The Moral Psychology Handbook
edited by
John Doris & the Moral Psychology Research Group

Oxford University Press, 2011

5

Altruism

STEPHEN STICH, JOHN M. DORIS, AND ERICA ROEDDER

1. Philosophical Background

People sometimes behave in ways that benefit others, and they sometimes do this while knowing that their helpful behavior will be costly, unpleasant, or dangerous. But at least since Plato's classic discussion in the second book of the *Republic*, debate has raged over *why* people behave in these ways. Are their motives altruistic, or is their behavior ultimately motivated by self-interest? According to Thomas Hobbes, who famously advocated the latter option,

> No man giveth but with intention of good to himself, because gift is voluntary; and of all voluntary acts, the object is to every man his own good; of which, if men see they shall be frustrated, there will be no beginning of benevolence or trust, nor consequently of mutual help. (Hobbes, 1651/1981: ch. 15)

This selfish or *egoistic* view of human motivation has had no shortage of eminent advocates, including La Rochefoucauld (1665/2007), Bentham (1824: 392–393), and Nietzsche (1881/1997: 148).[1] Egoism is also arguably the dominant view of human motivation in much contemporary social science, particularly in economics (see Grant, 1997; Miller, 1999). Dissenting voices, though perhaps fewer in number, have been no less eminent. Butler (1726: esp. *Sermon* XI), Hume (1751/1975: 272, 298), Rousseau (1754/1985) and Adam Smith (1759/1853) have all argued that, sometimes at least, human motivation is genuinely *altruistic*.

Although the issue that divides egoistic and altruistic accounts of human motivation is substantially empirical, competing answers may have profound

[1] Interpretation of historical texts is, of course, often less than straightforward. While there are passages in the works of each of these philosophers that can be interpreted as advocating egoism, scholars might debate whether these passages reflect the author's considered option. Much the same is true for the defenders of altruism mentioned below.

consequences for moral theory.[2] For example, Kant (1785: sec. 1, para. 12) famously argued that a person should act "not from inclination but from duty, and by this would his conduct first acquire true moral worth." But egoism maintains that *all* human motivation is ultimately self-interested, and if so, people *can't* act "from duty" in the way that Kant urged. Thus, if egoism were true, Kant's account would entail that *no* conduct has "true moral worth."[3] The same difficulty obtains for Aristotle's insistence, in *Nicomachean Ethics* ii 4, that acting virtuously requires choosing the virtuous action "for its own sake." Once again, if all actions are ultimately self-interested, it is unclear how any action can meet Aristotle's criterion, and we must consider the possibility of skepticism about virtuous action.

Whether or not such difficulties can be ameliorated, there can be little doubt that they resonate widely. It is easy to find philosophers who suggest that altruism is required for morality or that egoism is incompatible with morality—and easier still to find philosophers who claim that *other* philosophers think this. In the standard reference work we consulted (LaFollette, 2000a), examples abound:

> Moral behavior is, at the most general level, altruistic behavior, motivated by the desire to promote not only our own welfare but the welfare of others. (Rachels, 2000: 81)

> [O]ne central assumption motivating ethical theory in the Analytic tradition is that the function of ethics is to combat the inherent egoism or selfishness of individuals. Indeed, many thinkers define the basic goal of morality as "selflessness" or "altruism." (Schroeder, 2000: 396)

> Philosophers since Socrates worried that humans might be capable of acting only to promote their own self-interest. But if that is all we can do, then it seems morality is impossible. (LaFollette, 2000b: 5)

If these philosophers are right, and egoism is true, moral skepticism may be in order.

Additionally, if egoism is true, then we face a dilemma in answering the venerable question, "Why should I be moral?" If this question requests a justification that can actually motivate an agent to act morally, then egoism poses strong constraints on how it can be answered: the answer will have to ground moral motivation in the agent's self-interest. On the other hand, the question "Why should I be moral?" might be construed as inquiring after a merely theoretic justification—one that might or might not motivate a person to act morally. If the question is construed in this way, and if egoism is true, then there may well be a disconnect between moral theory and moral motivation. Our best moral theories may well answer the question "Why be moral?" by appealing to symmetry between the self and other or ideals of social harmony. Unfortunately, these appeals—while perhaps enlightening—will be motivationally moot.

Yet another cluster of issues surrounds the principle *"ought" implies "can."* Depending on the modal strength of egoism, it could turn out that persons simply cannot be motivated by anything other than self-interest. If this is true, and if "ought" implies "can," then it cannot be said that humans *ought* to be motivated by anything other than self-interest.

There are related implications for political philosophy. If the egoists are right, then the *only* way to motivate prosocial behavior is to give people a selfish reason for engaging in such behavior, and this constrains the design of political institutions intended to encourage civic-minded behavior. John Stuart Mill, who like Bentham before him was both a utilitarian and an egoist, advocated a variety of manipulative social interventions to engender conformity with utilitarian moral standards.[4]

Since the empirical debate between egoists and altruists appears to have such striking implications, it should come as no surprise that psychologists and other scientists have done a great deal of work aimed at determining which view is correct. But before we turn to the empirical literature, it is important to get clearer on what the debate is really about. We shall begin, in Section 2, with a brief sketch of a cluster of assumptions about human desires, beliefs, actions, and motivation that are widely shared by historical and contemporary authors on both sides in the debate. With this as background, we'll be able to offer a more sharply focused account of the debate.[5] In Section 3, our focus will be on links between evolutionary theory and the egoism/altruism debate. There is a substantial literature employing evolutionary theory on each side of the

[2] A note on terminology: we shall use the terms "egoism" and "altruism" for views about the nature of human motivation that will be explained in more detail below. Other authors prefer to call these views "psychological egoism" and "psychological altruism" to distinguish them from normative claims about how people should behave, and from an evolutionary notion of altruism that we'll discuss in Section 3. Since both evolutionary altruism and psychological altruism will be considered in Section 3, we'll use "psychological altruism" for the latter notion in that section.

[3] Kant appears to consider this possibility: "A cool observer, one that does not mistake the wish for good, however lively, for its reality, may sometimes doubt whether true virtue is actually found anywhere in the world" (Kant, 1785: sec. 2.)

[4] For example, Mill suggests instilling "hope of favor and the fear of displeasure from our fellow creatures or from the Ruler of the Universe" (Mill, 1861/2001: ch. 3). Another proposal was to instill a feeling of conscience: "a pain, more or less intense, attendant on violation of duty . . . . This feeling [is] all encrusted over with collateral associations . . . derived from . . . fear; from all the forms of religious feeling; from self-esteem . . . and occasionally even self-abasement" (ibid.).

[5] For more on the history of the debate between egoists and altruists, see Broad (1930); MacIntyre (1967); Nagel (1970); Batson (1991), chs. 1–3; Sober & Wilson (1998), ch. 9.

issue. However, it is our contention that neither camp has offered a convincing case. We are much more sanguine about recent research on altruism in social psychology, which will be our topic in Section 4. Although we don't think this work has resolved the debate, we shall argue that it has made illuminating progress—progress that philosophers interested in the question cannot afford to ignore.

2. Desires and Practical Reasoning

As we understand it, the egoism/altruism debate is typically structured by three familiar assumptions about the nature of desire and practical reasoning. First, parties to the debate typically assume that genuine actions are caused by desires. If Albert wants (or desires) to raise his hand, and if this desire causes his hand to go up, then this behavior counts as an *action*.[6] If, by contrast, Beth has no desire to raise her hand but it goes up anyway because of a nervous tic, or because Albert lifts it, then this behavior does not count as an action. Within this debate the term "desire" is used in a very inclusive way; it is a general term covering many motivational states. Donald Davidson (1963: 685–686) famously characterized all such motivational states as *pro-attitudes* (we prefer the term *desires*) and emphasized that this was meant to include states such as wantings, urges, promptings, and yens, enduring character traits like a taste for loud company, and passing fancies like a sudden desire to touch a woman's elbow!

A second assumption is that desires and beliefs can interact to generate a chain of new desires via a process that is often called *practical reasoning*. Thus, for example, if Cathy wants an espresso, and if she acquires the belief that the Starbucks on Main Street is the best place to get an espresso, this may cause her to desire to go to the Starbucks on Main Street. If she believes that the best way to get to Main Street is to take the number 43 bus, this, along with the newly formed desire to go to Starbucks, may cause a desire to take the number 43 bus. And so on.[7] Cathy's desire to take the bus and her desire to go to Starbucks are *instrumental* desires. She has them only because she wants that espresso.

But, and this is the final assumption, not all desires can be instrumental desires. If we are to avoid circularity or an infinite regress, there must be some

[6] Actually, not just any causal link between the desire and the behavior will do, and a great deal of philosophical work has been devoted to specifying the appropriate connection. See Grice (1971); Davidson (1980); Bratman (1987); Velleman (1989); Wilson (2002).

[7] For a classic statement of this conception of action, see Goldman (1970).

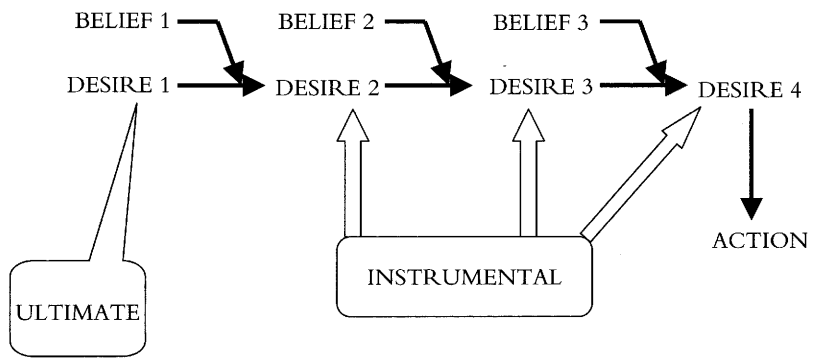

Figure 5.1. Practical reasoning
A causal process in which pre-existing desires and beliefs interact to generate a chain of new instrumental desires

desires that are *not* generated by our belief that satisfying them will facilitate satisfying some other desire. These are our "ultimate" desires; the states of affairs that will satisfy them are desired, as is often said, *for their own sake*. Figure 5.1 depicts all this in a format that will come in handy later on.

With this background in place, it can be seen that the debate between egoists and altruists is a debate about *ultimate* desires. Egoists maintain that *all* of our ultimate desires are selfish, and although altruists concede that some of our ultimate desires are selfish, they insist that people can and do have ultimate desires for the well-being of others. However, this account of the debate leaves a number of sticky issues yet to be addressed.[8]

One of these concerns the distinction between ultimate and instrumental desires. Our explanation of the distinction suggests that all desires fall into one category or the other, but it certainly seems to be possible for some desires to be both. Consider, for example, the desire to avoid pain. This is often cited as a paradigm case of an ultimate desire. But now suppose that David is in pain and desires not to be. Suppose too that David believes his mother is deeply distressed by the fact that he is in pain, and that David *also* wants his pain to end in order to alleviate his mother's distress. Should we count this desire as ultimate or instrumental? It seems to be both. In the context of the egoism/altruism debate we think that both sides should agree that a

[8] This account of the debate raises interesting questions about the role of emotions. There are complex issues here. However, in this chapter we shall be assuming that, for any emotion, if that emotion is to serve as the origin of action, it must cause some ultimate motivational state (i.e. an ultimate desire), or include some motivational component (again, an ultimate desire). We can then ask whether that ultimate desire is self- or other-oriented. This, in turn, will determine whether the action is egoistic or altruistic.

desire counts as ultimate if it has *any* ultimate "component" and that desires like David's clearly have an ultimate component. So if people's desires for the well-being of others ever have an ultimate component, then the altruist wins. But this still leaves us with the problem of saying more precisely what this talk about "ultimate components" amounts to.

Some progress can be made by appealing to counterfactuals: a person's desire has an ultimate component if she would continue to have the desire even if she no longer believed that satisfying the desire would lead to the satisfaction of some other desire. This is only a first pass, since, as clever philosophy majors know, it is easy to construct counter-examples to simple counterfactual analyses like this one.[9] For our purposes, however, no more nuanced account is required.

Another issue is how we are to interpret the notion of *self-interested* desires and desires *for the well-being of others*. According to one influential version of egoism, often called *hedonism*, there are only two sorts of ultimate desires: the desire for pleasure and the desire to avoid pain.[10] 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, power, and prestige.[11] While these desires are unproblematically self-interested, there are lots of other examples whose status is less clear. Is a desire for friendship self-interested? How about a desire to be involved in a mutually loving relationship? Or a desire to discover the cure for AIDS? Whether or not people have ultimate desires for any of these things is an open question. If they do, then hedonism is simply mistaken, since hedonism claims that there are only two sorts of ultimate desires. But what about non-hedonistic egoism; would it be refuted if people have ultimate desires like these? Given the conceptual uncertainties, we're inclined to think that egoism is vague. And similarly for altruism: if Albert has an ultimate desire that Beth be happy or that Cathy be cured of AIDS, then clearly the altruists are right and the egoists are wrong, since ultimate desires like these, if they exist, surely count as desires for the well-being of others. But suppose Albert has the ultimate desire that he and Beth be involved in a mutually loving relationship, or that *he* (as opposed to his rival in the next lab) cure Cathy of AIDS. Would *these* ultimate desires

[9] Martin (1994); Lewis (1997).

[10] Perhaps the most famous statement of hedonism is found in the first two sentences of Jeremy Bentham's *An Introduction to the Principles of Morals and Legislation* (1798/1996: 11): "Nature has placed mankind under the governance of two sovereign masters, *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do."

[11] Hedonists do not deny that many people desire such things, but they insist that these are not *ultimate* desires. According to the hedonist, people want wealth, power, and the rest only because they believe that being wealthy, powerful, etc. will lead to more pleasure and less pain.

suffice to show that altruism is right and egoism is wrong? Here again there is no clear and well-motivated answer. So altruism, like egoism, is vague. Fortunately, this vagueness need not be an insuperable problem in evaluating the competing theories, since there is no shortage of clear cases on both sides. If all ultimate desires are *clearly* self-interested, then the egoists are right. But if there are any ultimate desires that are clearly for the well-being of others, then the altruists are right and the egoists are wrong.

Before we turn to the empirical literature, there is one final complication that needs to be discussed. Despite the vagueness of the categories of self-interested desires and desires for the well-being of others, there seems to be a wide range of desires that clearly fall into *neither* category. Examples include the desire that great works of art be preserved and the desire that space exploration be pursued. Perhaps more interesting for moral theory are examples like the desire to do one's moral duty and the desire to obey God's commandments. Whether anyone holds these as *ultimate* desires is debatable. But if people do, then egoism is mistaken, since these desires are not self-interested. Interestingly, however, if ultimate desires like these exist, it would *not* show that altruism is true, since these desires are not desires for the well-being of others. Of course, a person who held the ultimate desire to do his moral duty might also believe that it was his duty to alleviate the suffering of the poor, and that belief might generate a desire to alleviate the suffering of the poor, which *is* a clear example of a desire for the well-being of others. But this lends no support to altruism, because the *ultimate* desire, in this case, is not for the well-being of others.

