Unto Others

The Evolution and Psychology of Unselfish Behavior

Elliott Sober

David Sloan Wilson

Harvard University Press
Cambridge, Massachusetts
London, England

7

Three Theories of Motivation

The present chapter will map out three psychological theories of motivation—hedonism, egoism, and altruism. We intend to proceed carefully, since evaluation of these theories is often short-circuited by biased definitions and spurious arguments. Just as it is easy to solve the evolutionary problem of altruism by defining "selfishness" as *whatever evolves,* so it is easy to solve the psychological problem of motivation by defining "self-interest" as *whatever people want.* More subtle biases can and do enter discussion of the psychological issue. The care we take here in formulating our hypotheses will pay dividends in subsequent chapters when we examine different attempts to resolve the motivational question.

In the course of delimiting the three motivational theories, we will have to address a variety of related issues. How is altruism related to morality? How is it related to the emotions of empathy and sympathy? Does egoism assume that agents always rationally calculate what is in their best interests? These details are important to our project because they help specify exactly what the traits are that need to be investigated. Before we can evaluate whether people ever have altruistic ultimate motives, or ask what evolutionary theory has to say about this motivational question, we must have a clear view of the phenotypes that require analysis.

Defining Hedonism

The discussion at the end of the last chapter of what it means for a desire to be ultimate or instrumental makes it easy to define hedonism. Hedonism says that the only ultimate desires that people have are the desires to obtain pleasure and avoid pain. All other desires are purely instrumental with respect to these two ends. Construed in this way, hedonism is a descriptive theory, not a normative one; it does not recommend actions or say whether it is good or bad that people are as they are. The theory merely attempts to describe how the mind is structured.[1]

When referring to our aversion to pain, hedonism uses the term *pain* quite inclusively. In ordinary parlance, there are many aversive sensations besides pain—nausea, dizziness, anxiety, depression, and a host of others. The hedonist has no trouble with the idea that avoiding nausea can be an end in itself. Yet, it may sound a little odd to say that nausea is a type of pain. The solution is to understand the term *pain* as encompassing all aversive sensations. Any sensation that a person dislikes experiencing the hedonist will dub an instance of "pain."

Similar remarks pertain to the word *pleasure*. If pleasure is a sensation, it is very hard to say which particular sensation it is. People "take pleasure" in a variety of experiences. One can enjoy the taste of a peach, but also be pleased to learn that others are doing well (or ill). In what sense do these two experiences "feel the same"? Hedonism need not insist that they involve a single type of sensation. Both can count as instances of pleasure, if "pleasure" names any experience that a person enjoys having (Sidgwick 1907).

The distinctive feature of hedonism is that it says that ultimate desires are always *solipsistic*. What we ultimately care about is limited to states of our own consciousness; what goes on in the world outside the mind is of instrumental value only.

Defining Egoism

Egoism maintains that the only ultimate goals an individual has are *self-directed;* people desire their own well-being, and nothing else, as an end in itself. If you care about the well-being of others, this is only because you think the well-being of others is instrumentally related to benefiting yourself. Strictly speaking, the ultimate desires postulated by egoism say nothing about the situation of others. Malevolence is as alien to the egoist's basic outlook as benevolence is (Butler 1726).

Egoism's use of the term *self-directed* requires clarification, but two properties of the egoistic theory should be clear from the outset. First, an irreducible concern for the welfare of others is incompatible with egoism. Second, an irreducible concern for obtaining pleasure and avoiding pain is quite consistent with egoism. In other words, the altruism hypothesis is incompatible with psychological egoism, whereas hedonism is a species of egoism. Although all hedonists are egoists, not all egoists are hedonists. An egoist may have the desire to accumulate money as an end in itself, but a hedonist may not. The same would be true of the ultimate goal of climbing Mount Everest. Egoists need not have states of their own consciousness as the only things they care about as ends in themselves.

To determine whether a desire is "self-directed," we must attend to the propositional content that the desire has. If Sam wants to eat an apple, this is a self-directed desire, because the proposition *Sam eats an apple* mentions Sam but no one else. Similarly, if Sam wants Aaron to eat an apple, this is an other-directed desire, because the proposition *Aaron eats an apple* mentions Aaron but not Sam himself. An egoistic ultimate desire is self-directed; an altruistic ultimate desire is other-directed.

This construal of egoism encounters problems when we consider desires that mention both self and other. Suppose that Aaron wants to be famous, not as a means to anything else but as an end in itself. A little reflection on what "famous" means indicates that the content of Aaron's desire involves a relation between self and other; Aaron wants others to know who he is. Even though Aaron's desire isn't purely self-directed, it may sound odd to conclude that Aaron is not an egoist.[2] A parallel difficulty arises if we define altruism as the claim that some of our ultimate desires are purely other-directed. Suppose Sam wants the apples to be distributed equally between himself and Aaron, not as a means to some further end but as an end in itself. Even though Sam's desire isn't purely other-directed, it may sound odd to conclude that this desire is not altruistic.

