychological hedonists[6] – let them care only about attaining pleasure and avoiding pain, but let them be so constituted that they feel good when their children do well and feel bad when their children do ill. And of course, there is a pluralistic solution to consider as well – let parents have altruistic *and* hedonistic motives, both of which motivate them to take care of their children. (305)

"Broadly speaking," they continue, "there are three considerations that bear on this question"(305). The first of these is *availability*; for natural selection to cause a trait to increase in frequency, the trait must have been available in an ancestral population. The second is *reliability*. Since parents who fail to provide care run a serious risk of never having grandchildren, we should expect that natural selection will prefer a more reliable solution to a less reliable one. The third consideration is *energetic efficiency*. Building and maintaining psychological mechanisms will inevitably require an investment of resources that might be used for some other purpose. So, other things being equal, we should expect natural selection to prefer the more efficient mechanism. There is, Sober and Wilson maintain, no reason to think that a psychologically altruistic mechanism

[6] Sober and Wilson cast their argument as contest between altruism and *hedonism* because "[b]y pitting altruism against hedonism, we are asking the altruism hypothesis to reply to the version of egoism that is most difficult to refute."(297)

would be less energetically efficient than a hedonist mechanism, nor is there any reason to think that an altruistic mechanism would have been less likely to be available. When it comes to reliability, on the other hand, they think there is a clear difference between a psychologically altruistic mechanism and various possible hedonistic mechanisms: an altruistic mechanism would be more reliable, and thus it is more likely that the altruistic mechanism would be the one that evolved.

To make their case, Sober and Wilson offer a brief sketch of how hedonistic and altruistic mechanisms might work, and then set out a variety of reasons for thinking that the altruistic mechanism would be more reliable. However, it has long been my conviction that in debates about psychological processes, the devil is often in the details. So rather than relying on Sober and Wilson's brief sketches, I will offer somewhat more detailed accounts of the psychological processes that might support psychologically altruistic and psychologically egoistic parental behavior. After setting out these accounts, I'll go on to evaluate Sober and Wilson's arguments about reliability.

Figure 2 is a depiction of the process underlying psychologically altruistic behavior. In Figure 2, the fact that the agent's child needs help (represented by the unboxed token of 'My child needs help' in the upper left) leads to the belief, *My child needs help*. Of course, formation of this belief requires complex perceptual and cognitive processing, but since this part of the story is irrelevant to the issue at hand, it has not been depicted. The belief, *My child needs help*, along with other beliefs the agent has leads to a belief that a certain action, A*, is the best way to help her child. Then, via practical reasoning, this belief and the *ultimate* desire, *I do what will be most helpful for*

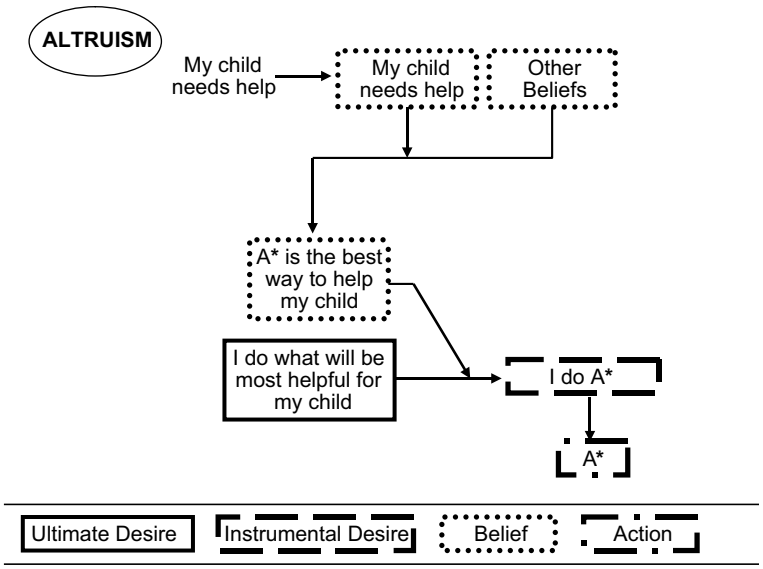

Figure 2. The process underlying psychologically altruistic behavior.

my child, leads to the desire to do A*. Since in this altruistic account the desire, *I do what will be most helpful for my child*, is an ultimate desire, it is not itself the result of practical reasoning. The hedonistic alternatives I'll propose retain all the basic structure depicted in Figure 2, but they depict the desire that *I do what will be most helpful for my child* as an instrumental rather than an ultimate desire.

