Chapter 7

Psychological Egoism

Elliott Sober

Psychological egoism is a theory about motivation that claims that all of our ultimate desires are self-directed. Whenever we want others to do well (or ill), we have these other-directed desires only instrumentally; we care about others only because we think that the welfare of others will have ramifications for our own welfare. As stated, egoism is a descriptive, not a normative, claim. It aims to characterize what motivates human beings in fact; the theory does not say whether it is good or bad that people are so motivated.

Egoism has exerted a powerful influence in the social sciences and has made large inroads in the thinking of ordinary people. Economists typically think of human beings as being moved by "rational self-interest," where this excludes any irreducible concern for the welfare of others. And ordinary folks often claim that people help others only because this makes them feel good about themselves, or because they seek the approval of third parties.

It is easy to invent egoistic explanations for even the most harrowing acts of self-sacrifice. The soldier in a foxhole who throws himself on a grenade to save the lives of his comrades is a fixture in the literature on egoism. How could this act be a product of self-interest if the soldier knows that it will end his life? The egoist may answer that the soldier realizes in an instant that he would rather die than suffer the guilt feelings that would haunt him if he saved himself and allowed his friends to perish. The soldier prefers to die and then have no sensations at all rather than live and suffer the torments of the damned. This reply may sound forced, but it remains to be seen what grounds we have for regarding it as false.

The criticisms that have been leveled against psychological egoism can be divided into three categories. First, there is the claim that it is not a genuine theory at all. Second, there is the allegation that it is a theory that is refuted by what we observe in human behavior. Third, there is the idea that, although egoism is a theory that is consistent with what we observe, there are other, extra-evidential considerations that suggest that it should be rejected in favor of an alternative theory, motivational pluralism, according to which human beings have both egoistic and altruistic

ultimate desires. All three types of criticism will be considered in what follows, but first we need to state the theory more carefully.

1 Clarifying Egoism

When egoism claims that all our ultimate desires are self-directed, what do "ultimate" and "self-directed" mean?

There are some things that we want for their own sakes; other things we want only because we think they will get us something else. The familiar means/end relation that links one desire to another also allows desires to be chained together – Sarah may want to drive her car because she wants to get to the bakery, she may want to go to the bakery because she wants to buy bread, etc. The crucial relation that we need to define is this:

> *S* wants *m* solely as a means to acquiring *e* if and only if *S* wants *m*, *S* wants *e*, and *S* wants *m* only because she believes that obtaining *m* will help her obtain *e*.

An ultimate desire is simply a desire that someone has for reasons that go beyond its ability to contribute instrumentally to the attainment of something else. Consider pain. The most obvious reason that people want to avoid pain is simply that they dislike experiencing it. Avoiding pain is one of our ultimate goals. However, many people realize that being in pain reduces their ability to concentrate, so they may sometimes take an aspirin in part because they want to remove a source of distraction. This shows that the things we want as ends in themselves we also may want for instrumental reasons.

When psychological egoism seeks to explain why one person helped another, it isn't enough to show that *one* of the reasons for helping was self-benefit; this is quite consistent with there being another, purely altruistic, reason that the individual had for helping. Symmetrically, to refute egoism, one need not cite examples of helping in which only other-directed motives play a role. If people sometimes help for both egoistic and altruistic ultimate reasons, then psychological egoism is false.

Egoism and altruism both require the distinction between self-directed and other-directed desires. This distinction is to be understood in terms of a desire's propositional content. If Adam wants the apple, this is elliptical for saying that Adam wants it to be the case that *he has the apple*. This desire is purely self-directed, since its propositional content mentions Adam, but no other agent; I assume that Adam does not regard the apple as an agent. In contrast, when Eve wants *Adam to have the apple*, this desire is purely other-directed; its propositional content mentions another person, Adam, but not Eve herself. Egoism claims that all of our ultimate desires are self-directed; altruism, that some are other-directed. The fact that Eve has an other- directed desire is not enough to refute egoism; one must ask *why* Eve wants Adam to have the apple.

A special version of egoism is psychological hedonism. The hedonist says that the only ultimate desires that people have are attaining pleasure and avoiding pain. Hedonism is sometimes criticized for holding that pleasure is a single type of sensation – that the pleasure we get from the taste of a peach and the pleasure we get from seeing those we love prosper somehow boil down to the same thing (LaFollette 1988). However, this criticism does not apply to hedonism as I have described it. The salient fact about this theory is its claim that people are motivational solipsists; the only things they care about ultimately are states of their own consciousness. Egoists need not be hedonists. If people desire their own survival as an end in itself, they may be egoists, but they are not hedonists.

There are desires that are neither purely self-directed nor purely other-directed. If Phyllis wants to be famous, this means that she wants others to know who she is. This desire's propositional content involves a relation between self and others. If Phyllis seeks fame solely because she thinks this will be pleasurable or profitable, then she may be an egoist (depending on what her other ultimate desires happen to be). But what if she wants to be famous as an end in itself? There is no reason to cram this possibility within egoism or altruism; to include some ultimate relational desires, but not others, within egoism, runs the risk of making the theory appear ad hoc or unclear (Kavka 1986); the same point also applies to altruism. So let us recognize *relationism* as a possibility distinct from both.

A fourth possibility involves desires that mention neither self nor other. The desire that some general moral principle be upheld falls into this category. When a utilitarian desires the greatest good for the greatest number, the desire is impersonal; the desire covers all sentient beings, presumably including the desirer himself, but the desire's content singles out neither self nor specific others. For this reason, I suggest that it is neither altruistic nor egoistic. Just as was true with respect to relational desires, the defender of psychological egoism can grant that there are desires concerning general moral principles that are not self-directed; the question is whether we have these desires instrumentally or as ends in themselves.