These problems might disappear if the desires just mentioned were merely instrumental. If Aaron wants to be famous only because he

thinks this will provide him with pleasurable experiences, then he is an egoist. Similarly, if Sam wants the apples to be distributed equally in part because he has the ultimate goal that Aaron's situation be improved, then he is an altruist. Unfortunately, these suggestions evade the problem posed. How should we categorize desires that have relational facts as their propositional contents *when these desires are ultimate?*

To cram this variety of ultimate desire into egoism, or into altruism, would be difficult to justify, and the attempt to do so might appear to bias the case in favor of one position or the other. In consequence, we propose to add a third category to egoism and altruism; *relationism* is the view that people sometimes have ultimate desires that certain relational propositions (connecting self and specific others) be true. If the reader thinks that some cases of relationism are properly viewed as subspecies of altruism or of egoism, we invite the reader to adjust the conceptual taxonomy we are suggesting. Our assessment of these theories will not be affected by such amendments.[3]

We have defined egoism so that the distinction between conscious and unconscious desires plays no role. An individual whose ultimate desires are *conscious* and an individual whose ultimate desires are *unconscious* will both be egoists if those ultimate desires are all directed solely toward self-benefit. The fact that individuals sincerely want to help others, and do not consciously experience this desire as involving a sacrifice or a conflict with their authentic selves, does not tell us what their ultimate motives really are. There is nothing in the egoism hypothesis that prohibits other-directed desires from being fully integrated into the agent's personality. Those desires must be instrumental, but people need not experience them as alien intrusions.

We hope it is obvious that the egoism hypothesis we have described is not the same as the view that might be called "vulgar egoism." Vulgar egoism maintains that people are moved solely by the goal of securing *material* benefits. We think it is obvious that this version of egoism is too narrow; the desire for material benefit is *one* motive that people have, and so it helps explain *some* aspects of human behavior. But there is much that it cannot explain. Egoism, unlike vulgar egoism, deploys a wider notion of self-benefit, one that includes internal (psychological) payoffs as well as external (material) ones.

Egoism is sometimes criticized for viewing happiness as a one-dimensional state (LaFollette 1988), but this criticism does not apply to the version of the theory we have described. If someone wants to discover a cure for cancer, to climb Mount Everest, and to experience the euphoria of romantic infatuation, the egoist need not pretend that these three goals somehow boil down to the same thing. Each of these desires is self-directed; if they exhaust what the agent's ultimate desires are, then the agent is an egoist, regardless of whether these desires reduce to a deeper unity. Symmetrically, the altruism hypothesis is not committed to the idea that people think of the welfare of others as a simple and one-dimensional matter.

We have one last comment on our definition of egoism. It is better to describe egoism as holding that all ultimate ends are self-directed than to say that it views all ultimate ends as "selfish" (Henson 1988). When Jim wants his tooth to stop aching, it is misleading to say that Jim is being selfish in feeling this way. The term *selfish* carries with it overtones of disapproval. This is not, strictly speaking, what egoism maintains; egoism is a descriptive, not a normative, theory.

Short-term and Long-term Egoism

Consider Ronald. He is now deliberating about whether he will continue to smoke cigarettes. He realizes that any single cigarette will have only a negligible impact on health; but over the long run, smoking many cigarettes may well have a devastating effect. Ronald is all too cognizant of the pleasure he gains from each and every cigarette. As he thinks about whether to stop smoking altogether, his mind drifts to a more immediate question. He holds an unlit cigarette before him and considers whether he should light it and smoke. According to Ronald's way of seeing things, smoking produces a short-term benefit. But he also realizes that over the long term smoking is apt to impose a substantial cost. If Ronald cares only about the here and now, he will light up. If he cares enough about the long-term quality of his life, he will stop smoking, starting today.

Our interest here is not to predict what Ronald will do but to ask what egoism says about this problem. Notice that the short-term and the long-term considerations that weigh with Ronald are both self-directed. Wanting the pleasure produced by nicotine is a self-directed

desire, but so is the desire to be healthy and long-lived. Although egoism says that our ultimate desires are always self-directed, it does not say how much importance people assign to present versus future benefits. Some may see this lack of specificity as a defect in the egoistic theory; this is a criticism that we will analyze in Chapter 9. For now, we merely register our opinion that this flexibility in egoism is unobjectionable. Whatever Ronald does in the situation just described will be consistent with egoism. Egoism does not say which specific desires people have as their ultimate ends; it says merely that they strive for a certain *type* of ultimate goal.[4]

Defining Altruism

The altruism hypothesis maintains that people sometimes care about the welfare[5] of others as an end in itself. Altruists have irreducible other-directed ends.

The word *sometimes* marks a logical difference between the altruism hypothesis and the hypotheses of hedonism and egoism. Hedonism and egoism are claims about *all* the ultimate desires an individual has, whereas altruism makes no such universal claim. Egoism says that all ultimate desires are self-directed, but the theory we are calling altruism does not say that all ultimate desires are other-directed. Of course, it is possible to construct a monolithic theory of this type, but no one would believe for a moment that it is true. Rather, we should construe altruism as part of a *pluralistic* theory of motivation that maintains that people have ultimate desires about others as well as about themselves. Egoism and hedonism, on the other hand, are rightly understood as (relatively) *monistic* doctrines.[6]

The thesis of altruism, as we understand it, says that some people at least some of the time have the welfare of others as ends in themselves. This does not entail that *most* people are altruistic *all* the time, or that *some* people are altruistic *most* of the time, or that the altruism that people sometimes experience is especially *strong*. A person who is prepared to make a very small sacrifice in order to provide a huge benefit for someone else may be altruistic, but this person will be less altruistic than someone who is prepared to make a larger sacrifice. Our version of altruism is quite compatible with the existence of widespread selfishness (a point to which we will return).