The simplest way to do this is via what I'll call *Future Pain Hedonism*, which maintains that the agent believes she will feel bad in the future if she does not help her child now. Figure 3 is my sketch of Future Pain Hedonism. In it, the content of the agent's ultimate desire is hedonistic: *I maximize my pleasure and minimize my pain*. The desire, *I do what is most helpful for my child*, is an instrumental desire, generated via practical reasoning from the ultimate hedonistic desire along with the belief that *If I don't do what is most helpful for my child I will feel bad*.

Figure 4 depicts another, more complicated, way in which the desire, *I do what is most helpful to my child*, might be the product of hedonistic practical reasoning, which I'll call *Current Pain Hedonism*. On this account, the child's need for help causes the parent to feel bad, and the parent believes that if she feels bad because her child needs help and she does what is most helpful, she will stop feeling bad. This version of hedonism is more complex than the previous version, since it includes an affective state – feeling bad – in addition to various beliefs and desires, and in order for that affective state to influence practical reasoning, the parent must not only experience it, but know (or at least believe) that she is experiencing it, and why.

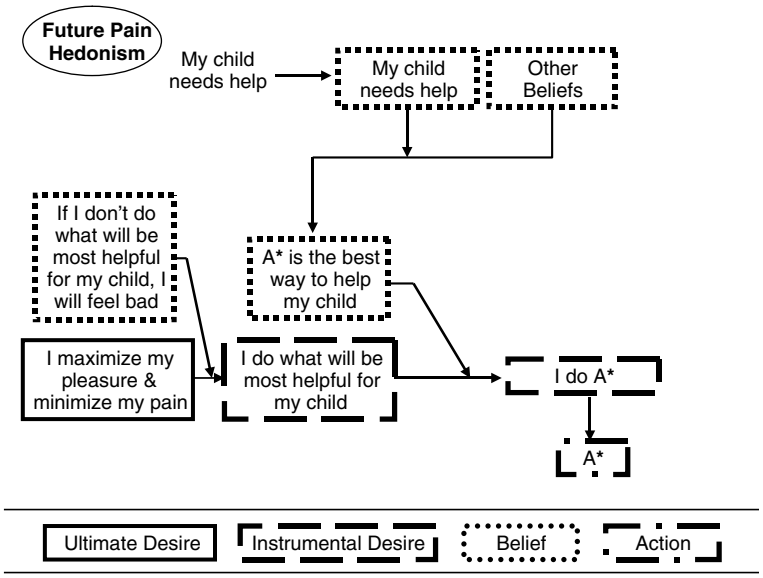

Figure 3. The process underlying future pain hedonism.

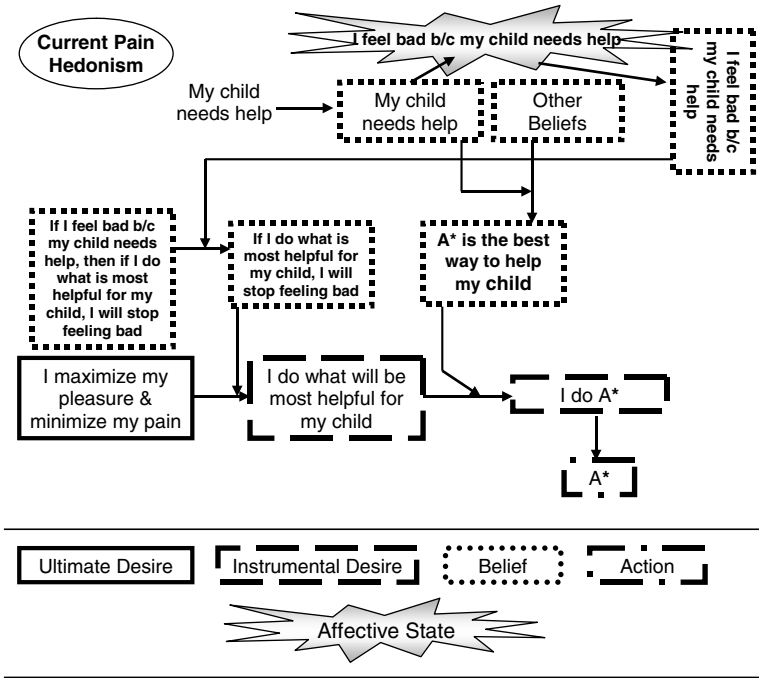

Figure 4. The process underlying current pain hedonism.