With egoism characterized in the way I have suggested, it obviously is not entailed by the truism that people act on the basis of their own desires, nor by the truism that they seek to have their desires satisfied. The fact that Joe acts on the basis of Joe's desires, not on the basis of Jim's, tells us whose desires are doing the work; it says nothing about whether the ultimate desires in Joe's head are purely self-directed. And the fact that Joe wants his desires to be satisfied means merely that he wants their propositional contents to come true; Joe's desire that it rain tomorrow is satisfied if and only if it rains tomorrow (Stampe 1994). If there is rain, the desire is satisfied, whether or not Joe knows that it is. To want one's desires satisfied is not the same as wanting the feeling of satisfaction that sometimes accompanies a satisfied desire.

Egoism is sometimes criticized for attributing too much calculation to spontaneous acts of helping. People who help in emergency situations often report doing so "without thinking" (Clark and Word 1974). However, it is hard to take such reports literally when the acts involve a precise series of complicated actions that are

well-suited to an apparent end. A lifeguard who rescues a struggling swimmer is properly viewed as having a goal and as selecting actions that advance that goal. The fact that she engaged in no ponderous and self-conscious calculation does not show that no means/end reasoning occurred. In any case, actions that really do occur without the mediation of beliefs and desires fall outside the scope of both egoism and altruism. People jerk their legs when their knees are tapped with hammers, but that refutes neither theory.

A related criticism is that egoism assumes that people are more rational than they really are. However, recall that egoism is simply a claim about the ultimate desires that people have. As such, it says nothing about how people decide what to do on the basis of their beliefs and desires. Theorists who assume that egoism is true also often assume that people are rational calculators; however, theories are not convicted by a principle of guilt by association. The assumption of rationality is no more a part of psychological egoism than it is part of motivational pluralism.

If egoism holds that all ultimate desires are self-directed, what are we to say of someone whose ultimate goal is his own destruction? And if altruism holds that some of our ultimate desires are other-directed, what are we to make of Iago, who has the ultimate goal of destroying Othello? It is jarring to say that a depressed person bent on suicide is an egoist, or that Iago is an altruist. What we need to add to both theories is the idea of what is good (or apparently good). Egoists seek their own benefit; altruists want others to do well. Although these additions to the theories bring them more in line with ordinary usage of the terms "egoism" and "altruism," they do not materially affect the substantive task of determining which theory is true. The crux of the problem is to tell whether all ultimate desires are self-directed.

It may strike some readers that the problem is easy. Individuals can merely gaze within their own minds and determine by introspection what their ultimate motives are. Perhaps advocates of egoism are right about themselves and advocates of motivational pluralism are right about themselves; both sides err only when they generalize beyond their own cases. An implicit assumption, in both philosophical and psychological explorations of this topic, is that people are basically the same. If egoism is false, it is false for practically everyone (sociopaths, perhaps, excepted). And if it is true, it is true because it characterizes a basic feature of human nature.

However, the fact that earlier work in psychology and philosophy often ignored the possibility of individual variation is no reason to build this into our understanding of the problem. Why, then, should we not say that advocates of egoism know their own hearts and that defenders of altruism know theirs? The reason is that there is no independent reason to think that the testimony of introspection is to be trusted in this instance. Introspection is misleading or incomplete in what it tells us about other facets of the mind; no one has shown why the mind must be an open book with respect to this question about ultimate motives. The problem, if it can be solved, must be solved in some other way.

2 Is Egoism Empirically Testable?

One standard philosophical objection to egoism is that it is not a testable hypothesis. As the example of the soldier in the foxhole suggests, it seems that egoism can accommodate any behavior whatever. Whether people are nasty or nice to each other, the theory can explain why. This claim about the flexibility of egoism is then linked to a Popperian criterion concerning what it takes for a statement to be scientific, with the conclusion drawn that egoism is not a genuine scientific theory at all. It is, despite appearances, empirically vacuous.

This argument is flawed in two ways. The first pertains to its sanguine confidence that no observation could ever disconfirm egoism. The fact that the theory can accommodate the soldier in the foxhole and other behaviors that have been considered by philosophers hardly suffices to justify this global claim. As it happens, the experimental work in social psychology on altruism and egoism shows that the relevant observational evidence extends beyond the existence of instances of helping behavior (Batson 1991; Schroeder et al. 1995). In addition, the Duhemian point that theories are testable only in conjunction with background assumptions should lead us to draw back from the charge of untestability. If two theories make the same predictions against one background framework, they may make different predictions against another. How do we know that new background theories will never be developed that allow egoism to be put to the test? The charge of untestability presupposes that we have an omniscient grasp of the future of science.

The second defect in this argument is that it neglects to notice that the charge of untestability is a two-edged sword. The argument is advanced as a reason for rejecting egoism. What, then, are we to accept as a positive account of motivation? Presumably, motivational pluralism is supposed to be the acceptable alternative. However, this cannot be where the argument leads. If egoism is untestable, then so is motivational pluralism. As flexible as egoism is in its ability to accommodate observations, pluralism is more flexible still. After all, pluralism deploys all the variables that egoism invokes, and then some. The two theories are related to each other in the same way that "$y = f(x)$" and "$y = g(x,w)$" are related.

The reason egoism appears to be untestable is that it is an *ism*. It does not provide *specific* explanations for behaviors, but merely indicates the *kind* of explanation that all behaviors will have. This is why it is possible for egoism to be retained even when specific egoistic explanations are found wanting. Why did George donate all that money to charity? A defender of egoism might suggest that George did so because he wanted to improve his business contacts by impressing others. However, suppose one then learns that George donated the money anonymously. This refutes the specific egoistic explanation just described, but it isn't hard to invent another. George made the donation because it made him feel good and because he knew that if he did not, he would experience pangs of guilt. The pattern here is typical – hedonism is the position to which egoists standardly retreat. If external benefits don't suffice to explain, one invokes internal, psychological benefits instead.