It might be suggested that the altruism hypothesis, so construed, is too modest a theory to be of much interest. If the theory claims merely that people sometimes have irreducibly altruistic motives, but says nothing about how strong or pervasive those motives are, why is it worth discussing? To be sure, there is more to the psychology of altruism than the altruism hypothesis we have just identified. However, we believe that this hypothesis is fundamental because more ambitious claims about the importance of altruism are committed to this modest thesis. In addition, this apparently modest claim is precisely what psychological egoism denies—it is the very nub of the issue.

Who are the others about whom altruistic individuals have ultimate concerns? The most obvious examples involve desires that focus on the welfare of another *person*. But consider people who care irreducibly about "the environment," meaning the well-being of the entire earth (both living and nonliving). Is this altruism? And what about people who care about a nation, a religion, an ethnic group, or a cultural tradition, not just as means but as ends in themselves?[7] True, such concerns sometimes count as "selfless." But are they altruistic? Although we will concentrate on the altruistic regard that human beings may have for other human beings, there is no reason to rule out these other candidates. The principal point about altruism, as we understand it, is that it attributes to people ultimate desires concerning the welfare of individuals other than themselves. We are prepared to be quite liberal concerning what might count as an "individual." This decision won't affect the main arguments we will advance in what follows; readers are invited to construe altruism more narrowly if they so wish.

The altruism hypothesis says that we have other-directed ultimate desires, whereas psychological egoism says that all of our ultimate desires are self-directed. But egoism and altruism are usually understood to involve more than this. For example, if Iago views the destruction of Othello as an end in itself, then one of Iago's ultimate aims is other-directed. Nonetheless, it would be odd to call Iago an altruist, since his other-directed desire is *malevolent*. Similarly, people whose only ultimate aim is to harm or destroy themselves would not normally be counted as egoists, since what they want for themselves is not their own good. This is why our definitions of egoism and altruism go beyond the distinction between self-directed and other-

directed ultimate desires. Egoists ultimately desire only what they think will be good for themselves; altruists have ultimate desires concerning what they think will be good for others.

The benevolent intentions associated with altruism can take two forms. An altruist may want others to have what they actually want for themselves; alternatively, an altruist may want for others something they have never thought of, or have considered and rejected. If Sheila buys Oscar a particular book simply because Oscar has been wanting that book, we may have a case of the first type. If Stanley wants Olivia to take her medicine even though she does not want to do so, we may have a case of the second.

The concept of altruism is sometimes restricted to cases in which an individual helps others without any expectation of receiving an *external* benefit, such as money or power (Macaulay and Berkowitz 1970); this definition entails that individuals are altruistic when they help just because they think that helping will make them feel good. We disagree with this way of defining the concept. To see why, consider a heroin addict whose every action is ultimately aimed at securing the pleasant states of consciousness that the drug produces. The addict, as so far described, is a hedonist. But now let us perform a thought experiment. Let us place this person in an environment in which the only way to get the drug is by helping people. This is very different from the real world that addicts usually inhabit, but the hypothetical situation is worth considering for the light it sheds on the conceptual issue. An addict who helps others only because the effect of helping is a drug-induced euphoria is not thereby an altruist. The same point applies to people in the real world who do not take heroin; if they are "hooked" on helping because of the pleasure that helping affords and the pain it allows them to avoid, their actions do not make them altruists. We must not muddy the waters by treating altruism as a form of hedonism.

Another definition of altruism also merits comment. Altruism is sometimes defined by saying that individuals have the ultimate desire that other people's desires be satisfied. This resembles a formulation used in the social sciences according to which an altruist is someone whose utility function "reflects" the utility functions of others. These definitions entail that altruists must have representations of the mental states of others. We believe that this formulation of the theory is stronger than it should be. Suppose Stanley wants Olivia to take the medicine simply because Stanley believes that it would be good for

her. Stanley is not thinking about Olivia's *desires* but about her *health*. Perhaps Olivia doesn't want to take the medicine; she is so despondent that she doesn't even want to get well. Nonetheless, Stanley is an altruist because he has Olivia's well-being as an end in itself. Maybe altruists are psychologists, as discussed in Chapter 6; however, we don't think this is true as a matter of definition.

Now that hedonism, egoism, and the pluralistic theory of motivation that embeds the altruism hypothesis have been formulated, let us reflect on how they are logically related. First we note an asymmetry. Hedonism entails egoism, but egoism does not entail pluralism; indeed, egoism and pluralism (which includes altruism) are incompatible. On the other hand, there is a symmetry here, which can be identified by considering what each hypothesis entails about the ultimate motives that might be involved in explaining why an individual performs a particular action. These views are *nested*. Suppose Lois helps someone. Hedonism says that Lois did this because she cares ultimately about the state of her own consciousness, and about nothing else. Egoism grants that this may be *part* of the explanation, but says that it need not be the whole story. According to egoism, Lois helped because she cares ultimately about her own situation, not about the welfare of others. Psychological pluralism grants that this may be *part* of the explanation, but denies that it must be the whole truth. The transition from hedonism to egoism, and from egoism to pluralism, involves canceling restrictions on the set of ultimate desires that might explain an agent's behavior.

Our construal of egoism and altruism entails that these two types of motivation are not exhaustive. Besides ultimate desires for one's own well-being, and ultimate desires concerning the well-being of this or that other individual, there are additional possibilities to consider. We have mentioned *relationism*, which maintains that people ultimately desire that certain relations between self and specific others obtain. Later in this chapter, we will argue that the ultimate desire to uphold a general moral principle should be regarded as neither altruistic nor egoistic.