The topography has gotten a bit complicated; Figure 5.2 may help keep the contours clear. In that figure, we've distinguished four sorts of desires. Hedonism maintains that all ultimate desires fall into category 1. Egoism maintains that all ultimate desires fall into category 2, which has category 1 as

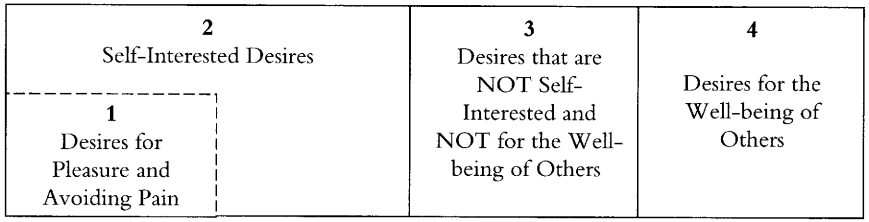

Figure 5.2. Four sorts of desires
Hedonism maintains that all ultimate desires are in category 1; egoism maintains that all ultimate desires are in category 2; altruism maintains that some ultimate desires are in category 4.

a subset. Altruism claims that *some* ultimate desires fall into category 4. Finally, if there are ultimate desires that fall into category 3 but none that fall into category 4, then both egoism and altruism are mistaken.

3. Altruism and Evolution

Readers familiar with some of the popular literature on the evolution of morality that has appeared in the last few decades might suspect that contemporary evolutionary biology has resolved the debate between egoists and altruists. For some readers—and some writers—seem to interpret evolutionary theory as showing that altruism is "biologically impossible." However, we maintain that the large literature on evolution and altruism has done very little to advance the philosophical debate. In this section, we shall make the case for this claim. We'll begin with arguments that purport to show that considerations drawn from evolutionary theory make altruism unlikely or impossible (except, perhaps, in a very limited range of cases). We'll then turn to a cluster of arguments, offered by Elliott Sober and David Sloan Wilson, that try to establish the opposite conclusion: evolutionary theory makes it appear quite likely that psychological altruism is true.

3.1. *Evolutionary Arguments against Altruism*

In discussions of evolution and altruism it is important to bear in mind a crucial distinction between two very different notions of altruism, which (following Sober and Wilson) we'll call *evolutionary altruism* and *psychological altruism*. Psychological altruism is the notion that has been at center stage in philosophical debates since antiquity. In most of this chapter, when we use the word "altruism," we are referring to psychological altruism. But in this section, to avoid confusion, we'll regularly opt for the longer label "psychological altruism." As we explained in Section 2, an organism is psychologically altruistic if and only if it has ultimate desires for the well-being of others, and a behavior is psychologically altruistic if and only if it is motivated by such a desire. By contrast, a behavior (or a behavioral disposition) is *evolutionarily altruistic* if and only if it reduces 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.[12] Since an

[12] Attempting a more precise account raises difficult issues in the philosophy of biology (e.g. Beatty, 1992); fortunately, our purposes do not require greater precision.

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.[13]

It is important to see 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 if it is entirely devoid of mental states 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 ultimate desires for the welfare of its own offspring. Behaviors resulting from that desire will be psychologically altruistic but not evolutionarily altruistic, since typically such behaviors will increase the inclusive fitness of the parent.

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 implicated in evolutionarily altruistic behavior with genes that did not foster evolutionarily altruistic behavior, and thus the evolutionarily altruistic behavior would disappear. In recent years, there has been a great deal of discussion of this problem. Some theorists have offered sophisticated models purporting to show how, in appropriate circumstances, evolutionary altruism could indeed evolve,[14] while others have maintained that evolutionary altruism is extremely unlikely to evolve in a species like ours, 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."

What is important for our purposes is that, even if the skeptics who doubt the existence of evolutionary altruism are correct, this would not resolve the philosophical debate over egoism and altruism since, as we have noted above, it entails nothing at all about the existence of psychological altruism. Unfortunately, far too many writers, including perhaps Ghiselin himself, have

[13] Some writers, including Sober and Wilson, define evolutionary altruism in terms of *individual* fitness (a measure of how many descendants an individual has) rather than inclusive fitness. For present purposes, we prefer the inclusive fitness account, since it facilitates a more plausible statement of the evolutionary argument aimed at showing that psychological altruism is possible only in very limited domains.

[14] See, for example, Sober & Wilson (1998), Part I.

made the mistake of assuming that the arguments against evolutionary altruism show that the sort of altruism that is of interest to moral theorists does not exist.

Although these considerations show that evolutionary theorists have no reason to deny that organisms can be psychological altruists, some authors have suggested that evolutionary theory permits psychological altruism only in very limited domains. The reasoning proceeds as follows: there are only two ways that a disposition to engage in behavior that helps other organisms but lowers one's own chance of survival and reproduction can evolve. One of these is the case in which the recipients of help are one's own offspring, or other close kin. Kin selection theory, pioneered by W. D. Hamilton (1963, 1964a, 1964b) makes it clear that in appropriate circumstances, genes leading to costly helping behavior will tend to spread throughout a population, provided that the recipients of the help are relatives, since this sort of helping behavior increases the number of copies of those genes that will be found in future generations. The other way in which a disposition to help can evolve requires that episodes of helping behavior are part of a longer-term reciprocal strategy in which the organism that is the beneficiary of helping behavior is subsequently disposed to help its benefactor. Building on ideas first set out in Trivers's (1971) classic paper on "reciprocal altruism," Axelrod and Hamilton (1981) described a simple "tit-for-tat" strategy, in which an organism helps on the first appropriate opportunity and then helps on subsequent opportunities if and only if the partner helped on the previous appropriate opportunity. They showed that tit-for-tat would be favored by natural selection over many other strategies, including a purely selfish strategy of never offering help but always accepting it. Since psychological altruism will lead to helping behavior, it is argued, psychological altruism can evolve only when a disposition to helping behavior can. So it is biologically possible for organisms to have ultimate desires to help their kin, and to help non-kin with whom they engage in ongoing reciprocal altruism. But apart from these special cases, psychological altruism can't evolve.

Versions of this influential line of thought can be found in many places (see, e.g., Nesse, 1990; Wright, 1994: chs. 8 & 9; Rachels, 1990: ch. 4). However, we think there is good reason to be very skeptical about the crucial assumption, which maintains that dispositions to helping behavior can evolve *only* via kin selection or reciprocal altruism. It has long been recognized that various sorts of group selection, in which one group of individuals leaves more descendants than another group, can lead to the evolution of helping behavior. Until recently, though, the reigning orthodoxy in evolutionary biology has been that the conditions under which group selection can act are highly unlikely

to occur in natural breeding populations, and thus group selection is unlikely to have played a substantial role in human evolution. This view has been boldly challenged by Sober and Wilson (1998), and while their views are very controversial, we think that the extent to which group selection played a role in human evolution is very much an open question.

Much less controversially, Boyd and Richerson (1992) have developed models demonstrating that helping behavior (and, indeed, just about *any* sort of behavior) can evolve if informal punishment is meted out to individuals who do not help in circumstances when they are expected to. In such circumstances, psychological altruism could be favored by natural selection. More recently, Sripada (2007) has argued that ultimate desires for the well-being of others could evolve via a rather different route. As Sripada observes, there are many situations in which people are better off if they act in a coordinated way, but where no specific way of acting is best. Driving on the right (or the left) is an obvious example: it matters not a whit whether the convention is to drive on the right or left, but failure to adopt *some* convention would be a disaster. In these situations several different "coordination equilibria" may be equally adaptive. To enable groups to reap the benefits of acting in a coordinated way, Sripada argues, natural selection may well have led to the evolution of a psychological mechanism that generates ultimate desires to adhere to locally prevailing customs or practices. And since some of those locally prevailing customs may require helping others, some of the ultimate desires produced by that psychological mechanism might well be psychologically altruistic. If Boyd and Richerson and Sripada are right, and we believe they are, then evolutionary theory gives us no reason to suppose that psychological altruism must be restricted to kin or to individuals involved in reciprocal exchanges. So, contrary to the frequently encountered presumption that evolutionary biology has resolved the debate between psychological egoists and psychological altruists in favor of egoism, it appears that evolutionary theory offers little succor to the egoists.

3.2. *Evolutionary Arguments for Altruism*[15]

In stark contrast with writers who think that evolutionary arguments show that psychological altruism is unlikely or impossible, Sober and Wilson (1998) believe that there are evolutionary arguments *for* the existence of psychological altruism. "Natural selection," they maintain, "is unlikely to have given us

[15] Much of this section is based on Stich (2007). We are grateful to Elliott Sober and Edouard Machery for helpful comments on this material.

purely egoistic motives."[16] 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 arguments, 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 aside 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:

> 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)

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 humans, 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

[16] Sober & Wilson (1998: 12). For the remainder of this section, all quotes from Sober & Wilson (1998) will be identified by page numbers in parentheses.

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[17]—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" (ibid.). 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 over 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 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, we believe that in debates about psychological processes, the devil is often in the details. So rather than relying on Sober and Wilson's brief sketches, we shall offer somewhat more detailed accounts of the psychological processes that might support hedonistic and psychologically altruistic parental behavior.

[17] 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). For expository purposes, we shall follow their lead here, although we are not committed to their evaluation of hedonism.

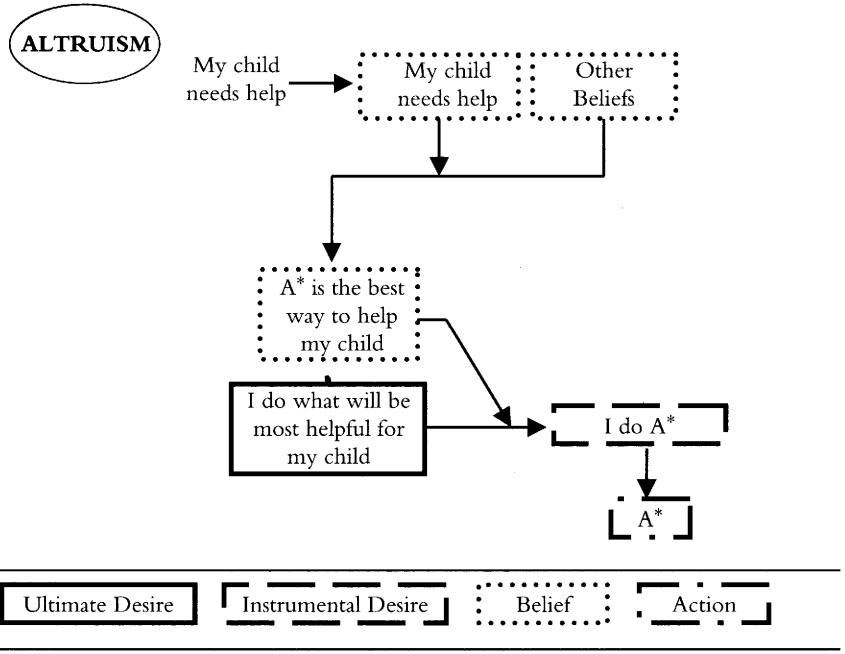

Figure 5.3. The process underlying psychologically altruistic behavior

After setting out these accounts, we'll go on to evaluate Sober and Wilson's arguments about reliability.

Figure 5.3 is a depiction of the process underlying psychologically altruistic behavior. In Figure 5.3, 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 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 we shall propose retain all of the basic structure depicted in Figure 5.3, 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 we shall call *future pain hedonism*, which maintains that the agent believes she will feel bad in the future if she does not help her

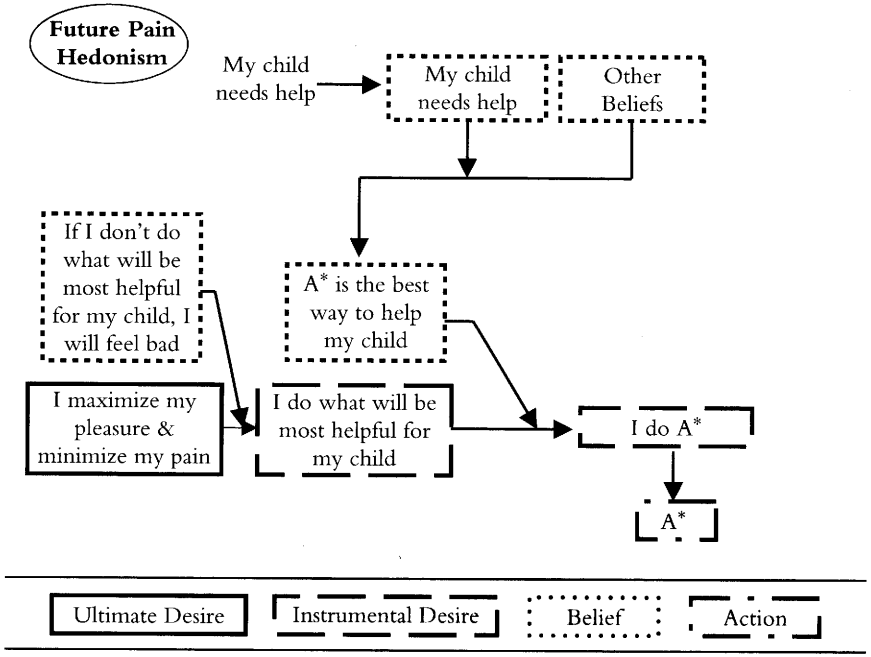

Figure 5.4. The process underlying future pain hedonism

child now. Figure 5.4 is our 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 5.5 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 we'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.