That egoism is a claim about a *type* of explanation, and therefore is distinct from the *specific explanations* that are of the type required, is a pattern that arises in many debates about *isms*. Consider adaptationism in evolutionary biology. Adaptationists emphasize the importance of natural selection in explaining the observed traits of organisms. Because this *ism*, by itself, does not provide a specific explanation for any trait, it remains possible for a biologist to continue to be an adaptationist even after a specific adaptationist explanation is found wanting. Why did wings evolve in insects? The hypothesis that wings evolved as an adaptation for flying is thrown in doubt by the fact that very small wing buds provide no lift whatever; although 5 percent of an eye can still function as a light sensor, 5 percent of a wing does nothing to get an organism off the ground. However, wing buds are found in some flightless insect species; they function as thermo-regulators. This suggests an alternative adaptationist hypothesis – that insect wings started to evolve because they initially promoted thermo-regulation and then continued to evolve because they then facilitated flight. And if this hypothesis is challenged, the adaptationist can cast about for a third alternative. It is no good to reject adaptationism because it has this sort of flexibility; the alternative *ism*, evolutionary pluralism, claims that natural selection is one among several important causes of evolution. As flexible as adaptationism is, pluralism is more flexible still.

3 Butler's Stone

As noted above, even though hedonism is a special version of egoism, hedonistic explanations are often what egoists invoke when a nonhedonistic explanation is found wanting. If George didn't donate money to charity to make business contacts, perhaps he did so for the warm glow of satisfaction that the donation provided. For this reason, arguments that attempt to refute hedonism have a special location in the dialectical landscape. Although refuting hedonism is not sufficient to refute egoism, it would make an important contribution to that larger enterprise.

Many philosophers have thought that Joseph Butler (1692–1752) refuted hedonism once and for all (Broad 1965; Feinberg 1984; Nagel 1970) in the following passage:

> That all particular appetites and passions are towards *external things themselves*, distinct from the *pleasure arising from them*, is manifested from hence; that there could not be this pleasure, were it not for that prior suitableness between the object and the passion: there could be no enjoyment or delight from one thing more than another, from eating food more than from swallowing a stone, if there were not an affection or appetite to one thing more than another. (Butler 1965 [1726]: 227)

I'll call this argument *Butler's stone*. Although Butler does not explicitly say in this passage that hedonism is false, let us construe the argument with this as its conclusion:

1 People sometimes experience pleasure.
2 When people experience pleasure, this is because they had a desire for some external thing, and that desire was satisfied.

Hedonism is false.

I don't propose to challenge the first premiss. However, I think the second premiss is false and that the conclusion does not follow from the premisses.

The second premiss is over-stated; although some pleasures are the result of a desire's being satisfied, others are not (Broad 1965: 66). One can enjoy the smell of violets without having formed the desire to smell a flower, or something sweet. Since desires are propositional attitude, forming a desire is a cognitive achievement. Pleasure and pain, on the other hand, are sometimes cognitively mediated, but sometimes they are not. Notice that this defect in the argument can be repaired; Butler does not need to say that desire satisfaction is the one and only road to pleasure.

The transition from premisses to conclusion is where the argument really goes wrong. Consider the causal chain from a *desire* (the desire for food, say), to an *action* (eating), to a *result* – pleasure. Because the pleasure traces back to an antecedently existing desire, it will be false that the resulting pleasure caused the desire (on the assumption that cause must precede effect). However, this does not settle how two *desires* – the *desire for food* and the *desire for pleasure* – are related. In particular, it leaves entirely open what caused the desire for food. Hedonism says that people desire food *because* they want pleasure (and think that food will bring them pleasure). Butler's stone concludes that this causal claim is false, but for no good reason. The crucial mistake in the argument comes from confusing two quite different items – the *pleasure* that results from a desire's being satisfied and the *desire for pleasure*. Even if the occurrence of pleasure presupposed that the agent desired something besides pleasure, nothing follows about the relationship between the *desire for pleasure* and the desire for something else (Sober 1992; Stewart 1992; Sober and Wilson 1998). Hedonism does not deny that people desire external things; rather, the theory tries to explain why that is so.

It is curious that this argument has been interpreted so widely as refuting hedonism. At the end of the sermon in which the stone passage occurs, Butler says this: "Let it be allowed, though virtue or moral rectitude does indeed consist in affection to and pursuit of what is right and good, as such; yet, that when we sit down in a cool hour, we can neither justify to ourselves this or any other pursuit, till we are convinced that it will be for our happiness, or at least not contrary to it" (Butler 1965 [1726]: 240). And if we return to the language of the stone argument itself, we see that Butler is making a claim about the content of "particular appetites and passions." Read narrowly, the argument says merely that if people desire pleasure, their desires do not fall under that rubric; the argument does not say that people never desire pleasure nor does it say that the desire for pleasure is never ultimate. Did Butler fail to refute hedonism in the stone argument because he wasn't even trying to do so?

4 The "Paradox" of Hedonism and its "Irrationality"

Individuals who focus exclusively on attaining pleasure or happiness inevitably fail to get what they want. They are like stockbrokers who think only that they should buy low and sell high. People who have an end in view but never consider what means they should use to pursue their goal surely will fail to get what they want. This has led some philosophers to claim that pleasure and happiness are attainable only as byproducts of becoming absorbed in specific activities. They also have suggested that this fact about pleasure and happiness constitutes a paradox for hedonism – the word "paradox" indicating that we are supposed to find here a flaw in hedonism as a psychological theory (Butler 1965 [1726]; Feinberg 1984).

The obvious reply to this criticism is that there is nothing in hedonism that says that people must be monomaniacs. Hedonism says that people have attaining pleasure and avoiding pain as their only *ultimate* goals; it does not say that attaining pleasure and avoiding pain are the only goals (ultimate *or* proximate) that people ever have. Hedonists reflect on which activities are most apt to bring pleasure and prevent pain, and decide what to do on that basis (Sidgwick 1922 [1907]). Furthermore, if hedonistic monomaniacs always fail to get what they want, what follows from this? Even if this entailed that people *should* not be hedonists, it does not show that people are not hedonists *in fact*. Recall that hedonism is a descriptive, not a normative, theory.