Empathy, Sympathy, and Personal Distress

Empathy and sympathy are emotions. When they occur, do they trigger altruistic desires? Common sense suggests that they do; empa-

thy and sympathy sometimes elicit helping behavior, and it makes sense to see this behavior as tracing back to the desire to improve the other person's situation. The causal chain seems to be this:

Emotions of empathy and sympathy ⟶ Desire to help ⟶ Helping

Psychologists have reached the same conclusion; see Batson (1991, pp. 93–96) for a review. However, even if empathy and sympathy have these effects, the question remains of whether the resulting desire to help is ultimate or instrumental. Perhaps empathy and sympathy are able to evoke altruistic desires because people don't like experiencing these emotions and therefore wish to do what they can to extinguish them. Thus, the existence of empathy and sympathy does not resolve the debate between psychological egoism and altruism. Nonetheless, it is worth getting clear on what empathy and sympathy are and why they differ from the altruistic desires they sometimes cause.

The term *empathy* entered English in 1909 as E. B. Titchener's translation of *Einfühlung* (Wispé 1987); since then, its meaning has gone through several metamorphoses in different branches of psychology and it has been absorbed into everyday English as well. The term *sympathy* has an older provenance, but it too has been used in different ways and has been expropriated as a term of art in various psychological theories. Although the definitions we will suggest coincide with ordinary and scientific usage in some respects, they depart from that usage in others. Given the fact that both terms have been put to multiple uses, it would be quixotic to expect a single pair of definitions to agree with what each and every person using the terms has meant. Rather, our response to the present Babel of meanings is to try to single out what is fundamental. In any event, the categories we will describe are more important than the labels we will use to name them.

Empathy is sometimes contrasted with sympathy by saying that empathy involves identifying with others, whereas sympathy involves a more detached variety of emotional connection. What does "identification" mean here? The idea is sometimes explained by saying that empathy makes the boundary between self and other disappear. We believe that this is almost always a poetic overstatement (so

does Batson 1991). When Barbara learns that Bob's father has just died, she may empathize with Bob without losing sight of the fact that they are two different people, not one and the same person. As much as Barbara's heart goes out to Bob, Barbara understands perfectly well who it is who has just lost a parent. When people confuse the real misfortunes of others with their own more fortunate situation, we do not praise them for their ability to empathize; empathy is not the inability to keep track of who is who.

An everyday example of what it means to "identify with" another individual is provided when people talk of identifying with a sports team. This doesn't involve the delusion of believing that one is identical with the New York Yankees. Rather, it means regarding one's self as part of a whole to which the team also belongs and caring about the fate of that whole. Good deeds performed by Yankees make one proud, foul deeds make one ashamed. "What they do reflects on me," or so Yankees fans seem to feel. This same pattern of thinking is present in deeper and more pervasive types of identification—with family, clan, ethnicity, nationality, and religion. The "I" is defined by relating it to a "we." Human beings don't simply *belong* to groups; they *identify* with them. This is an important fact about human experience.

Whether or not empathy entails identification, we suggest that empathy involves sharing the emotion of another. Barbara's empathy involves her feeling sad because Bob is sad. Of course, it is perfectly possible that Barbara may empathically connect with one of Bob's emotions, but not with another. Suppose that Bob feels both sad and guilty about his father's death and that Barbara empathizes with the sadness, but not with the guilt, of which she is unaware. No precise degree of similarity between the two individuals' overall emotional states is built into the concept of empathy (Eisenberg and Miller 1987; Eisenberg and Strayer 1987).

Empathy is sometimes said to require "perspective-taking." We agree that when people feel empathy, they typically have some understanding of why others see the world as they do. However, we don't want to build this in as a requirement. If *O* feels terror or sadness, *S* may see that this is so without knowing what it is in *O*'s situation that elicited these emotions. *S* may respond empathically; *S* may "feel for" *O* with the end result that their emotions match. Empathy requires

one to understand *that* the other person is experiencing an emotion; it need not involve a deeper grasp of *why* this is so.

Although empathy involves emotion matching, more is required. Suppose that O feels so depressed and anxious that he is unable to think of anyone but himself; S learns about this, and this information somehow causes S to go into precisely the same state. S is now matching O's emotion, but S is not empathizing with O. The reason is that S is not even thinking about O. It is one thing for S to feel sad, quite another for S to feel sad *for* O. The same distinction can be drawn with respect to other emotions; for example, when S feels frightened, this doesn't necessarily involve S's feeling frightened *for* O. This point about empathy, as well as the other observations we have made, are consolidated in the following definition:

> S empathizes with O's experience of emotion E if and only if O feels E, S believes that O feels E, and this causes S to feel E for O.[8]

What does it mean for one individual to feel a certain emotion "for" another? Here we may exploit the idea that belief involves the formation of representations that have propositional content (Chapter 6). If S feels sad for O, then S forms some belief about O's situation and feels sad that this proposition is true. When Barbara empathizes with Bob, he is the focus of her emotion; she doesn't just feel the same emotion that Bob experiences. Rather, Barbara feels sad *that Bob's father has just died;* she feels sad about what has made Bob sad.