In their attempt to show that natural selection would favor an altruistic process over the hedonistic alternatives, Sober and Wilson offer a number

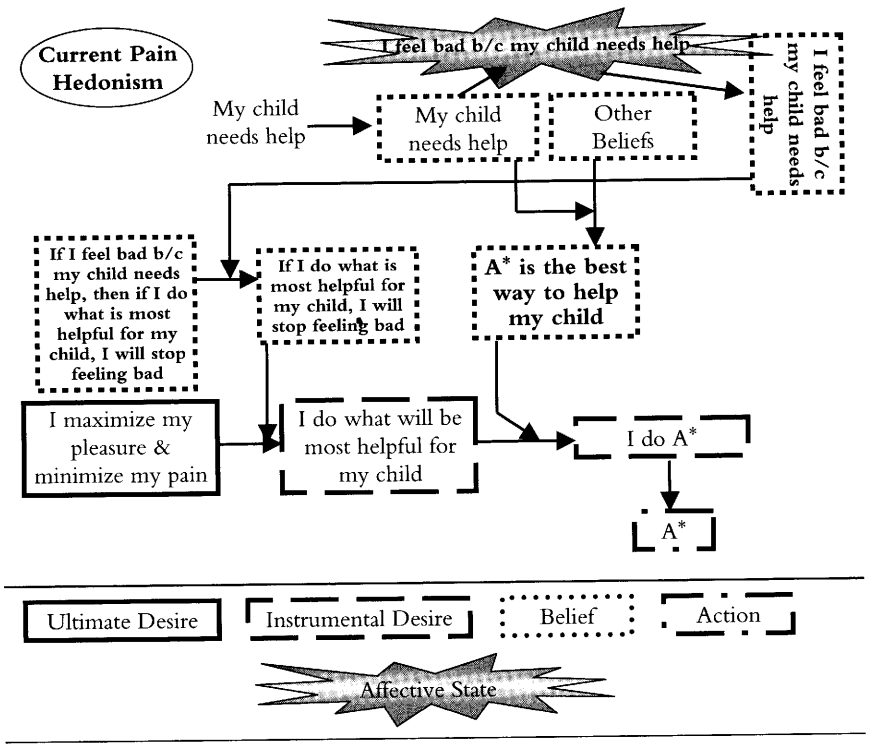

Figure 5.5. The process underlying current pain hedonism

of arguments, all of them focused on the more complicated current pain hedonism, although 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, we shall start with three that we don't find very plausible; we'll then take up one that we 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 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 . . . [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.[18] (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, we don't think that either of these facts offers much reason to believe 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 crucial roles 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 that you are in danger. Since natural selection has clearly opted for an emotion mediation system in these cases rather than relying on an

[18] 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 find suspect. We should also note that for the sake of simplicity, we'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.

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 any.

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 ran 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. Edouard Machery has pointed out to us 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 ensure that the ultimate desire leading to parental care usually prevails will *also* require some tricky engineering.

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" (for *pluralism*), in which *both* hedonistic motivation and altruistic motivation play a role, over a "monistic" system that relies on hedonism alone. The central idea is that, in 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, emphasis 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" (Ibid.).

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 that befell the Columbia space shuttle and a myriad of other technical catastrophes vividly illustrate, 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 weighed 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.[19]

Sober and Wilson's fourth argument raises what we 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 might have been more careful 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 5.4 or 5.5 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, we think that Sober and Wilson are onto something important. Both versions of hedonism that we'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 5.4, the

[19] Even if there were some reason to think that natural selection would prefer a redundant system, this would not, by itself, constitute an argument for altruism. Redundancy can exist in an entirely hedonistic system, for instance one that includes both current and future pain hedonism.

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 5.5, 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.[20] The version of altruism depicted in Figure 5.3 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.

We think this is an original and powerful argument for psychological altruism. Ultimately, however, we are not persuaded. To explain why, we'll have to clarify what the altruist and the egoist are claiming. Psychological altruists, recall, maintain that people have ultimate desires for the well-being of others, while psychological egoists believe that all desires for the well-being of others are instrumental, and that all of our ultimate desires are self-interested. As depicted in Figure 5.1, an instrumental desire is a desire that is produced or sustained entirely by a process of practical reasoning in which a desire and a belief give rise to or sustain another desire. In our discussion of practical reasoning (in Section 2), while a good bit was said about desire, nothing was said about the notion of *belief*; it was simply taken for granted. At this point, however, we can no longer afford to do so. Like other writers in this area, including Sober and Wilson, we 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

[20] Note that the vulnerability to disruption we're considering now is likely to be a much more serious problem than the vulnerability that was at 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 except by accident.

belief-like states that are "stickier" than this. Once they are in place, these "stickier" belief-like states are hard to modify by acquiring or changing other beliefs. They are also often unavailable to introspective access. In Stich (1978), they were termed *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 typically have 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 is the *core beliefs* posited by Carey and Spelke (Carey & Spelke, 1996; Spelke, 2000, 2003). These 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, although 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 5.1, some examples of practical reasoning might have the structure shown in Figure 5.6. 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 consider whether DESIRE 2 in Figure 5.6 is instrumental or ultimate. As we noted in Section 1, 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

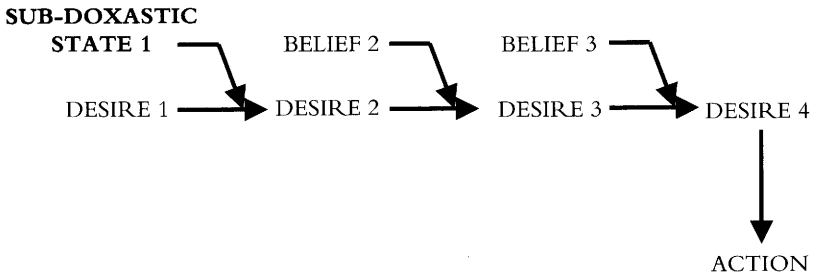

Figure 5.6. An episode of practical reasoning in which a sub-doxastic state plays a role

lead to the satisfaction of some other desire. In Figure 5.6, the agent has DESIRE 2 only because he thinks that satisfying the desire will lead to the satisfaction of DESIRE 1. So it looks as if the natural answer to our question is that DESIRE 2 is instrumental; the only ultimate desire depicted in Figure 5.6 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 versions of hedonism rely on empirical beliefs that might all too easily be undermined by other beliefs the agent might acquire. Suppose, however, that in Figures 5.4 and 5.5, 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 like DESIRE 2 in Figure 5.6, which are 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, i.e.

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.

As we've seen, Sober and Wilson contend that when an instrumental desire does not lead to the expected hedonic payoff, 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 we 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 ensure 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. Our 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).

Our central claim in this section has been that evolutionary theory offers little prospect for movement in philosophical debates between psychological egoism and psychological altruism. In 3.1 we saw that evolutionary considerations don't rule out psychological altruism or restrict its scope, and in 3.2 we've argued that Sober and Wilson's arguments—by far the most sophisticated attempt to make an evolutionary case *for* psychological altruism—are not convincing.

4. The Social Psychology of Altruism

We now turn from theoretical considerations purporting to make the existence of altruism seem likely (or unlikely) to attempts at directly establishing the existence of altruism through experimental observation. The psychological literature relevant to the egoism vs. altruism debate is vast, but in this section we shall focus primarily on the work of Daniel Batson and his associates, who have done some of the most important work in this area.[21] Batson, along with many other researchers, begins by borrowing an idea that has deep roots in philosophical discussions of altruism. Although the details and the terminology

[21] For useful reviews of the literature see Piliavin & Charng (1990); Batson (1991,1998); Schroeder et al. (1995); Dovidio et al. (2006). We are grateful to Daniel Batson for helpful discussion of the material in this section.

differ significantly from author to author, the central idea is that altruism is often the product of an emotional response to another's distress.

For example, Aquinas maintains that "mercy is the heartfelt sympathy for another's distress, impelling us to succour him if we can."[22] And Adam Smith tells us that "pity or compassion [is] the emotion we feel for the misery of others, when we either see it, or are made to conceive it in a very lively manner" and these emotions "interest [man] in the fortunes of others, and render their happiness necessary to him, though he derives nothing from it except the pleasure of seeing it."[23] While different writers have used different terms for the emotional response in question, Batson (1991: 58) labels it "empathy," which he characterizes as "an other-oriented emotional reaction to seeing someone suffer," and he calls the traditional idea that empathy leads to altruism the *empathy–altruism hypothesis*.

In this section we shall begin, in 4.1, by introducing Batson's account of empathy, and discussing some problems with that account. In 4.2, we'll evaluate Batson's claims about the causal pathway leading from perspective-taking to empathy and from empathy to helping behavior. In the remainder of the section, we'll look carefully at some of Batson's experiments that are designed to test the empathy–altruism hypothesis against a variety of egoistic alternatives. In a number of cases, we believe, the experiments have made important progress by showing that versions of egoism that have loomed large in philosophical discussion are not very promising. But in other cases we'll argue that Batson's experiments have not yet succeeded in undermining an egoistic alternative.

While the Batson group's experimental program is novel, the dialectical space is one, we dare say, that has exercised generations of introductory philosophy students: for many examples of helping behavior, the egoist and altruist can both offer psychological explanations consistent with their hypothesis, and the ensuing arguments concern which explanation is most plausible. Unfortunately, as generations of introductory philosophy students have found out, such arguments are bound to end in inconclusive speculation—so long as the competing explanations are not empirically evaluated. By showing how such evaluation may proceed, the Batson group has enabled progress in a shopworn debate.

4.1. *Empathy and Personal Distress*

Since there is no standardized terminology in this area, Batson's choice of the term "empathy" to label the emotion, or cluster of emotions, that plays

[22] Aquinas (1270/1917, II-II, 30, 3). [23] Smith (1759/1853: I, I, 1. 1).

a central role in his theory is inevitably somewhat *stipulative*—he might have chosen "compassion" or "sympathy" or even "pity" for the term of art he needs. Because of this, we are not concerned with the question of whether he uses "empathy" in ways consistent with common usage. However, we do think his characterization of empathy is neither as clear nor as detailed as one might hope. Much of what Batson says is aimed at contrasting empathy with a different cluster of affective responses to other people's suffering, which Batson calls "personal distress." According to Batson, empathy "includes feeling sympathetic, compassionate, warm, softhearted, tender, and the like, and according to the empathy–altruism hypothesis, it evokes altruistic motivation" (1991: 86). Personal distress, by contrast, is "made up of more self-oriented feelings such as upset, alarm, anxiety, and distress" (ibid.: 117). Elsewhere, he tells us that personal distress "includes feeling anxious, upset, disturbed, distressed, perturbed, and the like, and evokes egoistic motivation to have the distress reduced" (ibid.: 86). While these characterizations may suffice for designing experiments aimed at testing the view that empathy leads to altruistic motivation, they leave a number of important issues unaddressed.

One of these issues is often discussed under the heading of "congruence" or "homology." Sometimes when an observer becomes aware that another person (the "target" as we'll sometimes say) is experiencing an emotion, this awareness can cause a similar emotion in the observer. If, for example, you are aware that Ellen is frightened of the man walking toward her, you may also become frightened of him; if you learn that your best friend is sad because of the death of his beloved dog, this may make you sad as well. In these cases, the emotion evoked in the observer is said to be *congruent* or *homologous* to the emotion of the target. Since Batson describes empathy as a "vicarious emotion that is congruent with but not necessarily identical to the emotion of another" (1991: 86), it is tempting to suppose that he thinks empathic emotions are always at least similar to an emotion the target is experiencing (or similar to what the observer believes the target's emotion to be). But as both Sober and Wilson (1998: 234–5) and Nichols (2004: 32) have noted, *requiring* congruence in the emotion that allegedly gives rise to altruistic motivation may be unwise. Accident victims who are *obviously* unconscious sometimes evoke an "other-oriented emotional reaction" that might be characterized as "compassionate, warm, softhearted, tender, and the like," and people who feel this way are sometimes motivated to help the unconscious victim. But if empathy requires congruence, then the emotion that motivates people to help in these cases *can't* be empathy, since unconscious people aren't experiencing any emotions. We're inclined to give Batson the benefit of the doubt here, and interpret him—charitably, we believe—as holding that empathy is sometimes,

or often, a congruent emotion, but that it need not always be. Presumably personal distress is also sometimes congruent, as when one person's anxiety or alarm evokes anxiety or alarm in an observer. So empathy and personal distress cannot be distinguished by reference to congruence.

Another issue of some importance is whether empathy is always unpleasant or, as psychologists often say, aversive. According to Batson (1991: 87), "empathy felt for someone who is suffering will likely be an unpleasant, aversive emotion," and of course this is just what we would expect if empathy were often a congruent emotion. But as Sober and Wilson (1998: 235) note, we sometimes talk about empathizing with another person's pleasant emotions, and when the term is used in this way, one can have empathic joy as well as empathic sadness. Although Sober and Wilson are certainly right that ordinary language allows us to talk of empathizing with people's positive emotions as well as with their negative emotions, we think Batson is best understood as *stipulating* that, as he uses the term, empathy is a response to the belief that the target is suffering, and that it is typically aversive. Since personal distress is typically—or perhaps always—unpleasant, the distinction between them cannot be drawn by focusing on aversiveness.

That leaves "self-orientedness" vs. "other-orientedness" as the principal dimension on which personal distress and empathy differ. Thus the distinction is doing important theoretical work for Batson, although he does not tell us much about it. We believe that the distinction Batson requires becomes sufficiently clear when operationalized in his experimental work, and we shall not further tarry on the conceptual difficulty. But more conceptual and empirical work aimed at clarifying just what empathy and personal distress are would certainly be welcome.[24]

4.2. *Empathy, Perspective-Taking and Helping Behavior*

In order to put the empathy–altruism hypothesis to empirical test, it is important to have ways of inducing empathy in the laboratory. There is, Batson maintains, a substantial body of literature suggesting that this can indeed be done. For example, Stotland (1969) showed that subjects who were instructed to imagine how a target person felt when undergoing what subjects believed to be a painful medical procedure reported stronger feelings of empathy and showed greater physiological arousal than subjects who were instructed to watch the target person's movements. Krebs (1975) demonstrated that subjects who observe someone *similar to themselves* undergo painful experiences show more physiological arousal, report identifying with the target more strongly, and report feeling worse while waiting for the painful stimulus to begin than do subjects who observe the same painful experiences administered to someone who is not similar to themselves. Additionally, Krebs (1975) found subjects more willing to help at some personal cost when the sufferer was similar to themselves.

There is also evidence that the effects of empathy are focused on the specific distress that evokes it. Stotland's technique for manipulating empathy by instructing subjects to take the perspective of the person in distress was used by Dovidio et al. (1990) to induce empathy for a young woman, with subjects focusing on one of two quite different problems that the young woman faced. When given an opportunity to help the young woman, subjects in whom empathy had been evoked were more likely to help than subjects in a low empathy condition, and the increase in helping was specific to the problem that had evoked the empathy.

On the basis of these and other experiments, Batson concludes that the process of perspective-taking plays a central role in arousing empathy. According to Batson, "adopting the needy person's perspective involves imagining how that person is affected by his or her situation" (1991: 83), and "adopting the needy person's perspective seems to be a *necessary condition* for arousal of empathic emotion" (ibid.: 85, emphasis added). He goes on to assemble a list of ways in which perspective-taking, and thus empathy, can be induced.

> [A] perspective-taking set, that is, a set to imagine how the person in need is affected by his or her situation . . . may be induced by prior experience in similar situations, by instructions, or by a feeling of attachment to the other. In the psychology laboratory perspective taking has often been induced by instructions . . . In the natural stream of behavior also, perspective taking may be the result of instructions, including self-instructions (e.g., "I should walk a mile in his moccasins"), but it is more often the result either of prior similar experience ("I know just how you must feel") or of attachment." (Ibid.: 84)

Figure 5.7 is a sketch of the causal pathways that, on Batson's account, can lead to empathy.