The normative/descriptive distinction also is needed to evaluate the claim that egoism is irrational. Nagel (1970) defends this claim by contending that when egoists consider their own interests in deliberation, but not those of others, they neglect the fact that there is no property that they have and others lack that could justify this asymmetry. To evaluate whether egoists are irrational, we need to decide whether rationality should be understood "instrumentally" or "substantively." Instrumental rationality just means the ability to choose efficient means to achieve whatever ends one might have. The substantive notion means, not just that efficient means have been secured, but that the ends are praiseworthy, or at least are morally unobjectionable (Gibbard 1990). Efficient serial killers might be instrumentally rational, but they are not substantively rational. Regardless of which notion captures what the word "rational" means, the fact remains that this line of argument cannot show that people really have or are capable of having altruistic ultimate motives. If rationality just means instrumental rationality, then rationality does not entail altruism (or its possibility); and if rationality means substantive rationality, then even if rationality entails altruism, it needs to be shown that people really are substantively rational. Perhaps we *ought* to be rational and maybe we *ought* to be altruistic as well. This does not show that egoism is false as a descriptive thesis.

5 The Experience Machine

In the science fiction movie *Total Recall*, people centuries from now use their computer technology to go on "virtual vacations." Instead of going on a real vacation, they plug into a computer that provides a thoroughly convincing simulation of a real vacation. The movie quite plausibly suggests that people in the future often might choose to "vacation" in this way, especially if real trips to exotic locales are expensive and dangerous, while "virtual vacations" are cheap and completely convincing from an experiential point of view.

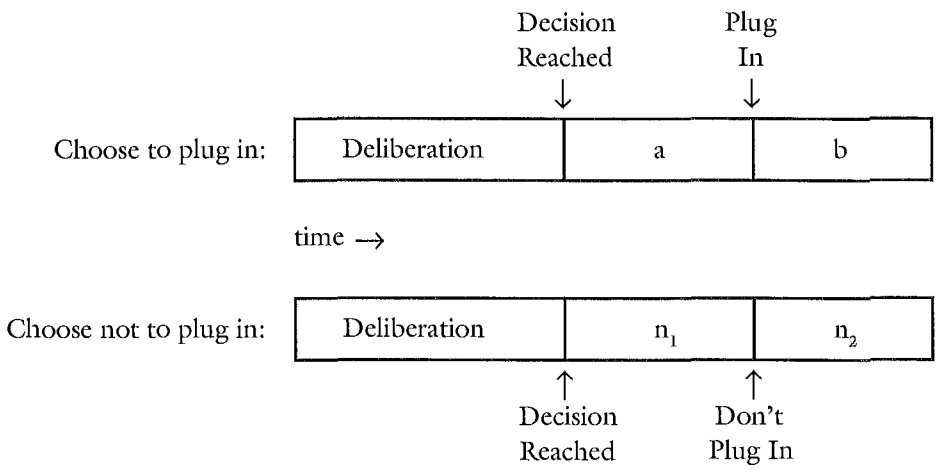
Robert Nozick wrote *Anarchy, State, and Utopia* considerably before *Total Recall* appeared. He there uses the idea of an "experience machine" to construct an argument that seems to show that hedonism is false (Nozick 1974: 42–5). Nozick's machine can be programmed to provide thoroughly convincing simulations of any real-life experience one might choose. Suppose you were offered the chance to plug into the experience machine for the rest of your life. The machine would be programmed to make you instantly forget that you had chosen to plug in and then would provide whatever sequence of experiences you would find maximally pleasurable and minimally painful. Of course, your beliefs about the type of life you are leading will be false. If you choose to plug into the experience machine, you will live your life strapped to a laboratory table with tubes and electrodes sticking into your body. You'll never *do* anything; however, the level of pleasure you'll experience, thanks to the machine, will be extraordinary.

If you were offered the chance to plug into the experience machine for the rest of your life, what would you do? Your first reaction might be to doubt that the machine will perform as promised; certainly no machine now on the market can deliver what this machine is said to be able to do, and this will remain true at least for the foreseeable future. However, for the sake of argument, try to set this hesitation to one side. Imagine yourself being offered the chance to plug in, and suppose that the machine will work as described. My guess is that many people, perhaps including yourself, would decline the opportunity of plugging in.

This fact about people seems to refute hedonism. Apparently, many people prefer to have a real life over a simulated one, even if real life brings less pleasure and more pain than the life they'd have if they plugged into the machine. It seems that people care irreducibly about how they are related to the world outside their own minds; it is false that the only things they care about as ends in themselves are pleasant states of consciousness.

Can hedonism explain why many people would decline the offer to plug into the machine? To see whether this is possible, we need to map out the sequence of events that will comprise your life if you choose to plug into the experience machine and the sequence of events that will occur if you do not. In both cases, the process begins with deliberation, which terminates in a decision. If you decide to plug in, there is a time lag between your decision and your actually being connected to the machine. The two time lines we need to consider are detailed in figure 7.1.

Figure 7.1



The four letters in these two time lines represent how pleasant your experiences will be during different temporal periods, depending on what you decide. If you choose to plug into the machine, you'll have an immense level of *bliss* (*b*) after you plug in. This will dwarf the amount of pleasure you'll experience in the same period of time if you decide not to plug in and to lead a normal life instead; $b > n_2$. If this were the only consideration involved, the hedonist would have to predict that people will choose to plug into the machine. How can hedonism explain the fact that many people make the opposite decision?

The hedonist's strategy is to look at earlier events. If you decided to plug into the machine, how would you feel before you were actually connected? Presumably, you would experience a great deal of *anxiety* (*a*). You'd realize that you were about to stop leading a real life. You will never again see the people you love; all of your projects and plans are about to be terminated. It is clear that you would have less pleasure during this period of time than you would if you rejected the option of plugging into the machine and continued with your real life instead; $a < n_2$.

If hedonists are to explain why people choose not to plug into the experience machine, and are to do this by considering just the pleasure and pain that subjects expect to come their way *after* they decide what to do, the claim must be that $a + b < n_1 + n_2$. Since b is far greater than n_2, this inequality will be true only if a is far far smaller than n_1. That is, hedonists seem compelled to argue that people reject the option of plugging in because the amount of pain they would experience between deciding to plug in and actually being connected to the machine is *gigantic* – so large that it dwarfs the pleasure they'd experience after they are connected.