In their attempt to show that natural selection would favor an altruistic process over the hedonistic alternatives, Sober and Wilson offer a number of arguments, all of them focused on the more complicated Current Pain Hedonism, though they think that "the argument would remain the same if we thought of the hedonist as acting to avoid future pain" (318). In discussing these arguments, I'll start with three that I don't find very plausible; I'll then take up one that I think poses a serious challenge to hedonism and leads to some important questions about how, exactly, psychological egoism and psychological altruism should be understood.

A first pair of arguments both focuses on the causal link between believing that one's child needs help and feeling an appropriate level of distress or pain. The worry raised by the first argument is that the link could occasionally fail.

> If the fitness of hedonism depends on how well correlated the organism's pleasure and pain are with its beliefs about the well-being of its children, how strong is this correlation apt to be? (315)...[W]e think 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 [altruism] is that its reliability does not depend on the strength of such correlations." (316, emphasis in the original)

The second argument focuses on the fact that, to do its job appropriately, the mechanism underlying the belief-to-affect link must not only produce pain or distress, it must produce *lots* of it.

> Hedonism assumes that evolution produced organisms – ourselves included – in which psychological pain is strongly correlated with having beliefs of various kinds. In the context of our example of parental care, the hedonist asserts that whenever the organism believes that its children are well off, it tends to experience pleasure; whenever the organism believes that its children are doing badly, it tends to feel pain. What is needed is not just that *some* pleasure and *some* pain accompany these two beliefs. The amount of pleasure that comes from seeing one's children do well must exceed the amount that comes from eating chocolate ice cream and from having one's temples massaged to the sound of murmuring voices. This may require some tricky engineering... To achieve simplicity at the level of ultimate desires, complexity is required at the level of instrumental desires. This complexity must be taken into account in assessing the fitness of hedonism.[7] (315)

Sober and Wilson are certainly right that current pain hedonism requires the affect generated by the belief that one's child is doing well or badly be of an appropriate magnitude, and that this will require some psychological engineering that is not required by the altruist process. They are also right that the mechanism responsible for this belief-to-affect link will not establish a perfect correlation between belief and affect; like just about any psychological mechanism it is bound to fail now and then.

However, I don't think that either of these facts offers much reason to think that natural selection would favor the altruistic process. To see why, let's first consider the fact that the belief-to-affect link will be less than perfectly reliable. It seems that natural selection has built lots of adaptively important processes by using links between categories of belief and various sorts of affective states. Emotions like anger, fear and disgust, which play a crucial role in regulating behavior, are examples of states that are often triggered by different sorts of beliefs. And in all of these cases, it seems (logically) possible to eliminate the pathway that runs via affect, and replace it with an ultimate desire to behave appropriately when one acquires a triggering belief. Fear, for example, might be replaced by an ultimate desire to take protective action when you believe

[7] It is perhaps worth noting that, *pace* Sober and Wilson, neither of these arguments applies to Future Pain Hedonism, since that version of hedonism does not posit the sort of belief-to-affect link that Sober and Wilson are worried about. I should also note that, for simplicity, in discussing these arguments I'll ignore the pleasure engendered by the belief that one's child is well off and focus on the pain or distress engendered by the belief that one's child is doing badly.

that you are in danger. Since natural selection has clearly opted for an emotion mediation system in these cases rather than relying on an ultimate desire that avoids the need for a belief-to-affect link, we need some further argument to show that natural selection would not do the same in the case of parental care, and Sober and Wilson do not offer one.

The second argument faces a very similar challenge. It will indeed require some "tricky engineering" to be sure that beliefs about one's children produce the right amount of affect. But much the same is true in the case of other systems involving affect. For the fear system to work properly, seeing a tiger on the path in front of you must generate quite intense fear – a lot more than would be generated by your belief that if you run away quickly you might stub your toe. While it no doubt takes some tricky engineering to make this all work properly, natural selection was up to the challenge. Sober and Wilson give us no reason to think natural selection was not up to the challenge in the case of parental care as well.[8]

A third argument offered by Sober and Wilson is aimed at showing that natural selection would likely have preferred a system for producing parental care, which they call 'PLUR', in which *both* hedonistic motivation and altruistic motivation plays a role, over a "monistic" system that relies on hedonism alone. The central idea is that, under many circumstances, two control mechanisms are better than one.