Although we are prepared to believe that Figure 5.7 depicts a number of possible routes to empathy, we are, for two reasons, skeptical about Batson's claim that perspective-taking is a *necessary condition* for arousing empathy. First, while the experimental evidence Batson cites makes it plausible that attachment and similarity to self can indeed lead to empathy, it does not rule out the possibility that these processes bypass perspective-taking and lead *directly* to empathy. Second, we know of no literature that takes on the task of showing that there are no *other* routes to empathy, so the existence of quite different

[24] Both Sober & Wilson (1998: 231–237) and Nichols (2004: ch. 2) have useful discussions, though we think much more remains to be done.

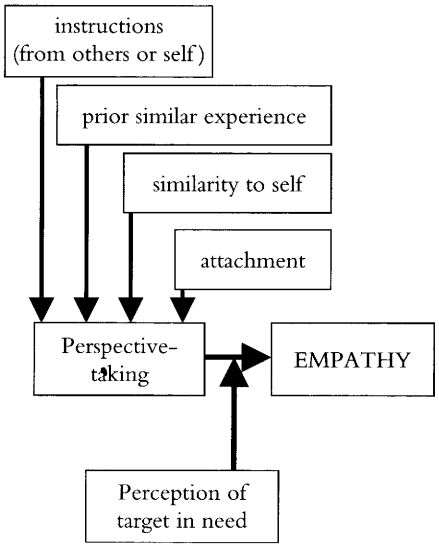

Figure 5.7. Batson's account of the causal pathways that can lead to empathy

causal pathways leading to empathy is largely unexplored. Neither of these reservations poses a major challenge to Batson's project, however, since what he really needs is the claim that—whatever the mechanism may be—the four factors at the top of Figure 5.7 can be used to induce empathy in experimental subjects. And we think that both the pre-existing literature and many of Batson's own experiments provide compelling support for that claim.

With this element in place, Batson's next step is to make the case that empathy leads to helping behavior. Here again, he relies in part on work by others, including the Krebs (1975) and Dovidio et al. (1990) studies cited earlier. Many of Batson's own experiments, some of which we'll describe below, also support the contention that empathy and empathy-engendering experimental manipulations increase the likelihood of helping behavior. Another important source of support for this conclusion is a meta-analysis of a large body of experimental literature by Eisenberg and Miller (1987). On the basis of these and other studies, Batson (1991: 95) concludes that "there is indeed an empathy-helping relationship; feeling empathy for a person in need increases the likelihood of helping to relieve that need."

4.3. *The Empathy–Altruism Hypothesis*

It might be thought that establishing a causal link between empathy and helping behavior would be bad news for egoism. But, as Batson makes clear, the fact

that empathy leads to helping behavior does not resolve the dispute between egoists and altruists, since it does not address the nature of the *motivation* for the helping behavior that empathy evokes. Egoists, of course, do not deny that people engage in helping behavior. Nor need they deny that seeing other people in distress can cause emotions like empathy or that these emotions can lead to helping behavior. The crucial question dividing egoists from altruists is: *how* does the emotion engender helping behavior?

The *empathy–altruism hypothesis* asserts that empathy causes a genuinely altruistic desire to help—an ultimate desire for the well being of the sufferer. It is important to note that the empathy–altruism hypothesis does *not* predict that agents who feel empathy for a target person will *always* help the target; people typically have various and conflicting desires, and not all conflicts are resolved in favor of empathy's urgings. Moreover, even when there are no conflicting desires, it will sometimes be the case that the agent simply does not know how to help. What the empathy–altruism hypothesis claims is that empathy evokes an ultimate desire that the target's distress be reduced. In favorable cases, this ultimate desire, along with the agent's background beliefs, will generate a plan of action. That plan will compete with other plans generated by competing, non-altruistic desires. When the altruistic desire is stronger than the competing desires, the altruistic plan is chosen and the agent engages in helping behavior. It is also important to keep in mind that the empathy–altruism hypothesis does not entail that people who feel little or no empathy will not want to help and will not engage in helping behavior, since an instrumental desire to help can be produced by a variety of processes in which empathy plays no role.

So the empathy–altruism hypothesis offers one account of the way in which empathy can lead to helping behavior. But there is also a variety of egoistic alternatives by which empathy might lead to helping behavior without generating an ultimate desire to help. Perhaps the most obvious of these is that empathy might simply be (or cause) an unpleasant experience, and that people are motivated to help because they believe that helping is the best way to *stop* the unpleasant experience that is caused by someone else's distress. Quite a different family of egoistic possibilities focuses on the rewards to be expected for helping and/or the punishments to be expected for withholding assistance. If people believe that others will sanction them if they fail to help in certain circumstances, or reward them if they do help, and if they believe that the feeling of empathy marks those cases in which social sanctions or rewards are most likely, then we would expect people to be more helpful when they feel empathy, even if their ultimate motivation is purely egoistic. A variation on this theme focuses on rewards or punishments that are self-generated or self-administered. If people believe that helping may make them feel good,

or that failing to help may make them feel bad, and that these feelings will be most likely to occur in cases where they feel empathy, then once again we would expect people who empathize to be more helpful, although their motives may be not at all altruistic.

For more than twenty-five years, Batson and his associates have been systematically exploring these and other options for explaining the link between empathy and helping behavior. Their strategy is to design experiments in which the altruistic explanation, which maintains that empathy leads to an ultimate desire to help, can be compared to one or another *specific* egoistic alternative. If the strategy succeeds, it does so by eliminating plausible egoistic competitors one at a time and by generating a pattern of evidence that is best explained by the empathy–altruism hypothesis. Batson (1991: 174) concludes, albeit tentatively, that the empathy–altruism hypothesis is correct.

> In study after study, with no clear exceptions, we find results conforming to the pattern predicted by the empathy–altruism hypothesis, the hypothesis that empathic emotion evokes altruistic motivation. At present, there is no egoistic explanation for the results of these studies . . . . Pending new evidence or a plausible new egoistic explanation for the existing evidence, the empathy–altruism hypothesis, however improbable, seems to be true.

Reviewing all of these studies would require a very long chapter indeed.[25] Rather than attempt that, we shall take a careful look at some of the best known and most influential experiments aimed at putting altruism to the test in the psychology laboratory. These will, we hope, illustrate both the strengths and the challenges of this approach to the egoism vs. altruism debate.

4.4. *The Empathy–Altruism Hypothesis vs. the Aversive-Arousal Reduction Hypothesis*

Of the various egoistic strategies for explaining helping behavior, among the most compelling is what Batson calls the "aversive-arousal reduction hypothesis." The simplest version of this idea claims that seeing someone in need causes an aversive emotional reaction—something like Batson's *personal distress*—and this leads to a desire to eliminate the aversive emotion. Sometimes the agent will believe that helping is the easiest way to eliminate the aversive emotion, and this will lead to an *instrumental* desire to help, which then leads to helping behavior.[26] However, this simple version of the aversive-arousal

[25] For excellent overviews of this research, see Batson (1991,1998).

[26] This idea, which is widely discussed in the social sciences, has venerable philosophical roots. In *Brief Lives*, written between 1650 and 1695, John Aubrey (1949) describes an occasion on which Thomas Hobbes gave alms to a beggar. Asked why, Hobbes replied that by giving alms to the beggar, he not only relieved the man's distress but he also relieved his own distress at seeing the beggar's distress.

reduction hypothesis cannot explain the strong effect that empathy-inducing factors have on helping behavior (as discussed in Section 4.2). To accommodate that effect, a more sophisticated version of the hypothesis must be constructed. A plausible suggestion is that the distress felt when we see someone in need is significantly greater when we also feel empathy. This increased distress might be explained by the fact that empathy itself is aversive, or it might be because personal distress is increased when we feel empathy, or perhaps both factors play a part. Whatever the cause, the egoist will insist that the increased helping in situations that evoke empathy is due to an ultimate desire to alleviate the increased distress. By contrast, the empathy–altruism hypothesis maintains that when people feel empathy, this evokes an ultimate desire to help, and that, in turn, sometimes leads to genuinely altruistic behavior.

Batson argues that manipulating *difficulty of escape* allows us to compare these two hypotheses experimentally. The central idea is that if a subject is motivated by an ultimate desire to help the target, that desire can be satisfied only by helping. However, if a subject is motivated by a desire to reduce his own distress, that desire can be satisfied either by helping *or* by merely escaping from the distress-inducing situation—for example, by leaving the room so that one is no longer confronted by the needy target. Assuming that subjects do whatever is easier and less costly, the aversive-arousal reduction hypothesis thus predicts that even subjects experiencing empathy will simply leave the needy target, provided escape is made easy enough.

Since the experimental designs are rather complex, it will be helpful to graphically illustrate the claims made by both the empathy–altruism hypothesis and the aversive-arousal reduction hypothesis. If a subject feels little or no empathy for a target, then Figure 5.8 depicts the processes underlying the subject's behavior according to *both* the aversive-arousal reduction hypothesis and the empathy–altruism hypothesis. In this low-empathy situation, personal distress is the only emotional reaction engendered by the perception of the target in need, and both helping and leaving are live options for reducing this distress.

Although the empathy–altruism hypothesis and the aversive-arousal reduction hypothesis agree about the case in which a subject feels no empathy for a target, the two hypotheses differ where a subject feels a significant amount of empathy for a target. Figure 5.9 depicts the processes underlying the subject's behavior according to Batson's version of the empathy–altruism hypothesis.

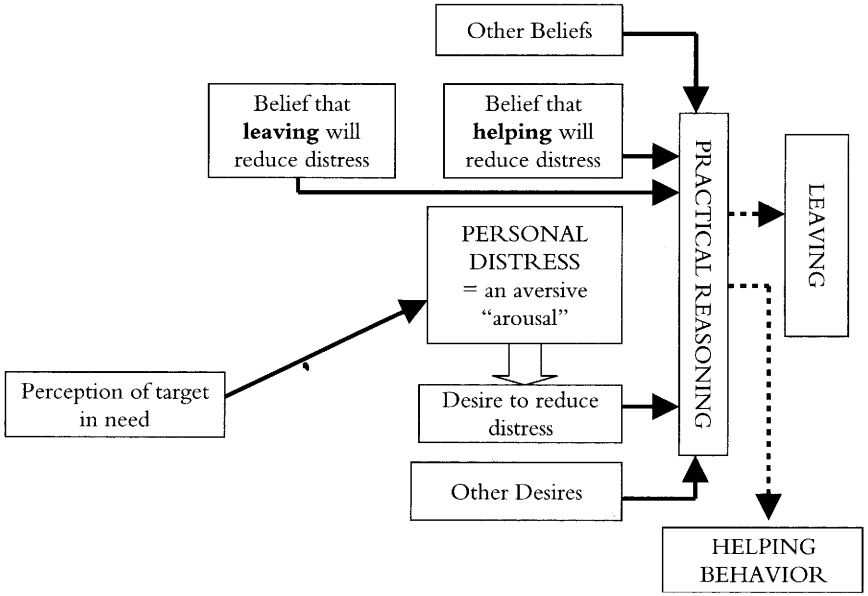

Figure 5.8. Low-empathy subject on both the empathy–altruism hypothesis and the aversive-arousal reduction hypothesis

Here, perception of the target's need leads to empathy and that produces an ultimate desire to help. Since leaving is not an effective strategy for satisfying this desire, helping is the likely behavior—although, of course, the subject might have some *other* desire that is stronger than the ultimate desire to help, so helping is not the only possible outcome.[27] The aversive-arousal reduction hypothesis, by contrast, depicts the processes underlying a high-empathy subject as in Figure 5.10. While the perception of the target in distress leads to empathy in this case too, the empathy simply heightens the subject's personal

[27] Since Batson holds that empathy is typically aversive, one might wonder why the egoistic motivation to reduce this aversive arousal plays no role in motivating helping behavior in Figure 5.9. The answer is that, for strategic reasons, Batson focuses on a "strong" version of the empathy–altruism hypothesis which maintains "not only that empathic emotion evokes altruistic motivation but also that all motivation to help evoked by empathy is altruistic . . . It is easy to imagine a weaker form of the empathy–altruism hypothesis, in which empathic emotion evokes both egoistic and altruistic motivation . . . The reason for presenting the strong form . . . is not because it is logically or psychologically superior; the reason is strategic. The weak form has more overlap with egoistic explanations of the motivation to help evoked by empathy, making it more difficult to differentiate empirically from these egoistic explanations" (1991: 87–88). Although it is not depicted in Figure 5.9, Batson's version of the empathy–altruism hypothesis would presumably also maintain that in many cases perception of a target person in distress would also generate some personal distress even in an agent who strongly empathizes with the target. And when the agent believes that leaving is the easiest way to reduce that personal distress, this will generate some motivation to leave.

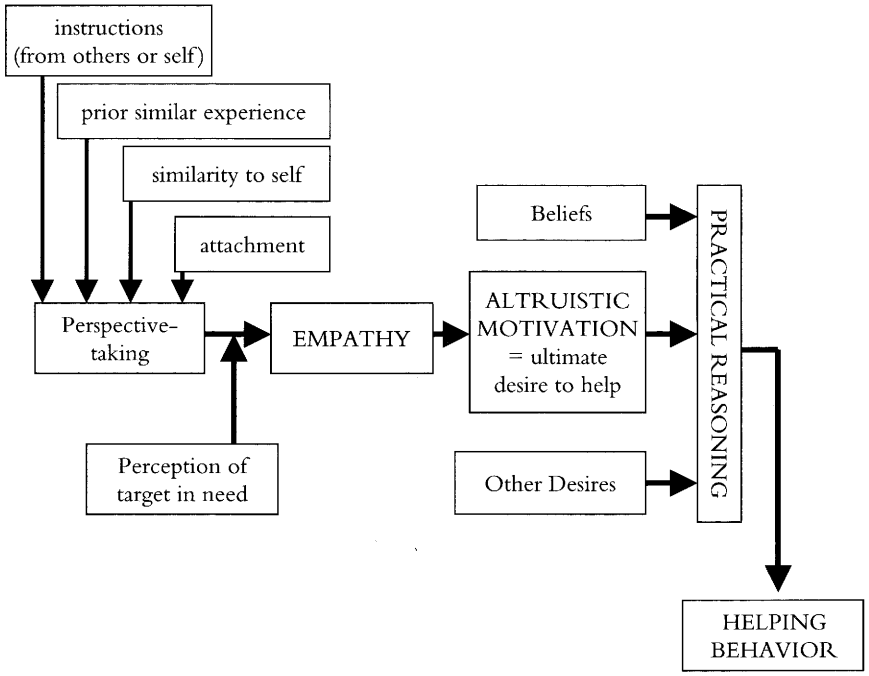

Figure 5.9. High-empathy subject on the empathy–altruism hypothesis

distress and thus strengthens his desire to reduce the distress. Since the subject believes that either helping or leaving will reduce distress, both of these actions are live options, and the subject will select the one that he believes to be easiest.