Our definition of empathy does not require that the shared emotion be negative. When O feels sorrow, S can empathize, but the same is true when O feels joy.[9] Herein we find one difference between empathy and sympathy. It sounds odd to talk about someone's sympathizing with another by feeling happy. If S sympathizes with O, then S must feel bad. There is, however, a more important difference. Consider the fact that your heart can go out to someone without your experiencing anything like a similar emotion. This is clearest when people react to the situations of individuals who are not experiencing emotions at all. Suppose Walter discovers that Wendy is being deceived by her sexually promiscuous husband. Walter may sympathize with Wendy, but this is not because Wendy feels hurt and betrayed. Wendy feels nothing of the kind, because she is not aware of her

husband's behavior. It might be replied that Walter's sympathy is based on his imaginative rehearsal of how Wendy would feel if she were to discover her husband's infidelity. Perhaps so—but the fact remains that Walter and Wendy do not feel the same (or similar) emotions. Walter sympathizes; he does not empathize.[10]

Even though sympathy does not require emotion matching, it still is not the same as a dispassionate grasp of someone else's misfortune. We therefore propose the following definition:

> S sympathizes with O precisely when S believes that something bad has happened to O and this causes S to feel bad for O.

This definition, like the one proposed for empathy, makes use of the idea that one person feels a certain emotion "for someone else." Feeling bad for someone requires that one feel bad. That is, one must experience an "aversive" emotion, such as sadness or anger. Aversive emotions are feelings that people dislike having, but this is not to deny that people often think they should be experiencing such emotions. When bad things happen to those we care about, aversive emotions are the ones we think it is appropriate for us to feel. We'd rather not experience them in the sense that we'd rather the world not contain situations that prompt them.

Thus defined, sympathy and empathy both differ from *personal distress,* a point first made in the psychology literature by Daniel Batson (see, e.g., Batson 1991). When the perceived misfortune of another causes one to feel bad and one quite forgets the other person, the self-focused emotion that results is neither empathy nor sympathy. Personal distress involves feeling bad without feeling bad *for someone else.* There is evidence that this difference between empathy and sympathy on the one hand and personal distress on the other is associated with various physiological differences. Sympathetic and empathic concern for others is associated with *lowered* heart rate, whereas personal distress (even when triggered by the situation of another person) is accompanied by *increased* heart rate; different facial expressions and degrees of skin conductance are associated with the emotional difference as well (Eisenberg and Fabes 1991). These physiological correlates of empathy and sympathy are consistent with the more general pattern of *somatic quieting,* which often accompa-

nies an individual's focusing attention on the external environment (Lacey 1967; Obrist et al. 1970).[11]

Our definitions entail that empathy and sympathy are emotions that involve a cognitive component; each requires the formation of a belief. How, then, should we describe infants a few days old, who often cry when they hear other infants crying (Simner 1971; Hoffman 1981b)? Perhaps the causal chain in such cases is something like the following:

O is unhappy ⟶ O cries ⟶ *S* is unhappy ⟶ *S* cries

S is unhappy *because* O is unhappy. Even if *S* forms a belief in this instance, it isn't so clear that *S* forms the belief that O is unhappy or that *S* believes that something bad has happened to O. We take no stand on the empirical question this raises about child development.[12] Right now, we merely point out a consequence of our definitions. Maybe reactive crying is a precursor of empathy and sympathy, rather than the genuine article (Hoffman 1981a; Eisenberg and Miller 1987; Eisenberg and Strayer 1987; Thompson 1987).

Although empathy and sympathy both require the formation of a belief, the requisite types of belief are different. Empathy entails a belief about the emotions experienced by another person. Empathic individuals are "psychologists" (Chapter 6); they have beliefs about the mental states of others. Sympathy does not require this. You can sympathize with someone just by being moved by their objective situation; you need not consider their subjective state. Sympathetic individuals have minds, of course; but it is not part of our definition that sympathetic individuals must be psychologists.[13]

Empathy and sympathy do not automatically entail the existence of altruistic desires. Nancy Eisenberg (personal communication) has suggested a simple way to see why. It is possible to enter these emotional states by thinking about problems that have already been solved. Suppose Wendy discovers her husband's infidelity, divorces him, and then creates a good life for herself. If she then recounts this sequence of events to Walter, he may find himself empathizing with the Wendy of a few years past. What does this empathy motivate Walter to do? Apparently, Walter empathizes without forming the desire that Wendy's situation be improved.

Even if empathy and sympathy are causes of altruism, other causes may be possible. Perhaps one can want the situation of another person to improve without feeling anything. Something like this more detached form of altruism may occur when people learn about disasters in distant places. People often feel empathy or sympathy when they meet suffering face-to-face; reading about suffering in the newspaper can fail to elicit this emotional reaction. Perfectly decent people are able to go about their daily activities after they learn of the horrible misfortunes that beset strangers. It isn't that the information fails to elicit desires concerning the welfare of others; what may be true is that the bad news fails to make them feel bad. The emotions of empathy and sympathy most commonly arise when people directly perceive individuals in trouble or have a personal connection with them that allows a third-person report to register powerfully. However, perhaps we have the ability to care about suffering that we neither see nor hear, and that afflicts individuals who are not our near and dear. It is possible that other-directed desires come into existence without the mediation of an empathic pathway (Karniol 1982).

Altruism and Morality

Morality and altruism are sometimes equated, both at the level of action and at the level of motivation. The first equation says that morality always requires us to sacrifice self-interest for the sake of others. The second says that to be motivated by an altruistic desire is the same thing as being motivated by a moral principle. Both these equations are mistaken; there *is* a relationship between morality and altruism, but we must consider the issues more carefully.