This suggestion is not plausible. The period of time between deciding to plug in and actually doing so can be made very brief, compared with the long stretch of years you'll spend attached to the machine and enjoying a maximally pleasurable

ensemble of experiences. I grant that people who decide to connect to the machine will experience sadness and anxiety during the brief interval between deciding to plug in and actually plugging in. But the idea that this negative experience swamps all subsequent pleasures just isn't credible.

To see why, let's consider a second thought experiment, suggested to me by William Talbott. Suppose you were offered $5,000 if you went through 10 seconds of a certain experience. The experience is believing that you had just decided to spend the rest of your life plugged into an experience machine. After your 10 second jolt of this experience, you will return to your normal life and will realize that you just had a "nightmare"; you then will receive the money as promised. I expect that many people would choose the 10 seconds just described because it will earn them $5,000. This shows that hedonism is mistaken if it claims that the experience of believing you will be plugged into an experience machine for the rest of your life is so horrible that no one would ever choose a life that included it.

The hedonist still has not been able to explain why many people would choose a normal life over a life plugged into the experience machine. The reason is that a hedonistic calculation seems to lead inevitably to the conclusion that $a + b > n_1 + n_2$. Does this mean that the hedonist must concede defeat? I think that the hedonist has a way out. Quite apart from the amount of pleasure and pain that accrues to subjects *after* they decide what to do, there is the level of pleasure and pain arising in the deliberation process itself. The hedonist can maintain that *deciding* to plug into the machine is so aversive that people almost always make the other choice. When people deliberate about the alternatives, they feel bad when they think of the life they'll lead if they plug into the machine; they feel much better when they consider the life they'll lead in the real world if they decline to plug in. The *idea* of life attached to the machine is painful, even though such a life would be quite pleasurable; the *idea* of real life is pleasurable, even though real life often includes pain. This hedonistic explanation of why people refuse to plug in exploits the distinction that Schlick (1939) drew between the pleasant idea of a state and the idea of a pleasant state.

To see what is involved in this suggestion, let's consider in more detail what goes through people's minds as they deliberate. They realize that plugging in will mean abandoning the projects and attachments they hold dear; plugging into the machine resembles suicide in terms of the utter separation it effects with the real world. The difference is that suicide means an end to consciousness, whereas the experience machine delivers (literally) escapist pleasures. Hedonism is not betraying its own principles when it claims that many people would feel great contempt for the idea of plugging in and would regard the temptation to do so as loathsome. People who decline the chance to plug in are repelled by the idea of narcissistic escape and find pleasure in the idea of choosing a real life.

One virtue of this hedonistic explanation is that it explains the results obtained in both the thought experiments described. It explains why people often *decline* to plug into the experience machine for the rest of their lives; it also explains why people offered $5,000 often *agree* to have ten seconds of the experience of believ-

ing that they have just decided to plug into the machine for the rest of their lives. In both cases, deliberation is guided, not so much by beliefs about which actions will bring *future* pleasure, but by the pleasure and pain that accompany certain thoughts *during the deliberation process itself.*

The problem of the experience machine resembles the problem of the soldier in the foxhole, discussed earlier. How can hedonism explain this act of suicidal self-sacrifice, if the soldier believes that he will not experience anything after he dies? The hedonist can suggest that there is a self-directed benefit that accrues *before* the act of self-sacrifice is performed. It is no violation of hedonism to maintain that the soldier decides to sacrifice his life because that decision is less painful than the decision to let his friends die. The problem of suicidal self-sacrifice and the problem posed by the experience machine can be addressed in the same way.

6 Burden of Proof

Philosophers sometimes maintain that a common sense idea should be regarded as innocent until proven guilty. That is, if a question is raised about whether some common sense proposition is true, and no argument can be produced that justifies or refutes it, then the sensible thing to do is to keep on believing the proposition. Put differently, the idea is that the burden of proof lies with those who challenge common sense.

This general attitude sometimes surfaces in discussion of egoism and altruism. The claim is advanced that the egoism hypothesis goes contrary to common sense. The common sense picture of human motivation is said to be pluralistic – people care about themselves, but also care about others, not just as means, but as ends in themselves. The conclusion is then drawn that if philosophical and scientific argumentation for and against egoism is indecisive, then we should reject egoism and continue to accept pluralism.

One objection to this proposed tie-breaker is that it is far from obvious that "common sense" is on the side of motivational pluralism rather than egoism. What is common sense? Isn't it just what people commonly believe? If so, it is arguable that egoism has made large inroads; it now seems to be a view that is endorsed by large numbers of people. Philosophers need to be careful not to confuse common sense with what they themselves happen to find obvious. As far as I know, no empirical survey has determined whether a pluralistic theory of motivation is more popular than psychological egoism.

Regardless of what people commonly believe about psychological egoism and motivational pluralism, I reject the idea that conformity with common sense is a tie-breaker in this debate. It does not have this status in physics or biology, and I see no reason why it should do so when the question happens to be philosophical or psychological in character. In fact, it is arguable that our intuitions in this domain are especially prone to error. People have a picture of their own motives and the mo-

tives of others. If certain types of self-deception – either regarding one's own motives or those of others – were advantageous, then evolution might have enshrined these falsehoods in the set of "obvious" propositions we call common sense. A philosophy informed by an evolutionary perspective has no business taking common sense at face value.

7 Parsimony

I so far have argued that hedonism and has not been refuted by philosophical arguments or by observed behavior; if this is right, then egoism has not been refuted either. This does not mean that egoism is true; after all, motivational pluralism has not been refuted either. In the light of this impasse, it is worth noting that social scientists often implicitly assume that if a behavior *can* be explained in egoistic terms, then it *ought* to be so explained. The fact that they have no direct argument in favor of this position seems not to be relevant. And the fact that the behavior also can be explained in terms of motivational pluralism also seems not to be relevant. However, why should egoism be the default hypothesis that we should assume is true unless we are forced to abandon it?