In designing experiments to compare the empathy–altruism hypothesis with the aversive-arousal reduction hypothesis, Batson must manipulate two distinct variables. To determine whether empathy is playing a role in producing helping behavior, he has to compare the behavior of low-empathy and high-empathy subjects. To determine whether ease of escape has any effect on the likelihood of helping behavior, he must arrange things so that leaving is significantly more costly for some subjects than for others. So there are four experimental conditions: low-empathy subjects where escape is either (1) easy or (2) hard, and high-empathy subjects where leaving is either (3) easy or (4) hard. Batson summarizes what he takes to be the predictions made by the aversive-arousal reduction hypothesis and by the empathy–altruism hypothesis in Tables 5.1 and 5.2 (Batson, 1991: 111). The crucial difference is in the upper-right quadrants, where escape is easy and empathy is high. Under these conditions, Batson maintains, the egoistic aversive-arousal reduction hypothesis predicts a

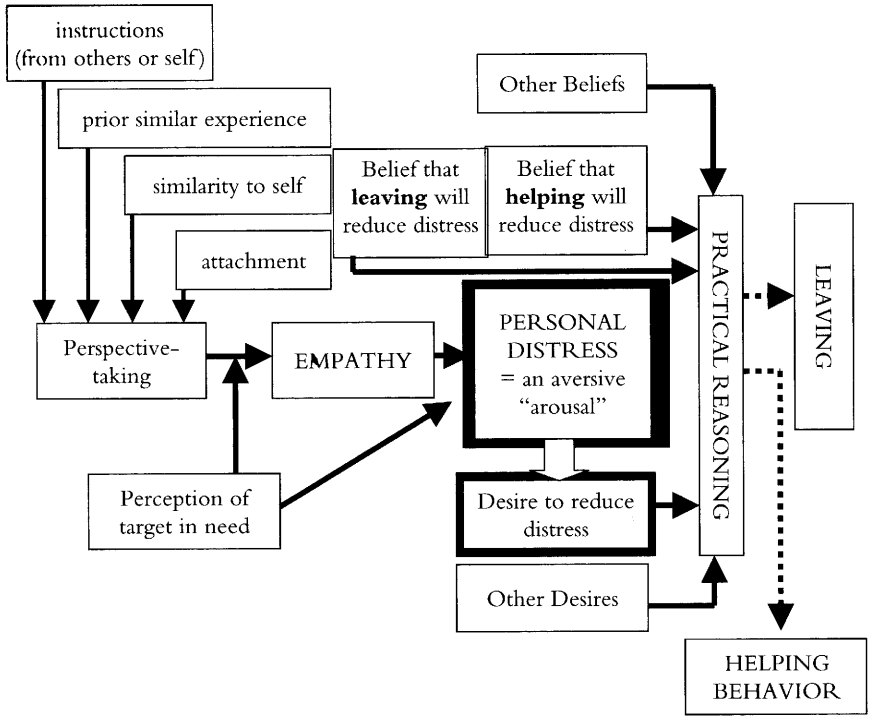

Figure 5.10. High-empathy subject on the aversive-arousal reduction hypothesis

Table 5.1. Predictions from the aversive-arousal reduction hypothesis for the rate of helping in an escape × empathy design

	Empathy	
Escape	Low	High
Easy	Low	**Low**
Difficult	High	High/Very High

low level of helping behavior, while the empathy–altruism hypothesis predicts that the rate of helping will be high.

To test these predictions, Batson and his colleagues conducted a total of six experiments. In one of these (Batson et al., 1981: experiment 1), student subjects were required to watch, via what they believed to be closed-circuit TV, as another student subject, Elaine, attempted to perform a task while receiving electric shocks at random intervals. Observer subjects were told that their task would be to form and report an impression of how Elaine

Table 5.2. Predictions from the empathy–altruism hypothesis for the rate of helping in an escape × empathy design

	Empathy	
Escape	Low	High
Easy	Low	**High**
Difficult	High	High

performs under aversive conditions. Actually, what the subjects were viewing was a videotape. On the tape, Elaine is clearly finding the shocks very uncomfortable, and after her second trial at doing the task, she explains to Martha, the assistant overseeing the experiment, that she is unusually sensitive to mild electric shocks because of a childhood trauma. Martha then suggests that perhaps the observer subject might be willing to help Elaine by taking her place, and the experimenter asks whether the subject is willing to do that. To manipulate ease of escape, some subjects are told that if they decide not to take Elaine's place, they will be required to watch eight additional trials, while other subjects are told that if they decide not to take Elaine's place, they will be free to go, although Elaine will have to endure eight more trials. To manipulate the level of empathy that subjects feel for Elaine, subjects are given a copy of a personal values and interests questionnaire, allegedly filled out by Elaine, in order to help them form an impression of her performance. In the high-empathy condition, Elaine's values and interests are very similar to the subject's (which had been determined in a screening session several weeks before), while in the low-empathy condition, they are very different. The results, given in Table 5.3, clearly exhibit the pattern that Batson thinks is predicted by the empathy–altruism hypothesis, not the pattern that Batson thinks is predicted by the aversive-arousal reduction hypothesis.

In four additional experiments, Batson and his associates used a total of four different techniques to create the low- and high-empathy conditions, two

Table 5.3. Proportion of subjects agreeing to take shocks for Elaine (Batson et al., 1981: experiment 1)

	Empathy Condition	
Escape Condition	Low (Dissimilar Victim)	High (Similar Victim)
Easy	.18	.91
Difficult	.64	.82

techniques for manipulating ease of escape, and two different need situations.[28] The results in all of these experiments exhibited the same pattern. Intriguingly, in a sixth experiment, Batson attempted to break the pattern by telling the subjects that the shock level they would have to endure was the highest of four options, "clearly painful but not harmful." They reasoned that, in these circumstances, even if high-empathy subjects had an ultimate desire to help, this desire might well be overridden by the desire to avoid a series of very painful shocks. As expected, the pattern of results in this experiment fit the pattern in Table 5.1.

These are, we think, truly impressive findings. Over and over again, in well designed and carefully conducted experiments, Batson and his associates have produced results that are clearly compatible with what Batson has argued are the predictions of the empathy–altruism hypothesis, as set out in Table 5.2, and clearly incompatible with the predictions of the aversive-arousal reduction hypothesis, as set out in Table 5.1. Even the "clearly painful shock" experiment, which produced results in the pattern of Table 5.1, is comfortably compatible with the empathy–altruism hypothesis since, as we noted in our discussion of Figure 5.9, the empathy–altruism hypothesis allows that high-empathy subjects may have desires that are stronger than their ultimate desire to help the target, and the desire to avoid a painful electric shock is a very plausible candidate.

There is, however, a problem to be overcome before we conclude that the aversive-arousal reduction hypothesis cannot explain the findings that Batson has reported. In arguing that Table 5.1 reflects the predictions made by the aversive-arousal reduction hypothesis, Batson assumes that escape will alleviate personal distress (and the aversive component of empathy) in both low- and high-empathy situations, and that subjects *believe* this, although the belief, along with many other mental states and processes posited in Figures 5.8, 5.9, and 5.10, may not be readily available to introspection. We might call this the *out of sight, out of mind* assumption.[29] But, elaborating on an idea suggested by Hoffman (1991) and Hornstein (1991), an advocate of egoism might propose that although subjects do believe this when they have little empathy for the target, *they do not believe it when they have high empathy for the target*. Perhaps high-empathy subjects believe that if they leave the scene they will continue to be troubled by the thought or memory of the distressed target and thus that

[28] The experiments are reported in Batson et al. (1981), Toi & Batson (1982), and Batson et al. (1983).

[29] As Batson himself remarks, "[T]he old adage, 'Out of sight, out of mind,' reminds us that physical escape often permits psychological escape as well" (1991: 80).

physical escape will not lead to psychological escape. Indeed, in cases where empathy is strong and is evoked by attachment, this is just what common sense would lead us to expect. (Do you suppose that if you abandoned your mother when she was in grave distress, you would no longer be troubled by the knowledge of her plight?) But if the high-empathy subjects in Batson's experiments believe that they will continue to be plagued by distressing thoughts about the target even after they depart, then the egoistic aversive-arousal reduction hypothesis predicts that these subjects will be inclined to help in both the easy physical escape and the difficult physical escape conditions, since helping is the only strategy they believe will be effective for reducing the aversive arousal.[30] So neither the findings reported in Table 5.3 nor the results of any of Batson's other experiments would give us a reason to prefer the empathy–altruism hypothesis over the aversive-arousal reduction hypothesis, because both hypotheses would make the same predictions.

Is it the case that high-empathy subjects in experiments like Batson's believe that unless they help they will continue to think about the target and thus continue to feel distress, and that this belief leads to helping because it generates an egoistic instrumental desire to help? This is, of course, an empirical question, and until recently there was little evidence bearing on it. But a cleverly designed experiment by Stocks and his associates (Stocks et al., 2009) suggests that, in situations like those used in Batson's experiments, a belief that they will continue to think about the target does *not* play a significant role in causing the helping behavior in high-empathy subjects. The first phase of the experiment was a "psychological escape" manipulation. Half the subjects were told that they would soon be participating in a "deleting memories" training session that would permanently delete their memories of an audiotaped news segment that they were about to hear. The remaining subjects were told that they would soon be participating in a "saving memories" training session designed to permanently enhance the memories of the news segment they were about to hear. Then, using stimulus materials that we shall see in various of Batson's experiments, the experimenters played subjects a fictional college radio news segment about the plight of a fellow student, Katie Banks, whose parents have recently been killed in an automobile accident, leaving her to care for her younger brother and sister. In the interview, Katie mentions that she has begun a fundraising campaign to raise money for her college tuition and for living expenses for her siblings. If she is not successful, she

[30] The point emerges clearly in Figure 5.10. If the *Belief that leaving will reduce distress* is eliminated, then LEAVING is no longer a way to satisfy the *Desire to reduce distress*. So HELPING BEHAVIOR is the only way to satisfy this desire.

Table 5.4. Proportion of subjects agreeing to help Katie Banks (Stocks et al., under review: experiment 1)

	Empathy	
Psychological Escape	Low	High
Easy—memory deleted	.08	.67
Difficult—memory enhanced	.42	.58

will be forced to drop out of school and put her brother and sister up for adoption. Empathy for Katie was manipulated by using the Stotland-inspired technique—instructing some participants to imagine how Katie felt and others to try to remain as objective and detached as possible. After hearing the tape, subjects completed two questionnaires, one designed to assess the success of the empathy manipulation, the other designed to test the effectiveness of the psychological escape manipulation. Both manipulations were successful: crucially, subjects reported being quite confident that the memory training session would enhance or delete their memory of the Katie Banks interview they had just heard. Finally, subjects were given an unexpected opportunity to help Katie with her child care and home maintenance chores.

Stocks and his associates reasoned that if high-empathy subjects in the Batson experiments recounted earlier are egoists who help because they believe that they will continue to have distressing thoughts about the victim, then in this experiment high-empathy subjects who believed their memories of Katie would be enhanced by the training session would be highly motivated to help, while subjects who believed that their memories of Katie would soon be deleted would have little motivation to help. If, by contrast, empathy generates altruistic motivation, there should be little difference between those high-empathy subjects who believe their memories of Katie will be enhanced and those who believe that their memories of her will soon be deleted. The results, shown in Table 5.4, provide impressive confirmation of the prediction based on the empathy–altruism hypothesis.[31]

We believe that Batson's work on the aversive-arousal reduction hypothesis, buttressed by the Stocks et al. finding, is a major advance in the egoism vs. altruism debate. No thoughtful observer would conclude that these experiments

[31] Both the aversive-arousal reduction hypothesis and the empathy–altruism hypothesis predict that low-empathy subjects will behave egoistically, and thus that they will be less inclined to help when they believe that their memories of Katie will be deleted than when they believe their memories will linger. The fact that this prediction is confirmed is a further indication that the psychological escape manipulation was successful.

show that altruism is true, since, as Batson himself emphasizes, the aversive-arousal reduction hypothesis is just one among many egoistic alternatives to the empathy–altruism hypothesis, although it has been one of the most popular egoistic strategies for explaining helping behavior. But the experimental findings strongly suggest that in situations like those that Batson has studied, the empathy–altruism hypothesis offers a much better explanation of the subjects' behavior than the aversive-arousal reduction hypothesis (Batson, 1991: 127).

4.5. *The Empathy–Altruism Hypothesis vs. The Empathy-Specific Punishment Hypothesis*

As noted earlier, thinkers in the egoist tradition have proposed many alternatives to the hypothesis that empathy engenders genuine altruism. Although aversive-arousal reduction may be the most popular of these, another familiar proposal maintains that people engage in helping behavior because they fear they will be punished if they do not help. On one version of this view, the punishments are socially administered. If I don't help, the agent worries, people will think badly of me, and this will have negative effects on how they treat me. On another version, which to our mind is both more plausible and more difficult to assess experimentally, the punishments that people are worried about are self-administered. If she doesn't help, the agent believes, she will suffer the pangs of guilt, or shame, or some other aversive emotion. As they stand, neither of these egoist accounts can explain the fact that empathy increases the likelihood of helping, but more sophisticated versions are easy to construct.[32] They need only add the assumption that people think either social sanctions or self-administered sanctions for not helping are more likely when the target engenders empathy. We'll take up the social and self-administered variants in turn. We believe that currently available evidence supports the conclusion that the social version is incorrect, but we shall argue that the evidence regarding the self-administered version is inconclusive.

Following Batson, let's call the social variant of this hypothesis (the one that maintains that subjects believe socially administered sanctions to be more likely when the target engenders empathy) the *socially administered empathy-specific punishment hypothesis*. To test it against the empathy–altruism hypothesis, Batson and his associates designed an experiment in which they manipulated both the level of empathy that the subject felt for the target and the likelihood that anyone would know whether or not the subject had opted to help a

[32] The problem is similar to the one confronting the simple version of the aversive-arousal reduction hypothesis discussed at the beginning of 4.4, as is the solution.

Table 5.5. Predictions about the amount of helping on the socially administered empathy-specific punishment hypothesis

Would the helping choice be private or public?	Empathy	
	Low	High
Public	Low	High
Private	Low	**Low**

Table 5.6. Predictions about the amount of helping on the empathy–altruism hypothesis

Would the helping choice be private or public?	Empathy	
	Low	High
Public	Low	High
Private	Low	**High**

person in need. Obviously, if your decision is secret, you need have no fear of social sanctions. Thus the socially administered empathy-specific punishment hypothesis predicts that high-empathy subjects will be more likely to help when others will know whether or not they decide to help. On the empathy–altruism hypothesis, by contrast, high-empathy subjects are motivated by an ultimate desire to help, and thus their helping levels should be high whether or not others will know if they decided not to help. In the low-empathy condition, both hypotheses predict that levels of helping will be low. These predictions are summarized in Tables 5.5 and 5.6.