> PLUR postulates two pathways from the belief that one's children need help to the act of providing help. If these operate at least somewhat independently of each other, and each on its own raises the probability of helping, then the two together will raise the probability of helping even more. Unless the two pathways postulated by PLUR hopelessly confound each other, PLUR will be more reliable than HED [hedonism]. PLUR is superior because it is a *multiply connected control device*. (320, italics in the original)

Sober and Wilson go on to observe that "multiply connected control devices have often evolved." They sketch a few examples, then note that "further examples could be supplied from biology, and also from engineering, where intelligent designers supply machines (like the space shuttle) with backup systems. Error is inevitable, but the chance of disastrous error can be minimized by well-crafted redundancy" (320).

Sober and Wilson are surely right that well-crafted redundancy will typically improve reliability and reduce the chance of disastrous error. They are also right that both natural selection and intelligent human designers have produced lots of systems with this sort of redundancy. But, as the disaster which

[8] Edouard Machery has pointed out another problem with the "tricky engineering" argument. On Sober and Wilson's account, altruists will have many ultimate desires in addition to the desire to do what will be most helpful for their children. So to insure that the desire leading to parental care usually prevails will *also* require some tricky engineering.

befell the Columbia space shuttle vividly illustrates, human engineers also often design crucial systems *without* backups. So too does natural selection, as people with damaged hearts or livers, or with small but disabling strokes, are all too well aware. One reason for lack of redundancy is that redundancy almost never comes without costs, and those costs have to be weighted against the incremental benefits that a backup system provides. Since Sober and Wilson offer us no reason to believe that, in the case of parental care, the added reliability of PLUR would justify the additional costs, their redundancy argument lends no support to the claim that natural selection would prefer PLUR to a monistic hedonism, or, for that matter, to a monistic altruism.

Sober and Wilson's fourth argument raises what I think is a much more troublesome issue for the hedonistic hypothesis.

> Suppose a hedonistic organism believes on a given occasion that providing parental care is the way for it to attain its ultimate goal of maximizing pleasure and minimizing pain. What would happen if the organism provides parental care, but then discovers that this action fails to deliver maximal pleasure and minimal pain? If the organism is able to learn from experience, it will probably be less inclined to take care of its children on subsequent occasions. Instrumental desires tend to diminish and disappear in the face of negative evidence of this sort. This can make hedonistic motivation a rather poor control device." (314) ...[The] instrumental desire will remain in place only if the organism ... is trapped by an unalterable illusion. (315)

Sober and Wilson are not as careful as they should be here. When it turns out that parental care does not produce the expected hedonic benefits, the hedonistic organism needs to have some beliefs about *why* this happened before it can effectively adjust its beliefs and instrumental desires. If, for example, the hedonist portrayed in Figures 3 or 4 comes to believe (perhaps correctly) that it was mistaken in inferring that A* was the best way to help, then it will need to adjust some of the beliefs that led to that inference, but the beliefs linking helping to the reduction of negative affect will require no modification. But despite this slip, I think that Sober and Wilson are onto something important here. Both versions of hedonism that I've sketched rely quite crucially on beliefs about the relation between helping behavior and affect. In the case of Future Pain Hedonism, as elaborated in Figure 3, the crucial belief is: *If I don't do what will be most helpful for my child, I will feel bad.* In the version of Current Pain Hedonism sketched in Figure 4, it's: *If I feel bad because my child needs help, then if I do what is most helpful for my child, I will stop feeling bad.* These beliefs make empirical claims, and like other empirical beliefs they might be undermined by evidence (including misleading evidence) or by more theoretical beliefs (rational or irrational) that a person could acquire by a variety of routes. This makes the process underlying parental care look quite vulnerable to disruption and suggests that natural selection would likely opt for some

more reliable way to get this crucial job done.[9] The version of altruism depicted in Figure 2 fits the bill nicely. By making the desire, *I do what will be most helpful for my child*, an ultimate desire, it sidesteps the need for empirical beliefs that might all too easily be undermined.