What is a moral principle?[14] Moral principles, like all principles, properly so called, are *general*. They specify general criteria or relevant considerations for deciding what one ought to do. Consider, for example, a distributive principle that is central to Rawls's (1971) theory of justice: *the difference principle* says that the resources in a society may be allocated unequally only if this benefits those who are worst off. Notice that this principle does not mention specific individuals. It does not say that Earl should receive a government subsidy or that Sarah should have her taxes increased, though the principle, in conjunction with specific facts about Earl and Sarah, may have precisely this impli-

cation. In this respect, moral principles formally resemble scientific laws of nature. Newton's law of gravitation is a general principle because it covers all objects that have a certain property (mass); the principle does not mention the Earth and the Sun, although the principle, in conjunction with specific facts about the Earth and the Sun, entails that they generate a particular gravitational force.[15]

Moral principles, if they are general in the way just described, conform to an abstract universalizability criterion. They entail that if it is right for one individual to perform an action in a given circumstance, then it is right for anyone else who is relevantly similar to perform that action in the same circumstance.[16] Of course, moralities differ in what they take the relevant similarity to be. Different moral principles specify different criteria; what is good according to one may be abhorrent according to another. For example, a tribal morality may lay down obligations that one has to group members but not to outsiders. Other moralities may claim that one has certain obligations to all human beings, or to all sentient organisms. Despite such substantive differences, however, these moral systems have in common the fact that they set forth principles of the form "anyone with such-and-such characteristics is to be treated thusly." Universalizability is an invariant feature.[17]

If moral principles must be general, then it is clear that an individual can have altruistic desires without being motivated by moral principles. This is because altruistic desires are often directed at specific individuals, whereas moral principles, in virtue of their generality, are about no one in particular. Suppose two parents want their child to do well, not for egoistic reasons but because they take the well-being of their child to be an end in itself. It is possible that the parents have this altruistic desire[18] without embedding it in any moral system at all. They may never formulate the thought that all parents should care about their children; nor need they think that if some other child were theirs, they would have an obligation to take care of that child as well. Perhaps this is especially clear when the parents in question are *nonhuman animals*. Specific desires need not be accompanied by endorsements of general principles.

This difference between one's general moral principles and one's altruistic concerns is easy to discern with a little reflection. Consider two women, Alma and Beth, who know each other only slightly. Each has a child; unfortunately, each child dies. Alma will almost certainly grieve more acutely for her own child than she will for Beth's. And if Alma is honest, she will admit that she wanted her own child to survive more than she wanted Beth's to do so. But in spite of these feelings and desires, Alma may recognize that, from a moral point of view, what happened to her and her child is no worse than what happened to Beth and to Beth's child. Moral principles involve a kind of *im*personal assessment that differs from the personal perspective that frequently accompanies our emotions and desires.

Just as an individual can be an altruist without being moved by moral principles, the converse is also possible. People sometimes believe that moral principles are binding for reasons that have nothing to do with how obeying those principles will affect the well-being of others. Some may find this deontological position wrong-headed, but the fact remains that many people (including influential philosophers, such as Kant), rightly or wrongly, have been deontologists. Examples may be found in moral beliefs that are grounded in theistic convictions. Many people believe that certain actions are required simply because God commands them. You are supposed to act in certain ways, not because God will punish you if you do not, and not because your conformity will benefit other people (or God), but simply because of God's say-so. People who accept this idea act on principles, but this does not entail that they have altruistic ultimate motives.

Another gap between altruism and morality is worth noting. Altruistically motivated actions can be morally wrong. It is easiest to see this when helping someone involves harming a third party. Suppose Alan cheats Betty at cards because Alan wants to use the money to buy something for Carl. Alan may be motivated by an altruistic concern for Carl, but that may not be enough to morally justify the way he treats Betty. A macabre illustration of this point is provided by accounts of the training that Nazi concentration camp guards and physicians received. They were taught that they had to overcome their feelings of revulsion because the atrocities they were committing were for the good of the German people (Lifton 1986). If these individuals helped implement the Final Solution in part because they had the ultimate goal of helping the *Volk*, they provide a striking example of how psychological altruism can help underwrite moral evil.

Just as altruistically motivated actions can be immoral, it also is

possible for selfishly motivated actions to be the ones that morality requires. For example, consider the utilitarian maxim that goods should be distributed so as to maximize the collective happiness. Suppose a single dose of a medicine is available and that it will go to either Boris or Morris. Utilitarianism says that the drug should go to the person who will receive the greater benefit. If it is up to Boris to decide who will get the drug, and if the medicine would benefit him more, then utilitarian principles require him to take it. However, let's imagine that Boris doesn't think about utilitarian principles or about any other morality; he is merely a selfish person and so he keeps the drug for himself. According to utilitarianism, he has done the right thing (though not for the right reason). Moralities rarely require complete self-abnegation. Utilitarianism is an example; it says that self-interest counts no more *and no less* than the interests of others.[19] Moralities of this type have some implications that coincide with the dictates of egoistic desires, while other implications coincide with the dictates of altruism. Once again, it is the *im*personal character of moral principles that distinguishes them from the personal character of altruistic and self-interested desires.

We have argued that morality sometimes conflicts with self-interest and that at other times it conflicts with self-sacrifice. Although both clashes are possible, it is interesting that people often react differently to these two types of conflict. In our story about Boris and Morris, we said that utilitarianism requires Boris to take the medicine. If Boris fails to do this, and he selflessly gives the medicine to Morris, we might not feel great moral outrage. Our reactions would probably be different if morality required Boris to give the medicine to Morris, and Boris selfishly kept the drug for himself. Commonsense morality seems to set minimum standards concerning how much self-sacrifice is required, but it allows individuals to sacrifice *more* if they wish. We suspect that this is not a parochial feature of contemporary society but a fairly pervasive characteristic of people and societies generally. What might have led morality to take this form? Why doesn't morality place a lower bound on how much *selfishness* we are required to exhibit, but allow people to be more selfish if they wish? Surely the social function of morality is central to explaining this asymmetry—a point that connects with our discussion in the first part of this book of human beings as a group-selected species.