Subjects in the experiment were told that they were participating in an impression-formation study, and that they had been randomly assigned to the role of "listener" who would read confidential notes from another student participant, the "communicator." The listeners were given two notes from a fictional communicator, "Janet Arnold," an out-of-state freshman who reveals that she is very lonely and hoping to find a friend. After completing a form on which subjects indicated their impressions of Janet and their emotional responses to her note, they were presented with an unexpected opportunity to befriend Janet by volunteering to take part in a study of long-term relationships which would involve an initial meeting with Janet and periodic phone interviews about the development of their friendship. Participants who were willing to help were asked how much time they were willing to spend with Janet during the next month.

To manipulate empathy, Batson and his associates used the Stotland technique. Subjects assigned to the low-empathy condition were instructed to "Read the information in the communicator's notes carefully, trying to be as objective as possible. Try to focus on the techniques used to get the communicator's message across." Subjects assigned to the high-empathy condition were instructed to "Read the information in the communicator's notes carefully, trying to take the perspective of the individual writing the notes. Try to imagine how this individual feels about what he or she writes" (1991: 132).

Four steps were taken to manipulate the perceived potential for negative social evaluation. In the *low-potential* condition, where the subjects' choices were private (Batson 1991: 132):

(1) Subjects were told that they would be the only person ever to read the communicator's notes, and the notes were delivered in a sealed envelope.
(2) Subjects were assured that the questionnaires reporting their reactions to Janet's note would be anonymous.
(3) Subjects' responses to the invitation to participate in the long-term relationship study were returned to the professor in a sealed envelope, so no one would know if they did not want to help Janet. (Since they believed the professor had not read Janet's note, they could be expected to surmise that he would not know that Janet needed a friend.)
(4) In the letter inviting them to participate in the relationship study, subjects were told that Janet would be contacted "only if you want a meeting; if you do not, then we will not contact the communicator and he or she will not know that you were ever asked about actually meeting."

In the *high-potential* condition where the subjects' choices were public (Batson, 1991: 133):

(1) Subjects were told that the experimenter would (a) read the communicator's notes, (b) collect the subject's completed questionnaire, and (c) compare the content of the notes with the content of the questionnaire.
(2) No envelope was provided for their response to the invitation to participate in the long-term relationship study, so subjects could infer that the experimenter would know about Janet's need for a friend, about their emotional reaction to Janet, and would also know whether or not they had agreed to help Janet.
(3) Subjects were told that Janet would know if they decided not to help.

As Batson states, or perhaps understates, procedures in the high-potential (public) condition "were designed to highlight the possibility of negative social evaluation if the participant decided not to volunteer to spend time with Janet" (1991: 133).

Table 5.7. Amount of help offered Janet (Fultz et al., 1986: study 2)

Would the helping choice be private or public?	Empathy	
	Low	High
Public	.67	1.71
Private	1.29	2.44

The results, given in Table 5.7, indicate that the socially administered empathy-specific punishment hypothesis did not fare well. On the basis of this experiment and a similar experiment in which empathy for Janet was not manipulated but was measured by self-report, Batson concludes that the socially administered empathy-specific punishment hypothesis is not consistent with the experimental findings.

> Contrary to what the social-evaluation version of the empathy-specific punishment hypothesis predicted, eliminating anticipated negative social evaluation in these two studies did not eliminate the empathy-helping relationship. Rather than high empathy leading to more help only under high social evaluation, it led to more helping under both low and high social evaluation. This pattern of results is not consistent with what would be expected if empathically aroused individuals are egoistically motivated to avoid looking bad in the eyes of others; it is quite consistent with what would be expected if empathy evokes altruistic motivation to reduce the victim's need. (1991: 134)

Although two experiments hardly make a conclusive case, we are inclined to agree with Batson that these studies make the socially administered empathy-specific punishment hypothesis look significantly less plausible than the empathy–altruism hypothesis. High-empathy subjects were more likely to help *whether or not* they could expect their behavior to be socially scrutinized. So another popular egoist hypothesis has been dealt a serious blow. At least in some circumstances, empathy appears to facilitate helping independently of the threat of social sanction.[33]

There is, however, another version of the empathy-specific punishment hypothesis that must be considered, and we are less sanguine about Batson's

[33] These studies do not address a variant of the socially administered empathy-specific punishment hypothesis that might be called the "*divinely* administered empathy-specific punishment hypothesis." It is very plausible that subjects in the private low potential for social evaluation condition believed that no ordinary person would know if they declined to help someone in need. But these subjects might believe that *God* would know, that *He* would punish them for their failure to help, and that God's punishment would be particularly severe when they failed to help someone for whom they felt empathy. To the best of our knowledge, the divinely administered empathy-specific punishment hypothesis is a version of egoism that has not been explored empirically.

attempts to defend the empathy–altruism hypothesis against this version. According to the *self-administered* empathy-specific punishment hypothesis, people are motivated to help because they believe that if they don't help they will experience some negative self-regarding emotion, such as guilt or shame. As Batson (1991: 98) explicates the view, "we learn through socialization that empathy carries with it a special obligation to help and, as a result, an extra dose of self-administered shame or guilt if we do not. When we feel empathy we think of the impending additional self-punishments and help in order to avoid them . . . To test whether self-punishment underlies the empathy-helping relationship," Batson (1991: 134–135) notes, "high-empathy individuals must anticipate being able to escape . . . from negative self-evaluation." But, one might wonder, how is that possible? Here is the crucial passage in which Batson addresses the question and sets out his strategy.

> [I]f expectations of self-punishment have been internalized to the degree that they are automatic and invariant across all helping situations, then providing escape seems impossible. It seems unlikely, however, that many people—if any—have internalized procedures for self-punishment to such a degree. Even those who reflexively slap themselves with guilt and self-recrimination whenever they do wrong are likely to be sensitive to situational cues in determining when they have done wrong . . . And given the discomfort produced by guilt and self-recrimination, one suspects that most people will not reflexively self-punish but will, if possible, overlook their failures to help. They will dole out self-punishments only in situations in which such failures are salient and inescapable.
>
> If there is this kind of leeway in interpreting failure to help as unjustified and hence deserving of self-punishment, then expectation of self-punishment may be reduced by providing some individuals with information that justifies not helping in some particular situation. The justifying information probably cannot be provided directly by telling individuals not to feel guilty about not helping. Calling direct attention to the failure may have the reverse effect; it may highlight the associated punishments. The information needs to be provided in a more subtle, indirect way. (1991: 135)

Before considering how Batson and his associates designed experiments that attempt to implement this strategy, we want to emphasize a subtle but very important concern about the passage we've just quoted. In that passage, Batson slips back and forth between claims about people's *expectations* about self-punishment and their internalized procedures for administering it. The latter are claims about what people will actually feel if they fail to help in various circumstances, while the former are claims about what people *believe* they will feel. The distinction is an important one because in the debate between egoists and altruists it is the *beliefs* that are crucial. To see this, recall what is at issue. Both egoists and altruists agree that people sometimes desire to help

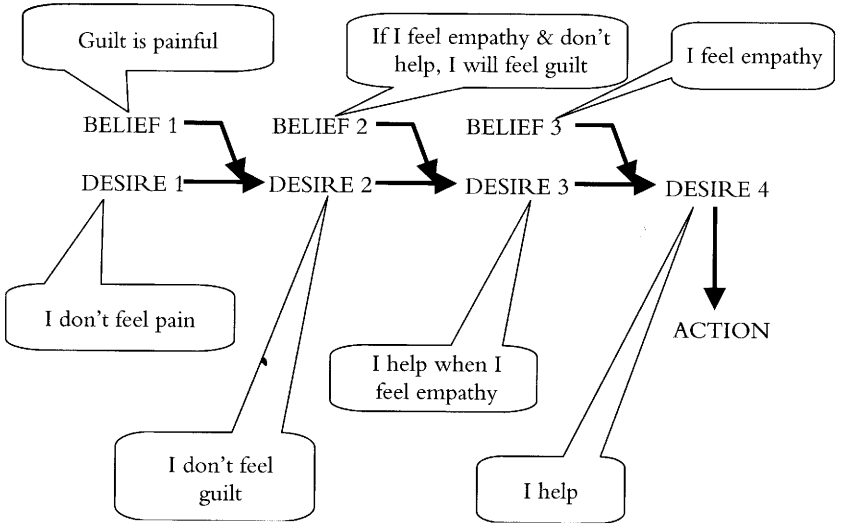

Figure 5.11. The self-administered empathy-specific punishment account of the practical reasoning leading to the instrumental desire to help

others, but egoists insist that these desires are always instrumental desires, while altruists maintain that, sometimes at least, these desires are not instrumental but ultimate. As we saw in Section 1, instrumental desires are desires that are produced or sustained via practical reasoning, a process in which desires and beliefs lead to new desires. So an egoist who advocates the self-administered empathy-specific punishment hypothesis would maintain that the desire to help is generated by an episode of practical reasoning something like the one in Figure 5.11. In that diagram, BELIEF 2 (*If I feel empathy and don't help, then I will feel guilty*) plays a central role, since it is this belief that is supposed to capture the agent's expectations about self-punishment. That belief may or may not be accurate—perhaps the agent won't actually feel guilty. But the accuracy of the belief is not relevant to the debate between egoists and altruists. What is crucial is just that the subject has some belief that, together with her desire not to feel guilt (DESIRE 2), will generate DESIRE 3 (*I help when I feel empathy*). The importance of all this will emerge as we review Batson's experiments and how he interprets them.

In designing experiments to test the empathy–altruism hypothesis against the self-administered empathy-specific punishment hypothesis, Batson's strategy is to provide some subjects with a justification for not helping. By doing this, Batson's goal is to alter his subjects' beliefs; he expects that subjects who have been given a justification for not helping will be less likely to believe that they

will feel guilty if they do not help. In one experiment, subjects were asked to listen to a tape of what they were told was a pilot broadcast of a campus radio show. The show is an interview with Katie Banks, whom we met earlier in our discussion of the Stocks et al. experiment. She is a student on their campus struggling to stay in school and keep her family together after her parents are killed in an automobile accident. To manipulate levels of empathy, the investigators again used the Stotland procedure. After listening to the tape, subjects are given an opportunity to pledge time to help Katie. To manipulate the justification that subjects have for not helping, some are given a sign-up sheet on which five of seven previous "participants" in the study have agreed to help, while others are given a sign-up sheet on which only two of the previous seven participants have agreed to help. By getting the second group to believe that most previous participants had not been willing to help, Batson and his associates intend to be providing just the sort of "subtle, indirect" justification for not helping that their strategy requires.[34]

According to Batson, in the low-empathy condition, both the empathy–altruism hypothesis and the self-administered empathy-specific punishment hypothesis will make the same prediction: helping will be high in the low-justification condition (where subjects have little justification for not helping), and low in the high-justification condition (where subjects have a lot of justification for not helping).[35] In the high-empathy condition, however, Batson maintains that the two hypotheses make different predictions. Empathy–altruism claims that empathy will lead to an ultimate desire to help, and thus it predicts that helping should be high whether or not the subject has good justification for not helping. Self-administered empathy-specific punishment, on the other hand, predicts that having or lacking a justification will affect the likelihood of helping. Helping will be high when there is little justification for not helping, but low when there is good justification for not helping. This is because in the former case subjects will believe they will feel very guilty if they do not help, but in the latter case they do not believe this.

The results in the Katie Banks experiment are given in Table 5.8. They conform, quite dramatically, to the "three high and one low" empathy–altruism

[34] One might worry, here, that the sign-up-sheet manipulation could have just the opposite effect. If lots of other people have already signed up to help Katie, some subjects might come to believe that they are justified in *not* helping, since she already has plenty of help.

[35] In order to derive this prediction for the low-justification condition, Batson must assume that, even without empathy, subjects "should be egoistically motivated to avoid general shame and guilt associated with a failure to help" (1991:136). So, as Batson is interpreting the self-administered empathy-specific punishment hypothesis, it maintains that even low-empathy individuals believe they will feel *some* guilt if they fail to help without justification, and high-empathy individuals believe they will feel significantly more guilt if they fail to help without justification.

Table 5.8. Proportion of participants volunteering to help Katie Banks (Batson et al., 1988: study 3)

	Empathy Condition	
Justification condition	Low	High
Low justification for not helping	.55	.70
High justification for not helping	.15	.60

hypothesis prediction. Moreover, "there was no evidence of the significant effect of the justification manipulation in the high-empathy condition predicted by the self-punishment version of the empathy-specific punishment hypothesis" (1991: 138). In a pair of additional experiments, Batson and his associates varied the helping opportunity, the justification manipulation and the empathy manipulation. In those experiments, too, "the pattern of helping was very much as predicted by the empathy–altruism hypothesis" (ibid.: 140). Once again, it appears there is good reason to prefer the empathy–altruism hypothesis over an egoistic empathy-specific punishment alternative. But we are not convinced.

Our concerns are not focused on the details of Batson's experiments, but on his account of what the self-administered empathy-specific punishment hypothesis predicts. To make the point, let's return to the practical reasoning diagram in Figure 5.11. In that diagram, BELIEF 2 portrays the agent believing: *If I feel empathy and don't help, I will feel guilt*. But if Batson is right, that account of BELIEF 2 needs to be modified, since if the agent really believed that, then the justification manipulation would have no effect. In order for the justification manipulation to disrupt the process portrayed in Figure 5.11, BELIEF 2 would have to be something like: *If I feel empathy & don't help, I will feel guilt UNLESS there is justification for not helping*. That belief, along with DESIRE 2 (*I don't feel guilt*), would lead to DESIRE 3 with the content: *I help when I feel empathy UNLESS there is justification for not helping*, and that is just what Batson needs to derive his predictions about what agents are likely to do in high-empathy situations, if the self-administered empathy-specific punishment hypothesis is correct. Suppose, however, that instead of this candidate for BELIEF 2, what agents actually believe is: *If I feel empathy & don't help, I will feel guilt EVEN IF there is justification for not helping*. In that case, DESIRE 3 would be: *I help when I feel empathy EVEN IF there is justification for not helping*, and the self-administered empathy-specific punishment hypothesis would predict high levels of helping in high-empathy subjects whether or not there is justification for not helping. Since this is just what the empathy–altruism hypothesis predicts, Batson's experiments would not be able to distinguish between the two hypotheses. So

clearly the content of agents' beliefs about the link between empathy and guilt is playing a crucial role in Batson's argument.

What grounds does Batson offer for assuming that subjects believe: *If I feel empathy & don't help, I will feel guilt UNLESS there is justification for not helping*, rather than: *If I feel empathy & don't help, I will feel guilt EVEN IF there is justification for not helping*? As best we can tell, the long passage we quoted earlier is his only attempt to justify the assumption. But that passage offers no evidence for the claims it makes about people's "internalized procedures for self-punishment." Moreover, even if Batson's speculations about those procedures are true, it would not justify the claim he really needs, namely that people have accurate *beliefs* about these internalized procedures. The prediction that Batson derives from the egoistic hypothesis requires assuming that people typically have a specific sort of belief about what they will feel if they fail to help when there is justification for not helping. That is an empirical assumption, of course, but it is one for which Batson offers no evidence. Batson might be able to protect his argument from our critique by showing that subjects really do have the belief that his prediction assumes. But until that is done, the experiments we have discussed give us no good reason to reject the self-administered empathy-specific punishment hypothesis.