I think this is both a powerful argument for psychological altruism and an original one, though ultimately I am not persuaded. To explain why, we'll have to clarify what the altruist and the egoist are claiming. Altruists, recall, maintain that people have ultimate desires for the well-being of others, while egoists believe that all desires for the well-being of others are instrumental, and that all of our ultimate desires are self-interested. An instrumental desire is a desire that is produced or sustained by a process of practical reasoning like the one depicted in Figure 1 in which a desire and a belief give rise to or sustain another desire. In the discussion of practical reasoning, in the 'Preliminaries' section, nothing was said about the notion of *belief*; it was simply taken for granted. Like other writers in this area, including Sober and Wilson, I tacitly adopted the standard view that beliefs are inferentially integrated representational states that play a characteristic role in an agent's cognitive economy. To say that a belief is *inferentially integrated* is to say (roughly) that it can be both generated and removed by inferential processes that can take any (or just about any) other beliefs as premises.

While inferentially integrated representational states play a central role in many discussions of psychological processes and cognitive architecture, the literature in both cognitive science and philosophy also often discusses belief-like states that are "stickier" than this. Once they are in place, these "stickier" belief-like states are harder to modify by acquiring or changing other beliefs. They are also typically unavailable to introspective access. In Stich (1978), they were dubbed *sub-doxastic states*. Perhaps the most familiar example of sub-doxastic states are the grammatical rules that, according to Chomsky and his followers, underlie speech production, comprehension and the production of linguistic intuitions. These representational states are clearly not inferentially integrated, since a speaker's explicit beliefs about them typically has no effect on them. A speaker can, for example, have thoroughly mistaken beliefs about the rules that govern his linguistic processing without those beliefs having any effect on the rules or on the linguistic processing that they subserve. Another important example are the *core beliefs* posited by the psychologists Susan Carey and Elizabeth Spelke (Carey and Spelke 1996; Spelke 2000, 2003). These

[9] Note that the vulnerability to disruption we're considering now is likely to be a much more serious problem than the vulnerability that was center stage in Sober and Wilson's first argument. In that argument, the danger posed for the hedonistic parental care system was that "the psychological pain that hedonism postulates" might not be "*perfectly* correlated with believing that one's children are doing badly" (316, emphasis in the original). But, absent other problems, a hedonistic system in which belief and affect were highly – though imperfectly – correlated would still do quite a good job of parental care. Our current concern is with the stability of the crucial belief linking helping behavior and affect. If that belief is removed the hedonistic parental care system simply crashes, and the organism will not engage in parental care at all, except by accident.

are innate representational states that underlie young children's inferences about the physical and mathematical properties of objects. In the course of development, many people acquire more sophisticated theories about these matters, some of which are incompatible with the innate core beliefs. But, if Carey and Spelke are correct, the core beliefs remain unaltered by these new beliefs and continue to affect people's performance in a variety of experimental tasks. Although sub-doxastic states are sticky and hard to remove, they do play a role in *inference-like* interactions with other representational states, though their access to other representational premises and other premises' access to them is limited. In *The Modularity of Mind*, Fodor (1983) notes that representational states stored in the sorts of mental modules he posits are typically sub-doxastic, since modules are "informationally encapsulated". But not all sub-doxastic states need reside in Fodorian modules.

Since sub-doxastic states can play a role in inference-like interactions, and since practical reasoning is an inference-like interaction, it is possible that sub-doxastic states play the belief-role in some instances of practical reasoning. So, for example, rather than the practical reasoning structure illustrated in Figure 1, some examples of practical reasoning might have the structure shown in Figure 5. What makes practical reasoning structures like this important for our purposes is that, since SUB-DOXASTIC STATE 1 is difficult or impossible to remove using evidence or inference, DESIRE 2 will be reliably correlated with DESIRE 1.

Let's now ask whether, in Figure 5, DESIRE 2 is instrumental or ultimate? As we noted earlier, the objects of ultimate desires are typically characterized as "desired for their own sakes" while instrumental desires are those that agents have only because they think that satisfying the desire will lead to the satisfaction of some other desire. In Figure 5, the agent has DESIRE 2 only because he thinks that satisfying the desire will lead to the satisfaction of DESIRE 1. So it looks like the natural answer to our question is that DESIRE 2 is instrumental; the only ultimate desire depicted in Figure 5 is DESIRE 1.