Satisficing and Irrationality

Hedonism, egoism, and the altruism hypothesis are all claims about the ultimate desires that people have. As such, none of them says anything directly about what people do. For any of these theories to generate predictions about an individual's behavior, it must be supplemented with assumptions concerning what the individual believes and the processes whereby beliefs and desires generate behavior.

One hypothesis that describes how beliefs and desires lead to action is that individuals are *rational maximizers*. This is the idea that people choose the action that their beliefs indicate will get them the most of what they want.[20] This assumption, much used in the social sciences, has come in for heavy criticism.

The idea that people are rational maximizers presupposes a kind of computational omniscience that mortal creatures do not possess. If the options under consideration are sufficiently complex, people may fail to figure out which option will provide the most of what they want. This problem led Herbert Simon (1981) to suggest *satisficing* as a more realistic principle. Individuals satisfice when they accept the first option that comes to mind that is *good enough*. A satisficer need not survey and analyze the entire field of alternatives. The savings in search time and computation can be considerable. Additional reasons to doubt that agents are rational maximizers come from the growing body of evidence that people systematically deviate from rational modes of inference. It isn't just the occasional lapse in attention that makes people draw an invalid conclusion from a set of assumptions. Rather, people often seem to reason by exploiting heuristics that work well in some contexts but lead to systematic error in others (Kahnemann, Slovic, and Tversky 1982).

Although psychological egoism is sometimes criticized for holding that people are rational maximizers, we feel that this objection is off the mark. *All* the motivational theories we are considering require a view about how beliefs and desires lead to action. Egoism is no more wedded to an unrealistic conception of this process than is the pluralistic theory in which the altruism hypothesis is embedded. In what follows, we usually will portray individuals as rational maximizers; however, we adopt this assumption strictly as a matter of convenience. When this idealization becomes problematic, adjustments can

be made. The point of importance is that the inadequacy of the idealization is a problem for *all* the theories we need to consider. What is a problem for everyone is no one's problem in particular.

How Desires Interact

Individuals often have more than one desire that is relevant in an episode of deliberation. How should we understand the idea that desires "interact" in the production of behavior?

Desires are said to "push" or "incline" agents in different "directions." When two desires conflict, it is the stronger one that determines the behavior that ensues. This commonsense description of how desires work together may not entirely capture the rich phenomenology of our inner lives. Yet, it is a highly serviceable idealization, one whose implications we want to explore. The idea is that desires are related to action in the way that component forces impinging on an object are related to its resulting motion in Newtonian mechanics. If you push a billiard ball due north and someone else pushes it due south, the direction of motion is determined by which component cause is stronger.

The idea of conflict between desires is especially relevant to understanding the concept of altruism. As noted earlier, the altruism hypothesis is best thought of as part of a pluralistic theory of motivation. We need to be able to conceptualize the conflicts that can arise between a concern for self-interest and a concern for the welfare of others. This will help us get clearer on what the altruism hypothesis does and does not entail.

Consider a hypothetical example. Suppose you are thumbing through a magazine one day and see an advertisement. The ad asks you to send a check for $25 to a charity that helps starving children. You feel that you could afford this donation. Of course, you could find other things to do with $25. But the picture in the ad is pathetic. You believe that a $25 contribution will make a real difference for the children (it won't solve the whole problem, of course). You think for a moment and then send a $25 check to the charity.

There are at least two motives that may have moved you to action. Perhaps your motive was altruistic; maybe you cared about the welfare of the children, not as a means to some benefit for yourself, but as an end in itself. On the other hand, perhaps your motive was selfish.

Maybe you wrote the check in order to obtain a glow of satisfaction—a nice feeling about yourself—and to avoid feeling guilty. Indeed, it is conceivable that your action was produced by both motives acting at once. We now will describe three possible relationships that might obtain among these two possible desires—an individual might have the altruistic desire but not the selfish desire, an individual might have the selfish desire but not the altruistic desire, and an individual might have both. This last category is the one of greatest interest, since it involves the kind of pluralism that allows the interaction of desires to be examined.

We will represent desires as *preferences*. If you want the children to be better off, this means that you prefer their being better off over their being worse off. If you want to feel the glow of satisfaction, this means that you prefer feeling the glow over not feeling it. Note that the first of these preferences is other-directed, while the second is self-directed.

The first preference structure we want to describe characterizes people who care nothing about the welfare of others; the only thing that matters to them is their own situation. This purely egoistic preference structure is depicted in the following 2-by-2 table. The table answers two questions about such individuals. What preference do they have as to whether they will receive some putative benefit (feeling good)? What preference do they have as to whether the children do better rather than worse? The numbers in the table represent the *ordering* of the agent's preferences; a situation with a higher number is preferred over a situation with a lower number. The absolute values have no meaning; we could have used the numbers "8" and "6" instead of "4" and "1." Think of the four cells in this table as four states that the world might occupy; the egoist's ranking of these four possibilities is as follows:

The Egoist

		Other: +	Other: −
Self	+	4	4
	−	1	1

Egoists care only that they receive more (+) rather than less (−) of the benefit that is at issue. It is a matter of indifference to such people whether the children do better (+) or worse (−). These individuals are not benevolent; they also are not malevolent. Their attitude to others is one of indifference.