One answer to consider is that egoism is more parsimonious – it postulates only one type of ultimate motive, whereas pluralism postulates two (Hume 1970 [1751]; Batson 1991). Even if we assume that parsimony marks not just an aesthetic difference between theories, but a reason for finding some theories more plausible than others, there still is a defect in this defense of egoism. The problem is that egoism is *less* parsimonious than pluralism when we consider how many causal beliefs the two theories postulate. When Sally wants Otto to do well, the defender of egoism counts this as an instrumental desire while the proponent of motivational pluralism may hold that Sally has this other-directed desire as an end in itself. But notice that the egoistic explanation attributes to Sally a causal belief – *that she stands to receive a benefit from Otto's doing well.* Motivational pluralism is not committed to saying that Sally has this belief. An egoist has a shorter list of ultimate desires than a pluralist, but the egoist has a longer list of causal beliefs. For this reason, it is quite unclear why psychological egoism should be regarded as the more parsimonious theory (Sober and Wilson 1998).

8 An Evolutionary Approach

Psychological motives are proximate mechanisms in the sense of that term used in evolutionary biology. When a sunflower turns towards the sun, there must be some mechanism inside the sunflower that causes it to do so. Hence, if phototropism is an adaptation that evolved because it provided organisms with certain benefits, then a proximate mechanism that causes that behavior also must have evolved. Simi-

larly, if certain forms of helping behavior in human beings are evolutionary adaptations, then the motives that cause those behaviors in individual human beings also must have evolved. Perhaps a general perspective on the evolution of proximate mechanisms will throw light on the specific problem of whether egoism or motivational pluralism was more likely to have evolved.

Pursuing this evolutionary approach does not presuppose that every detail of human behavior, or every act of helping, can be completely explained by the hypothesis of evolution by natural selection. Doubtless there are many facts about behavior and many instances of helping for which natural selection is not a relevant explanation. However, I want to consider a single fact about human behavior, and my claim is that selection is relevant to explaining it. The phenomenon of interest is that human parents take care of their children; the average amount of parental care provided by human beings is strikingly greater than that provided by parents in many other species. I will assume that natural selection is at least part of the explanation of why parental care evolved in our lineage. This is not to deny that human parents vary; some parents take better care of their children than others, and some even abuse and kill their offspring. Another striking fact about individual variation is that mothers, on average, expend more time and effort on parental care than fathers. Perhaps there are evolutionary explanations for these individual differences as well; the question I want to address here, however, makes no assumption as to whether this is true.

To tease out some general principles that govern how one might predict the proximate mechanism that will evolve to cause a particular behavior, I'll switch examples to a hypothetical mindless organism whose problem is to select items from its environment to eat. Some particles that float by in the liquid medium in which the organism lives contain protein; others contain poison. The organism has evolved a particular behavior – it tends to eat protein and avoid poison. What proximate mechanism might have evolved that allows it to do so?

First let's survey the range of possible design solutions that we need to consider. The most obvious design solution to this problem is for the organism to have a detector that distinguishes protein from poison. It captures a morsel that floats by, puts the particle in its detector, and then has the output of this detector wired to a behavior; the organism either eats the morsel, or spits it out. I'll call this the *direct* solution to the design problem; the organism needs to discriminate between protein and poison and this solution accomplishes that end by using a detector that detects that very contrast in properties.

It isn't hard to imagine other solutions to the design problem that are less direct. Suppose that protein tends to be red and that poison tends to be green in the organism's environment. If so, the organism could use a color detector to make the requisite discrimination. This design solution is *indirect*; the organism needs to distinguish protein from poison and accomplishes this by discriminating between two other properties that happen to be correlated with the target contrast. In general, there may be many indirect design solutions that the organism might exploit; there are as many indirect solutions as there are correlations between the protein/

poison distinction and other properties found in the environment. Finally, we may add to our list the idea that there can be pluralistic solutions to a design problem. In addition to the monistic solution of having a protein detector and the monistic solution of having a color detector, an organism might deploy both a protein detector *and* a color detector.

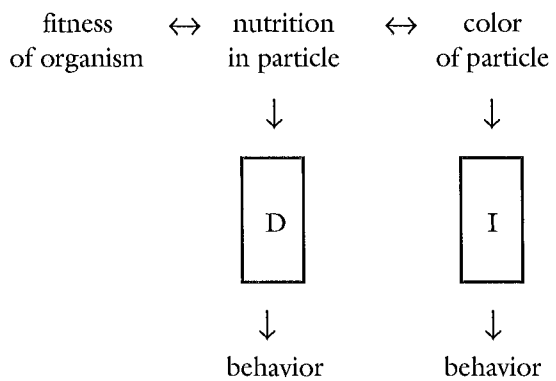
Given this multitude of possibilities, how might one predict which of them will evolve? Three principles are relevant here – *availability*, *reliability*, and *efficiency* (Sober 1994; Sober and Wilson 1998).

Natural selection acts only on the range of variation that exists ancestrally. A protein detector might be a good thing for the organism to have, but if that device was never present as an ancestral variant, then natural selection cannot cause that trait to evolve. So the first sort of information we'd like to have concerns which proximate mechanisms were *available* ancestrally.

Let's suppose for the sake of argument that both a protein detector and a color detector are available ancestrally. Which of them is more likely to evolve? Here we need to address the issue of *reliability*. Which device does the more reliable job of indicating which particles in the environment are good to eat? Without further information, not much can be said. A color detector may have any degree of reliability, and the same is true of a protein detector. There is no *a priori* reason why the direct strategy should be more or less reliable than the indirect strategy. However, there is a special circumstance in which they will differ. It is illustrated by figure 7.2.