The findings we have discussed so far are not, however, the only ones that Batson relies on in his critique of the self-administered empathy-specific punishment hypothesis. In another version of the Katie Banks experiment, Batson and his associates attempted to determine *what participants were thinking* when they made their decisions about whether to offer help. Here is Batson's explanation of the motivation for this experiment:

> The empathy-specific punishment hypothesis and the empathy–altruism hypothesis each postulate a different goal for the helping associated with feeling empathy: The goal is avoiding punishment in the former; the goal is relieving the victim's need for the latter. Each hypothesis assumes that the empathically aroused individual, when deciding to help, has one of these goals in mind. If this is true, then cognitions relevant to one of these goals should be associated with empathy-induced helping. (1991: 143)

So if we can find some way to determine what empathically engaged individuals are thinking about, Batson argues, we may be able to provide evidence for one or another of these hypotheses. Simply asking subjects what they were thinking about is methodologically problematic, first because the relevant thoughts might not have been fully conscious, and second because subjects might be unwilling, or otherwise unable, to provide accurate self-reports. But, as it happens, the well-known Stroop procedure provides a way of determining what people are thinking about without directly asking them. In the Stroop

procedure, subjects are shown words typed in various colors. Their task is to name the color of the type, and to do so as quickly as possible. Previous research had shown that the time taken to respond for a particular word (the "latency" in psychologists' jargon) will increase whenever a subject has been thinking about something related to that word (Geller & Shaver, 1976).

In the experiment, which for the most part repeated the Katie Banks procedure, the experimenter paused after telling subjects that they would be given an opportunity to sign up to help Katie, and administered a Stroop test. Some of the words used, like DUTY, GUILT, SHAME, SHOULD, were taken to be "punishment-relevant"; other words, including HOPE, CHILD, NEEDY, FRIEND, were classified as "victim-relevant"; and still other words, LEFT, RAPID, LARGE, BREATH, were classified as neutral. Here is Batson's account of the results and his interpretation of them:

> [T]he only positive association in the high-empathy condition was a correlation between helping and color-naming latency for the victim-relevant words . . . This was the correlation predicted by the empathy–altruism hypothesis. Contrary to the prediction of the empathy-specific punishment hypothesis, there was no evidence of a positive correlation in the high-empathy condition between helping and color-naming latency for the punishment-relevant words.
>
> In the low-empathy condition, in which empathic feelings had not been explicitly aroused, there was not a positive correlation between helping and latency for the victim-relevant words. This finding suggests that the positive correlation for victim-relevant words in the high-empathy condition was not due to some general characteristic of these words or their association with helping. The relationship seemed to be empathy specific. (1991: 147)

Although this is an ingenious experiment with intriguing results, we are again skeptical that the results favor one hypothesis over the other. To make the case for our skepticism, we'll argue that the self-administered empathy-specific punishment hypothesis can be plausibly construed in a way that it *does not* predict, of the agent, that she will be thinking punishment-relevant thoughts. Instead, it predicts exactly what Batson found: the agent thinks only victim-relevant thoughts.

Let's begin by returning to the quotation, two paragraphs back, in which Batson explains the motivation for the experiment. According to Batson, "The empathy-specific punishment hypothesis and the empathy–altruism hypothesis each postulate a different goal for the helping associated with feeling empathy: The goal is avoiding punishment in the former; the goal is relieving the victim's need for the latter" (1991: 143). It is important to see that this claim is plausible only if "goal" is understood to mean *ultimate desire*, and is clearly false if "goal" refers to any desire that may play a role in the process leading to helping behavior. On the empathy-specific punishment hypothesis, although the ultimate desire is avoiding punishment (or the pain that punishment engenders), this leads to an *instrumental* desire to relieve the victim's need. So it is misleading for Batson to claim that "each hypothesis assumes that the empathically aroused individual, when deciding to help, has one of these goals in mind" (1991: 143), since the empathy-specific punishment hypothesis maintains that the empathically aroused individual has *both* desires in mind—one as an ultimate desire and the other as in instrumental desire. This suggests that the empathy-specific punishment hypothesis should predict that high-empathy subjects have a longer latency for *both* punishment-relevant words and victim-relevant words.

If this is right, then we've located a slip in what Batson claims about the prediction of the empathy-specific punishment hypothesis. However, it might be thought that this does no serious harm to Batson's case, since the experimental results also contradict this new prediction. On the new prediction, high-empathy subjects should be thinking both punishment-relevant and victim-relevant thoughts. But they aren't. They are thinking only victim-relevant thoughts, which is what the empathy–altruism hypothesis predicts.

This is not the end of the story, however. For we believe that on one very natural interpretation of the empathy-specific punishment hypothesis, it will *not* predict that agents think punishment-relevant thoughts. To set out this interpretation, we'll need to introduce the idea of a *long-standing instrumental desire*. Consider what happens when you get your monthly electric bill. Typically, we'll assume, you pay it. Why? Well, because you have a desire to pay it. This is, to be sure, an instrumental desire, not an ultimate desire. You want to pay your electric bill because you believe that if you don't, they will turn off your electricity, and that would lead to lots of other unpleasant consequences which you want to avoid. The lights would go out; the heat or air conditioning would go off; your computer would stop working; it would be a real pain in the neck. But are any of these consequences on your mind when you reach for the checkbook to pay the electricity bill? Quite typically, we think, the answer is no. Rather, what happens is that you have a *long-standing desire* to pay the electric bill on time whenever it comes. This is an instrumental desire; there is nothing intrinsically desirable about paying the bill. So, like other instrumental desires, it was formed via a process of practical reasoning. But that was a long time ago, and there is no need to revisit that process every time you pay your bill. Most people, we think, have *lots* of desires like that. They are enduring desires that were formed long ago via practical reasoning. When the circumstances are appropriate, they are activated, they generate further instrumental desires, and ultimately they lead to action. And

all of this happens without either consciously or unconsciously revisiting the practical reasoning that led to the formation of the long-standing desire.

Let's return, now, to Figure 5.11, which sketches the motivational structure underlying helping behavior on one version of the empathy-specific punishment hypothesis. On the interpretation of the hypothesis that we are proposing, DESIRE 3 is a long-standing instrumental desire. It was originally formed via the egoistic process of practical reasoning sketched on the left side of Figure 5.11. But there is no need to repeat that process every time the desire is activated, any more than there is a need to reflect on the lights going out every time you pay your electric bill. On this interpretation of the empathy-specific punishment hypothesis, the agent will not activate thoughts about guilt and punishment when she decides to help. Instead, the only thoughts that need be activated are thoughts about the victim and how to help her. So, on this interpretation, the results of Batson's Stroop experiment are just what the empathy-specific punishment hypothesis would predict. Thus the experiment gives us no reason to prefer empathy–altruism over empathy-specific punishment.

In this section we have looked at two versions of the egoistic empathy-specific punishment hypothesis. We've argued that Batson's work poses a serious challenge to the version on which the punishment is delivered by others. But Batson's experiments do not make a convincing case against the version on which the punishment is *self-inflicted*, via guilt or some other aversive emotion. To address the problems we've elaborated, Batson needs to provide more convincing evidence about the beliefs that subjects invoke when they are making their decision about helping Katie Banks, and about the processes that generate and sustain the desires involved in that decision. This is a tall order, since it is no easy matter to get persuasive evidence about either of these. The need for such evidence makes it clear how challenging empirical work in this area can be. But, as illustrated by Batson's own work, as well as the Stocks et al. study discussed in the previous section, cleverly designed experiments can go a long way toward resolving issues that at first appear intractable. So the gap in Batson's case against the empathy-specific punishment hypothesis certainly gives us no reason to be skeptical about the experimental approach to the egoism vs. altruism debate.

4.6. *The Empathy–Altruism Hypothesis vs. The Empathy-Specific Reward Hypothesis*

In the previous section our focus was on the venerable idea that helping is motivated by fear of punishment. In this section, we'll take a brief look at the

equally venerable idea that helping is motivated by the expectation of reward. Like punishment, reward can come from two sources: others can reward us in various ways for our helpful actions, or we can reward ourselves—helping others can make us feel good. But just as in the case of punishment, the simple theory that people help others because they believe they will be rewarded offers no explanation for the fact that empathy increases helping behavior. To remedy this problem, the egoist can propose that "helping is especially rewarding when the helper feels empathy for the person in need" (Batson, 1991: 97). This is the view that Batson calls *the empathy-specific reward hypothesis*. Although the idea can be spelled out in a variety of ways, we'll begin with the version that claims "that we learn through socialization that *additional rewards* follow helping someone for whom we feel empathy; these rewards most often take the form of extra praise from others or a special feeling of pride in ourselves. When we feel empathy, we think of these additional rewards, and we help in order to get them" (ibid.).

To motivate an ingenious experiment designed to test this hypothesis against the empathy–altruism hypothesis, Batson notes that it is only one's own helping, or attempts to help, that make one eligible for the rewards that helping engenders; if someone else helps the target before you get around to it, the rewards are not forthcoming. This is plausible both in the case where the rewards are provided by others and in the case where the rewards are provided by our own self-generated feeling of pride. For surely we don't typically expect others to praise us because someone else has helped a person in distress, nor do we expect to feel pride if the target's distress is alleviated by a stranger, or by chance. So on the version of the empathy-specific rewards hypothesis that we are considering, we should predict that an empathically aroused agent will be pleased when he gets to help the target (either because he is feeling a jolt of pride or because he is looking forward to the rewards that others will provide), but he will not be pleased if he is unable to help the target.

For reasons that will emerge shortly, Batson distinguishes two cases in which the agent is unable to help. In the first, there is just nothing he can do to relieve the target's distress; in the second, someone (or something) else relieves the target's distress before the agent gets a chance to act. In both cases, Batson maintains, the egoistic reward-seeking agent has nothing in particular to be pleased about. Finally, Batson suggests that we can determine whether an agent is pleased by using self-report tests to assess changes in his mood—the more the mood has improved, the more pleased the agent is.

If the empathy–altruism hypothesis is correct, then agents are motivated by an ultimate desire that the target's distress be alleviated. So on this hypothesis, it shouldn't matter *how* the target's distress is alleviated. No matter how it

is accomplished, the agent's altruistic ultimate desire will be satisfied. If we assume that people are pleased when their ultimate desires are satisfied, then the empathy–altruism hypothesis predicts that empathically aroused individuals should be pleased—and thus have elevated moods—whenever the target is helped. They should be displeased, and exhibit a lower mood, only in the case where the target's distress is not alleviated.

To see which of these predictions was correct, Batson and his associates designed an experiment in which participants were told that they would likely have the chance to perform a simple task that would reduce the number of electric shocks that a peer would receive (Batson et al., 1988: study 1). Somewhat later, half of the participants learned, by chance, that they would not be performing the helping task after all, and thus that they could not help the other student. This divided the participants into two experimental conditions, "perform" and "not perform." Subsequently, half of the participants in each condition learned that, by chance, the peer was not going to get the shocks, while the other half learned that, by chance, the peer would still have to get the shocks. This yielded two more experimental conditions, "prior relief." and "no prior relief". All participants were also asked to self-report their level of empathy for the peer, so that high- and low-empathy participants could be distinguished. To assess mood change, the moods of all participants were measured both before and after the experimental manipulation. As we saw above, the version of the empathy-specific reward hypothesis that we're considering predicts that participants in the perform + no prior relief condition should indicate an elevated mood, since they were able to help the peer; it also predicts that participants in all the other conditions should not have an elevated mood, since for one reason or another they were unable to help, and thus were ineligible for the reward. The empathy–altruism hypothesis, by contrast, predicts an elevated mood in all three conditions in which the peer escaped the shocks: perform + no prior relief, perform + prior relief, and not perform + prior relief. The only condition in which empathy–altruism predicts low mood is the one in which the peer gets the shocks: not perform + no prior relief. In fact, the results fit the pattern predicted by the empathy–altruism hypothesis, not the pattern predicted by empathy-specific reward.

We are inclined to agree with Batson that this experiment shows that the version of the empathy-specific reward hypothesis we've been considering is less plausible than the empathy–altruism hypothesis.[36] However, there's

[36] Batson and colleagues (Batson et al., 1988) also did a Stroop experiment aimed at testing this version of the empathy-specific reward hypothesis. But for the reasons discussed in the previous section, we don't think the Stroop procedure is useful in this context.

another way of elaborating the self-administered version of the empathy-specific reward hypothesis that the experiment does not address. On the version of the empathy-specific reward hypothesis we've been considering, the self-administered rewards come from something like a "jolt of pride" that the empathically aroused agent feels when he helps the target. And, as Batson rightly notes, it is unlikely that an agent would expect to get *this* reward if he were in no way involved in the relief of the target's distress. But the jolt of pride story is not the only one that an egoist can tell about the self-administered reward an empathically aroused agent might anticipate when confronted with an opportunity to help; another option focuses on the vicarious *pleasure* that empathically aroused agents might expect to feel when the target's distress is alleviated. This account, the *empathic-joy hypothesis*, maintains that empathically aroused individuals "help to gain the good feeling of sharing vicariously in the needy person's joy at improvement" (Batson et al., 1991: 413). On this story, the actor's ultimate goal is egoistic; the desire to help is just instrumental.

There have been a number of experiments aimed at testing the empathic-joy hypothesis. All of them rely on manipulating subjects' expectations about the sort of *feedback* they can expect about the condition of the target. The central idea in two of these experiments[37] was that if the empathic-joy hypothesis is correct, then high-empathy subjects should be more highly motivated to help when they expect to get feedback about the effect of their assistance on the target's well-being than when they have no expectation of learning about the effect of their assistance. In the latter ("no-feedback") condition subjects won't know if the target's situation has improved and thus they can't expect to experience vicarious joy. On the empathy–altruism hypothesis, by contrast, high-empathy subjects are motivated by an ultimate desire for the well-being of the target, so we should not expect those anticipating feedback to be more likely to help than those not anticipating feedback. In both experiments, the Stotland technique was used to manipulate empathy, and in both cases the subjects who were instructed to imagine how the target felt failed to show a higher level of helping in the feedback condition than in the no-feedback condition. This looks like bad news for the empathic-joy hypothesis, but for two reasons, the situation is less than clear-cut. First, doubts have been raised about the effectiveness of the Stotland manipulation in these experiments.[38] Second, in one experiment there was an unexpected finding: while high-empathy subjects helped more than low-empathy subjects in the no-feedback condition, they actually helped *less* in the feedback condition.