If this is right, if desires like DESIRE 2 are instrumental rather than ultimate, then Sober and Wilson's evolutionary argument for psychological altruism is in trouble. The central insight of that argument was that both

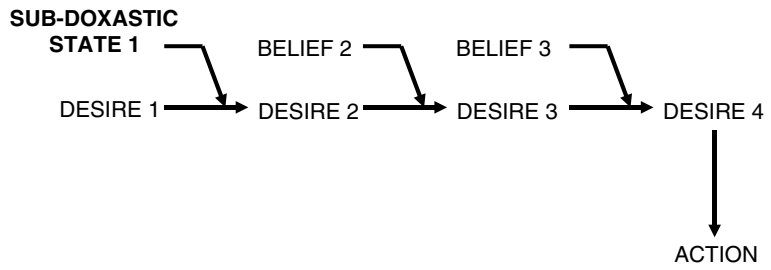

Figure 5. An episode of practical reasoning in which a sub-doxastic state plays a role.

versions of hedonism rely on empirical beliefs which might all too easily be undermined by other beliefs the agent might acquire. Suppose, however, that in Figures 3 and 4, the representations

If I don't do what will be most helpful for my child, I will feel bad

and

If I feel bad because my child needs help, then if I do what is most helpful for my child, I will stop feeling bad

are not beliefs but sticky sub-doxastic states. If we grant that desires produced or sustained by a desire and a sub-doxastic state count as instrumental desires, not ultimate desires, then the crucial desire whose presence Sober and Wilson sought to guarantee by making it an ultimate desire, viz.

I do what will be most helpful for my child

is no longer at risk of being undermined by other beliefs. Since the crucial desire is reliably present in both the Altruistic model and in both versions of the Hedonist model, natural selection can't prefer Altruism because of its greater reliability in getting a crucial job done.

In a passage from Sober and Wilson quoted earlier, they noted that when an instrumental desire does not lead to the expected hedonic pay-off, the "desire will remain in place only if the organism ... is trapped by an unalterable illusion" (315). But as a number of authors have noted, some illusions – or as I would prefer to put it, some belief-like representational states that are not strictly true – are conducive to fitness (Stich 1990; Plantinga 1993; Sober 1994; Godfrey-Smith 1996). In a variety of domains, it appears that natural selection has used sub-doxastic states and processes that have some of the features of mental modules to insure that those representations stay put and are not undermined by the systems that revise beliefs. Since natural selection often exploits the same trick over and over again, it is entirely possible that, when faced with the problem of assuring that parents were motivated to care for their children, this was the strategy it selected. My conclusion, of course, is *not* that parental care is subserved by an egoistic psychological process, but rather that Sober and Wilson's argument leaves this option quite open. Their analysis does not "provide... evidence for the existence of psychological altruism" (12).

Acknowledgements

I'm grateful to Elliott Sober, Edouard Machery, Kim Sterelny and an anonymous referee for helpful comments on earlier drafts of this paper.

References

Beatty J. 1992. Fitness: theoretical contexts. In: Keller E. and Lloyd E. (eds), Keywords in Evolutionary Biology, Harvard University Press, Cambridge, MA.

Carey S. and Spelke E. 1996. Science and core knowledge. Philos Sci 63(4): 515–533.

Fodor J. 1983. The Modularity of Mind. MIT Press. Cambridge, MA.

Ghiselin M. 1974. The Economy of Nature and the Evolution of Sex. University of California Press, Berkeley.

Godfrey-Smith P. 1996. Complexity and the Function of Mind in Nature. Cambridge University Press, Cambridge.

Goldman A. 1970. A Theory of Human Action. Prentice-Hall, Englewood-Cliffs, NJ.

Plantinga A. 1993. Warrant and Proper Function. Oxford University Press, Oxford.

Sober E. 1994. The adaptive advantage of learning and a priori prejudice. In: From a Biological Point of View. Cambridge University Press, Cambridge.

Sober E. and Wilson D.S. 1998. Unto Others: The Evolution and Psychology of Unselfish Behavior. Harvard University Press, Cambridge, MA.

Spelke E. 2000. Core knowledge. Am Psychol 55: 1233–1243.

Spelke E. 2003. Core knowledge. In: Kanwisher N. and Duncan J. (eds), Attention and Performance, Vol. 20: Functional Neuroimaging of Visual Cognition. Oxford University Press.

Stich S. 1978. Beliefs and sub-doxastic states. Philos Sci 45: 499–518.

Stich S. 1990. The Fragmentation of Reason. MIT Press, Cambridge, MA.

Stich S., Doris J. and Roedder E. (in preparation). Egoism *vs*. altruism. In: Doris J. et al. (eds), The Handbook of Moral Psychology. Oxford University Press (to be published).