The double arrows indicate correlation; gaining nutrition is correlated with an organism's fitness, and a particle's being red rather than green is correlated with its nutritional content (figure 7.2). In the diagram, there is no arrow from fitness to color except the one that passes through nutrition. This means that an organism's fitness is correlated with the color of the particles that it eats. There is no *a priori* reason that color should be relevant to fitness only by virtue of indicating nutritional content. For example, if eating red particles attracted predators more than

Figure 7.2



eating green ones does, then color would have two sorts of relevance for fitness. However, if nutrition "screens off" fitness from color in the way indicated, we can state the following principle about the reliability of the direct device D and the indirect device I:

(D/I) If nutrition and color are less than perfectly correlated, and if D detects nutrition at least as well as I detects color, then D will be more reliable than I.

This is the Direct/Indirect Asymmetry Principle. Direct solutions to a design problem aren't always more reliable, but they are more reliable in this circumstance.

A second principle about reliability also can be extracted from this diagram. Just as scientists do a better job discriminating between hypotheses if they have more evidence rather than less, so it will be true that organisms make more reliable discriminations if they have two sources of information about what to eat rather than just one:

(TBO) If nutrition and color are less than perfectly correlated, and if D and I are each reliable, though fallible, detectors of fitness, then D and I working together will be more reliable than either of them working alone.

This is the Two-is-Better-than-One Principle. It requires an assumption – that the two devices do not interfere with each other when they are both present in an organism; they function fairly independently.

The D/I Asymmetry and the TBO Principle pertain to the issue of reliability. Let us now turn to the third consideration that is relevant to predicting which proximate mechanism will evolve, namely *efficiency*. Even if a nutrition detector and a color detector are both available, and even if the nutrition detector is more reliable, it doesn't follow that natural selection will favor the nutrition detector. It may be that a nutrition detector requires more energy to build and maintain than a color detector. Organisms run on energy no less than automobiles do. Efficiency is relevant to a trait's overall fitness just as much as its reliability is.

With these three considerations in hand, let's return to the problem of predicting which motivational mechanism for providing parental care is likely to have evolved in the lineage leading to human beings. The three motivational mechanisms we need to consider correspond to three different rules for selecting a behavior in the light of what one believes:

(HED) Provide parental care if and only if doing so will maximize pleasure and minimize pain.

(ALT) Provide parental care if and only if doing so will advance the welfare of one's children.

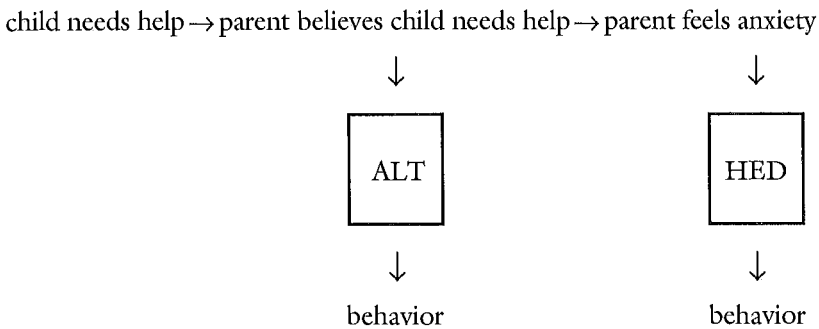
(PLUR) Provide parental care if and only if doing so will either maximize pleasure

and minimize pain, or will advance the welfare of one's children.

(ALT) is a relatively direct, and (HED) a relatively indirect, solution to the design problem of getting an organism to take care of its offspring. Just as an organism can find nutrition by detecting color, so it is possible in principle for a hedonistic organism to be built in such a way that it will provide parental care; what is required is that the organism be so constituted that providing parental care is the thing that maximizes its pleasure and minimizes its pain (or that the organism at least believes that this is so).

Let's consider how reliable these three mechanisms will be in a certain situation. Suppose that a parent learns that its child is in danger. Imagine that your neighbor tells you that your child has just fallen through the ice on a frozen lake. Figure 7.3 shows how (HED) and (ALT) will do their work.

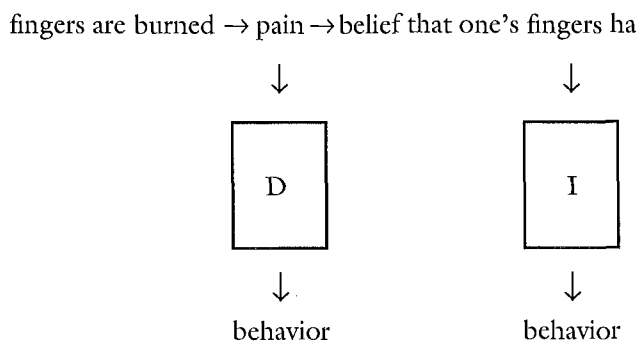
Figure 7.3



The altruistic parent will be moved to action just by virtue of believing that its child needs help. The hedonistic parent will not; rather, what moves the hedonistic parent to action is the feelings of anxiety and fear that are caused by the news, or the parent's belief that such negative feelings will continue unless the child's situation is improved. It should be clear from figure 7.3 that the (D/I) Asymmetry Principle applies. In the circumstance specified, (ALT) will be more reliable than (HED). And by the (TBO) Principle, (PLUR) will do better than both. In this example, hedonism comes in last in the three-way competition, at least as far as reliability is concerned.

The important thing about this example is that the feelings that the parent has are *belief mediated*. The only reason the parent feels anxiety and fear is that the parent *believes* that its child is in trouble. This is true of many of the situations that egoism and hedonism are called upon to explain, but it is not true of all of them. For example, consider the following situation in which pain is a direct effect, and belief a relatively indirect effect, of bodily injury (figure 7.4).

Figure 7.4



Now hedonism is a direct solution to the design problem; it would be silly to build the organism so that it is unresponsive to pain and withdraws its fingers from the flame only after it forms a belief about bodily injury. In this situation, *belief is pain-mediated* and the (D/I) Asymmetry Principle explains why a hedonistic focus on pain makes sense. However, the same principle indicates what is misguided about hedonism as a design solution when pain is belief-mediated, which is what occurs so often in the context of parental care.