[37] Smith et al. (1989) and Batson et al. (1991), experiment 1.

[38] For discussion, see Smith et al. (1989) and Batson et al. (1991).

In an effort to buttress the case against the empathic-joy hypothesis, Batson and his colleagues designed two additional experiments in which the rationale was rather different.[39] If the empathic-joy hypothesis is true, they reasoned, then if high-empathy subjects listen to a taped interview detailing the plight of a troubled target in the recent past and are then offered a choice between getting an update on how the target is doing and hearing about another person, there should be a linear relationship between the probability that the target has improved and the likelihood of choosing to get an update on the target, since the more likely it is that the target has improved, the more likely it is that the subject will get to experience the vicarious joy that he seeks. In both experiments, subjects were given what were alleged to be experts' assessments of the likelihood that the target would improve in the time between the first and second interviews. Neither experiment showed the sort of linear relationship that the empathic-joy hypothesis predicts.

We agree with Batson and colleagues' contention that these results "cast serious doubt on the empathic-joy hypothesis" (1991: 425). But as they go on to note, the experiments were not designed to test the empathic-joy hypothesis *against* the empathy–altruism hypothesis, since the latter hypothesis makes no clear prediction about the scenarios in question. Therefore, Batson and colleagues (1991: 425) are appropriately cautious, observing that the experiments "did not provide unequivocal support" for empathy–altruism. The bottom line, as we see it, is that while the empathic-joy hypothesis does not look promising, more evidence is needed before coming to a final judgment.

4.7. *The Social Psychology of Altruism: Summing Up*

Batson concludes that the work we have reviewed in this section gives us good reason to think that the empathy–altruism hypothesis is true.

> Sherlock Holmes stated: "When you have eliminated the impossible, whatever remains, *however improbable*, must be the truth." If we apply Holmes's dictum to our attempt to answer the altruism question, then I believe we must, tentatively, accept the truth of the empathy–altruism hypothesis. It is impossible for any of the three major egoistic explanations of the empathy-helping relationship—or any combination of these—to account for the evidence reviewed. (Batson, 1991: 174)

Although we don't believe that this conclusion is justified, we think it is clear that Batson and his associates have made important progress. They have shown that one widely endorsed account of the egoistic motivation underlying helping

[39] Batson et al. (1991), experiments 2 and 3.

behavior, the aversive-arousal reduction hypothesis, is very unlikely to be true in the sorts of cases used in their studies. They have also dealt a serious blow to both the socially administered empathy-specific punishment hypothesis and to several versions of the empathy-specific reward hypothesis. However, we think the jury is still out on the self-administered empathy-specific punishment hypothesis, and that the case against the empathy-specific reward hypothesis is not yet conclusive.

A worry of another sort emerges when we focus on Batson's claim that no "combination" of the three major egoistic explanations could explain the experimental data. An egoistic thesis that might be labeled *disjunctive* egoism maintains that when empathically aroused people try to help, they have a variety of egoistic motivations—they are sometimes motivated by the desire to reduce aversive arousal, sometimes by the desire to avoid socially or self-administered punishment and sometimes by the desire for socially or self-administered reward. Since all of Batson's experiments are designed to test empathy–altruism against *one or another* specific egoistic hypothesis, none of these experiments rules out this sort of disjunctive egoism. For it might be the case that in each experiment subjects are motivated by one of the egoistic goals that the experiment is *not* designed to rule out. We're not sure how seriously to take this concern, since it seems to require that nature is playing a shell game with the investigators, always relying on an egoistic motivation that the experiment is not designed to look for. But we do think the idea deserves more explicit attention than it has so far received in the literature.[40] Clearly, there is still much important work to be done on the social psychology of altruism.

5. Conclusion

Readers might be tempted to think, at this point, that our concluding section must be rather inconclusive. After all, we haven't claimed to have resolved the philosophically venerable egoism vs. altruism debate, and the scientific record appears somewhat equivocal, as indeed we've been at pains to show. But before we offer refunds, we should enumerate what we think we *have* learned.

Our first lesson is negative: contrary to what some writers have asserted, appeal to evolutionary theory does not generate movement in the philosophical debate about altruism. This may seem disappointing, especially given

[40] This worry about Batson's one-at-a-time strategy was noted, albeit briefly, in Cialdini (1991). For a helpful discussion, see Oakberg (unpublished ms.).

the fecundity of recent explications of philosophical ethics in the light of evolutionary theory (e.g. Joyce, 2006: Machery & Mallon, Chapter 1, this volume). Fortunately, our conclusions regarding the philosophical impact of experimental social psychology are rather more inspiring. Batson and associates have shown quite conclusively that the methods of experimental psychology can move the debate forward; it now looks as though certain venerable renderings of psychological egoism are not true to the contours of human psychology. Indeed, in our view, Batson and his associates have made more progress in the last three decades than philosophers using the traditional philosophical methodology of *a priori* arguments buttressed by anecdote and intuition have made in the previous two millennia. Their work, like other work recounted in this volume, powerfully demonstrates the utility of empirical methods in moral psychology.

References

Aquinas, T. (1270/1917). *The Summa Theologica* (Vol. 2, Part II). New York: Benziger Brothers.

Aubrey, J. (1949). *Aubrey's Brief Lives*, ed. Oliver Lawson Dick. Boston, MA: David R. Godine. Aubrey's sketch of Hobbes is available online at: http://www-groups.dcs.st-and.ac.uk/~history/Societies/Aubrey.html

Axelrod, R. & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211: 1390–1396.

—— (1991). *The Altruism Question: Toward a Social-Psychological Answer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Batson, C. D. (1998). Altruism and prosocial behavior. In D. T. Gilbert & S. T. Fiske (eds.), *The Handbook of Social Psychology*, Vol. 2. Boston, MA: McGraw-Hill, 282–316.

Batson, C. D., Duncan, B., Ackerman, P., Buckley, T., & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology*, 40: 290–302.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., & Isen, A. (1983). Self-reported distress and empathy and egoistic versus altruistic motivation for helping. *Journal of Personality and Social Psychology*, 45: 706–718.

Batson, C. D., Dyck, J., Brandt, R., Batson, J., Powell, A., McMaster, M., & Griffitt, C. (1988). Five studies testing two new egoistic alternatives to the empathy–altrusim hypothesis. *Journal of Personality and Social Psychology*, 55: 52–77.

Batson, C. D., Batson, G., Slingsby, J., Harrell, K., Peekna, H., & Todd, R. M. (1991). Empathic joy and the empathy–altruism hypothesis. *Journal of Personality and Social Psychology*, 61: 413–426.

Beatty, J. (1992). Fitness. In E. Keller & L. Lloyd (eds.), *Keywords in Evolutionary Biology*. Cambridge, MA: Harvard University Press.

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*, ed. J. H. Burns & H. L. A. Hart, with a new introduction by F. Rosen. Oxford: Oxford University Press, 1996.

—— (1824). *The Book of Fallacies*. London: Hunt.

Boyd, R. & Richerson, P. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13: 171–195. Reprinted in R. Boyd & P. Richerson, *The Origin and Evolution of Cultures*. Oxford: Oxford University Press, 2005.

Bratman, M. (1987). *Intention, Plans, and Practical Reasoning*. Cambridge, MA: Harvard University Press.

Broad, C. D. (1930). *Five Types of Ethical Theory*. New York: Harcourt, Brace.

Butler, J. (1726). *Fifteen Sermons Preached at the Rolls Chapel*. Sermons I, II, III, XI, XII, reprinted in S. Darwall (ed.), *Five Sermons Preached at the Rolls Chapel and A Dissertation Upon the Nature of Virtue*. Indianapolis, IN: Hackett, 1983.

Cialdini, R. B. (1991). Altruism or egoism? That is (still) the question. *Psychological Inquiry*, 2: 124–126.

Carey, S. & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, 63 (4): 515–533.

Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60 (23): 685–700.

—— (1980). Agency. In *Essays on Actions and Events*. Oxford, Clarendon Press, 43–61.

Dovidio, J., Allen, J., & Schroeder, D. (1990). The specificity of empathy-induced helping: Evidence for altruistic motivation. *Journal of Personality and Social Psychology*, 59: 249–260.

Dovidio, J., Piliavin, J., Schroeder, D., & Penner, L. (2006). *The Social Psychology of Prosocial Behavior*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eisenberg, N. & Miller, P. (1987). Empathy and prosocial behavior. *Psychological Bulletin*, 101: 91–119.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: Bradford Books/MIT Press.

Fultz, J., Batson. D., Fortenbach, V., McCarthy, P., & Varney, L. (1986). Social evaluation and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, 50, 761–769.

Geller, V. & Shaver, P. (1976). Cognitive consequences of self-awareness. *Journal of Experimental Social Psychology*, 12: 99–108.

Ghiselin, M. (1974). *The Economy of Nature and the Evolution of Sex*. Berkeley, CA: University of California Press.

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

Goldman, A. (1970). *A Theory of Human Action*. Englewood-Cliffs, NJ: Prentice-Hall.

Grant, C. (1997). Altruism: A social science chameleon. *Zygon*, 32 (3): 321–340.

Grice, H. P. (1971). Intention and certainty. *Proceedings of the British Academy*, 57: 263–279.

Hamilton, W. D. (1963). The evolution of altruistic behavior. *American Naturalist*, 97: 354–356.
—— (1964a). The general evolution of social behavior I. *Journal of Theoretical Biology*, 7: 1–16.
—— (1964b). The general evolution of social behavior II. *Journal of Theoretical Biology*, 7: 17–52.
Hobbes, T. (1981). *Leviathan*. Edited with an introduction by C. B. Macpherson. London: Penguin Books. First published 1651.
Hoffman, M. (1991). Is empathy altruistic? *Psychological Inquiry*, 2: 131–133.
Hornstein, H. (1991). Empathic distress and altruism: Still inseparable. *Psychological Inquiry*, 2: 133–135.
Hume, D. (1975). *Enquiry Concerning the Principles of Morals*, ed. L. A. Selby-Bigge, 3rd ed. revised by P. H. Nidditch. Oxford: Clarendon Press. Originally published 1751.
Joyce, R. (2006). *The Evolution of Mind*. Cambridge, MA: MIT Press.
Kant, I. (1785/1949). *Fundamental Principles of the Metaphysics of Morals*, trans. Thomas K. Abbott. Englewood Cliffs, NJ: Prentice-Hall/Library of Liberal Arts.
Krebs, D. (1975). Empathy and altruism. *Journal of Personality and Social Psychology*, 32: 1134–1146.
LaFollette, H. (2000a) (ed.). *The Blackwell Guide to Ethical Theory*. Oxford: Blackwell.
—— (2000b). Introduction. In LaFollette (2000a), 1–12.
La Rochefoucauld, F. (2007). *Collected Maxims and Other Reflections*, trans. E. H. Blackmore, A. M. Blackmore, & Francine Giguère. New York: Oxford University Press. Originally published 1665.
Lewis, D. (1997). Finkish dispositions. *Philosophical Quarterly*, 47: 143–158.
MacIntyre, A. (1967). Egoism and altruism. In P. Edwards (ed.), *The Encyclopedia of Philosophy*, Vol. 2. New York: Macmillan, 462–466.
Martin, C. B. (1994). Disposition and conditionals. *Philosophical Quarterly*, 44: 1–8.
Mill, J. S. (1861/2001). *Utilitarianism*. Indianapolis, IN: Hackett.
Miller, D. T. (1999). The Norm of Self-Interest. *American Psychologist*, 54: 1053–1060
Nagel, T. (1970). *The Possibility of Altruism*. Oxford: Oxford University Press.
Nesse, R. (1990). Evolutionary explanations of emotions. *Human Nature*, 1: 261–289.
Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford: Oxford University Press.
Nietzsche, F. (1997). *Daybreak: Thoughts on the Prejudices of Morality*, trans. R. J. Hollingsdale, ed. M. Clark & B. Leiter. Cambridge: Cambridge University Press. Originally published 1881.
Oakberg, T. (unpublished ms). A critical review of Batson's project and related research on altruism.
Piliavin, J. & Charng, H. (1990). Altruism—A review of recent theory and research. *Annual Review of Sociology*, 16: 27–65.
Plantinga, A. (1993). *Warrant and Proper Function*. Oxford: Oxford University Press.
Rachels, J. (2000). Naturalism. In LaFollette (2000a), 74–91.
Rousseau, J. J. (1985). *A Discourse on Inequality*. New York: Penguin. Originally published 1754.

Schroeder, D., Penner, L., Dovidio, J., & Piliavin, J (1995). *The Psychology of Helping and Altruism*. New York: McGraw-Hill.
Schroeder, W. (2000). Continental ethics. In LaFollette (2000a), 375–399.
Smith, A. (1759/1853). *The Theory of Moral Sentiments*. London: Henry G. Bohn.
Smith, K., Keating, J., & Stotland, E. (1989). Altruism revisited: The effect of denying feedback on a victim's status to emphatic witnesses. *Journal of Personality and Social Psychology*, 57: 641–650.
Sober, E. (1994). The adaptive advantage of learning and *a priori* prejudice. In *From a Biological Point of View*. Cambridge: Cambridge University Press.
Sober, E. & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.
Spelke, E. (2000). Core knowledge. *American Psychologist*, 55: 1233–1243.
—— (2003). Core knowledge. In N. Kanwisher & J. Duncan (eds.), *Attention and Performance, vol. 20: Functional neuroimaging of visual cognition*. Oxford: Oxford University Press, 29–56.
Sripada, C. (2007). Adaptationism, culture and the malleability of human nature. In P. Carruthers, S. Laurence, & S. Stich (eds.), *Innateness and the Structure of the Mind: Foundations and the Future*. New York: Oxford University Press, 311–329.
Stich, S. (1978). Beliefs and sub-doxastic states, *Philosophy of Science*, 45(4): 499–518.
—— (1990). *The Fragmentation of Reason*. Cambridge, MA: MIT Press.
—— (2007). Evolution, altruism and cognitive architecture: A critique of Sober and Wilson's argument for psychological altruism, *Biology & Philosophy*, 22 (2): 267–281.
Stocks, E., Lishner, D., & Decker, S. (2009). Altruism or psychological escape: Why does empathy promote prosocial behavior? *European Journal of Social Psychology*, 39: 649–665.
Stotland, E. (1969). Exploratory studies of empathy. In L. Berkowitz (ed.), *Advances in Experimental Social Psychology*, Vol. 4. New York: Academic Press, 271–313.
Toi, M. & Batson, C. D. (1982). More evidence that empathy is a source of altruistic motivation. *Journal of Personality and Social Psychology*, 43: 281–292.
Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46: 35–57.
Velleman, D. (1989). *Practical Reflection*. Princeton, NJ: Princeton University Press.
Wilson, G. (2002). Action. In *The Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2002/entries/action/>.
Wright, R. (1994). *The Moral Animal*. New York: Pantheon Books.