If hedonism is less reliable than either pure altruism or motivational pluralism, how do these three mechanisms compare when we consider the issues of evolutionary availability and efficiency? With respect to availability, I want to make this claim: *if hedonism was available ancestrally as a design solution, so was altruism.* The reason is that the two motivational mechanisms differ in only a very minor way. Both require a belief/desire psychology. And both the hedonistic and the altruistic parent want their children to do well; the only difference is that the hedonist has this propositional content as an instrumental desire while the altruist has it as an ultimate desire. If altruism and pluralism did not evolve, this was not because they were unavailable as variants for selection to act upon.

What about the question of efficiency? Does it cost more calories to build and maintain an altruistic or a pluralistic organism than it does to build and maintain a hedonist? I don't see why. What requires energy is building the hardware that implements a belief/desire psychology. However, it is hard to see why having one ultimate desire rather than two should make an energetic difference; nor is it easy to see why having the ultimate desire that your children to do well should require more calories than having the ultimate desire to avoid pain and attain pleasure. People with more beliefs apparently don't need to eat more than people with fewer. The same point seems to apply to the issue of how many, or which, ultimate desires one has.

In summary, hedonism is a less reliable mechanism than pure altruism or pluralism as a device for delivering parental care. And, with respect to the issues of availability and efficiency, we found no difference among these three motivational

mechanisms. This suggests that natural selection is more likely to have made us motivational pluralists than to have made us hedonists.

From an evolutionary point of view, hedonism is a very bizarre motivational mechanism. What matters in the process of natural selection is an organism's ability to survive and be reproductively successful. Reproductive success involves not just the production of offspring, but the survival of those offspring to reproductive age. So what matters is the survival of one's own body and the bodies of one's children. Hedonism, on the other hand, says that organisms care ultimately about the states of their own consciousness, and about that alone. Why would natural selection have led organisms to care about something that is peripheral to fitness, rather than have them set their eyes on the prize? If organisms were unable to conceptualize propositions about their own bodies and the bodies of their offspring, that might be a reason. After all, it can make sense for an organism to exploit the indirect strategy of deciding what to eat on the basis of color rather than on the basis of nutritional value, if the organism has no epistemic access to nutritional content. But if an organism is smart enough to form representations about itself and its offspring, this justification of the indirect strategy will not be plausible. The fact that we evolved from ancestors who were cognitively less sophisticated makes it unsurprising that avoiding pain and attaining pleasure are two of our ultimate goals. But the fact that human beings are able to form representations with so many different propositional contents suggests that evolution supplemented this list of what we care about as ends in themselves.

9 Concluding Comments

I have argued that past philosophical and psychological attempts to resolve the debate between egoism and motivational pluralism have not succeeded. It would be astonishing if this dispute about an apparently empirical matter could be resolved by arguments *a priori*. And, unfortunately, the observations that people casually make in ordinary life and that scientists make in the laboratory have not been decisive either; although some simple versions of egoism are refuted by what we observe, other versions of egoism can be constructed that seem to fit the available observations. Perhaps more sophisticated experiments and observations of behavior will answer the question. But for now, the situation in philosophy and psychology is one of stalemate.

Can evolutionary considerations break through this impasse? The argument of the previous section aims to establish that a purely egoistic set of motives is less likely to have evolved than a set of motives that includes both egoistic and altruistic ultimate desires. I do not suggest that this argument *proves* that people are motivational pluralists; there is much that remains unknown about the mind and how it evolved, and there is no guarantee that further details will not substantially alter the picture I have tried to develop. However, I do think that the argument suffices to

show that egoism does not deserve to be regarded as the default hypothesis that we should accept as long as it is consistent with what we observe. In my opinion, the weight of evidence favors pluralism, if only to a small degree.

References

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

Broad, C. D.: *Five Types of Ethical Theory* (Totowa, NJ: Littlefield, Adams, 1965).

Butler, J.: Fifteen Sermons upon Human Nature (1726); Reprinted in L. A. Selby-Bigge (ed.), *British Moralists: Being Selections From Writers Principally of the Eighteenth Century*, volume 1 (New York: Dover Books, 1965), pp. 180–241.

Clark, R. D. and Word, L. E. "Where is the apathetic bystander? Situational characteristics of the emergency," *Journal of Personality and Social Psychology*, 29 (1974): 279–87.

Feinberg, J.: "Psychological egoism," in S. Cahn, P. Kitcher, and G. Sher (eds.), *Reason at Work* (San Diego, CA: Harcourt Brace and Jovanovich, 1984), pp. 25–35.

Gibbard, A.: *Wise Choices, Apt Feelings* (Cambridge: Harvard University Press, 1990).

Hume, D. "On self love," in *An Enquiry Concerning the Principles of Morals* (1751); (Indianapolis: Hackett, 1970).

Kavka, G.: *Hobbesian Moral and Political Theory* (Princeton, NJ: Princeton University Press, 1986).

LaFollette, H.: "The truth in psychological egoism," in J. Feinberg (ed.), *Reason and Responsibility*, 7th edn. (Belmont, CA: Wadsworth, 1988), pp. 500–7.

Nagel, T.: *The Possibility of Altruism* (Oxford: Oxford University Press, 1970).

Nozick, R.: *Anarchy, State, and Utopia* (New York: Basic Books, 1974).

Schlick, M.: *Problems of Ethics* (New York: Prentice Hall, 1939).

Schroeder, D., Penner, L., Dovidio, J. and Piliavin, J.: *The Psychology of Helping and Altruism* (New York: McGraw Hill, 1995).

Sidgwick, H.: *The Methods of Ethics* (1907) 7th edn. (London: Macmillan, 1922).

Sober, E.: "Hedonism and Butler's stone," *Ethics* 103 (1992): 97–103.

——: "Did evolution make us psychological egoists?" In *From a Biological Point of View: Essays in Evolutionary Philosophy* (New York: Cambridge University Press, 1994), pp. 8–27.

—— and Wilson, D.S.: *Unto Others – the Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press, 1998).

Stampe, D.: "Desire," in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind* (Cambridge, MA: Basil Blackwell, 1994), pp. 244–50.

Stewart, R. M.: "Butler's argument against psychological hedonism," *Canadian Journal of Philosophy* 22 (1992): 211–21.