If you were an egoist, would you give the $25 to charity? Donating the money would have two effects. You would receive a glow of satisfaction and the children would be better off. In similar fashion, not donating the money also would have two effects. You would feel bad and the children would be worse off. (For simplicity, we are ignoring whether you prefer to retain the $25.) In this situation, there are two possible actions, giving or not giving, whose consequences are represented in the table by the upper-left entry (+ to self and + to other) and the entry in the lower-right (− to self and − to other). Given the options available, egoists will choose the first action; they will donate the $25 to charity. They therefore choose an action that benefits others. However, this benefit to others is not the goal of their action; it is a mere side-effect. If you were an egoist, you would help the starving, but your ultimate motive would be to make yourself feel good.

The second preference structure is the mirror image of the egoist's. Pure Altruists care nothing about their own situation; their only desire is that other people be better off:

The Pure Altruist

Self \ Other	+	−
+	4	1
−	4	1

What would Pure Altruists do if they had to choose between the actions represented by the entries in the upper-left and lower-right cells of the table? One action will benefit both self and other; the other will benefit neither. Pure Altruists choose the former action. A consequence of this choice is that Pure Altruists will feel good about

themselves, but this benefit to self is a side-effect of the act, not the act's real motive.[21]

Egoists have only one ultimate preference; the same is true of Pure Altruists. Now let's consider preference structures in which an irreducible concern for self coexists with an irreducible concern for others. There are two cases to consider; the first we call "E-over-A Pluralism":

The E-over-A Pluralist

Self \ Other	+	−
+	4	3
−	2	1

E-over-A pluralists prefer that they be better off rather than worse off (since 4 > 2 and 3 > 1 in the preference ranking). They also prefer that other people do better rather than worse (since 4 > 3 and 2 > 1). These individuals are pluralists because they have both self-directed and other-directed preferences.[22]

We call this preference structure "E-over-A" to describe what these individuals will do if their own welfare *conflicts* with the welfare of others. Suppose the agent faces a choice between two actions. The first action provides a benefit to self but prevents other people from benefiting; this is the outcome represented in the upper-right cell. The second action confers a benefit on others but deprives self of a benefit; this is the lower-left outcome. When self-interest and the welfare of others conflict, E-over-A Pluralists give priority to themselves (since 3 > 2). Their egoistic preference is stronger than their altruistic preference. Notice how E-over-A Pluralists and Egoists differ. Egoists do not care *at all* about the situation of others. E-over-A Pluralists *do* prefer that others do better rather than worse. However, both say "me first" when self-interest conflicts with the welfare of others.

The last preference structure we want to mention also is pluralistic, but the weight it gives to self and other is the reverse of the pluralism just described:

The A-over-E Pluralist

SELF \ OTHER	+	−
+	4	2
−	3	1

A-over-E Pluralists care about others and about themselves as well. When self-interest and the interests of others conflict, however, they sacrifice their well-being to advance the interests of others. When A-over-E Pluralists have to choose between upper-right and lower-left, they choose lower-left (since $3 > 2$).[23]

The four preference structures just described—Egoism, Pure Altruism, E-over-A Pluralism, and A-over-E Pluralism—all produce the same behavior when the choice is between upper-left and lower-right; all agents choose an action that benefits both self and other over an action that benefits neither. This choice situation is one in which self-interest *coincides* with the welfare of others. However, when self-interest and the welfare of others *conflict*, Egoists and E-over-A Pluralists perform one action, while Pure Altruists and A-over-E Pluralists perform the other.

Of these four preference structures, only one is consistent with the egoism hypothesis, whereas three are consistent with the altruism hypothesis. This uneven split is due to the fact that egoism is a (relatively) monistic theory, whereas the altruism hypothesis is compatible with pluralism. The altruism hypothesis says that people sometimes have an irreducible regard for the welfare of others; whether they have other irreducible preferences is left open. It therefore is a mistake to interpret the altruism hypothesis as saying that people sometimes help for *purely* other-directed reasons; the hypothesis does not rule out the possibility that instances of helping are sometimes or even always accompanied by ultimate motives that are self-directed.

You can see from these four preference structures why it is inaccurate to describe the altruism hypothesis as saying that people are sometimes "disposed" to sacrifice self-interest to benefit others. This is not true of E-over-A Pluralism. In that preference structure, there is

an irreducible desire that others do better rather than worse; however, this preference is so *weak,* compared with the desire for self-benefit, that the individual will never produce self-sacrificial behavior. E-over-A Pluralists are *not* disposed to sacrifice self-interest; yet, they have irreducibly altruistic motives.[24] The altruism hypothesis leaves unspecified whether altruistic ultimate desires are stronger or weaker than ultimate desires that are self-interested.

The two pluralistic preference structures show why the terms "because" and "only because" must be used carefully in describing how motives are related to behavior. When self-interest and the welfare of others coincide (that is, when the choice is between upper-left and lower-right), it is true that pluralists help *because* doing so benefits self. However, it will not be true that they help *only because* helping benefits self. The latter, exclusive, claim—that self-interest is the *only* motive—is true of Egoists and of them alone.

In setting out this typology, we are not suggesting that people fall into the same category in every situation they encounter. Even if people are sometimes A-over-E Pluralists, this does not mean that they always are. A person may be willing to sacrifice self-interest for the sake of others in some situations, but not in others. What the debate over egoism and altruism requires us to ask is whether there are *any* circumstances in which concern for others is anything more than instrumental.

Applying these different preference structures to the simple example of donating money to charity helps clarify why it is so difficult to infer what someone's ultimate motives are from what the person does. When self-interest and the welfare of others *coincide,* all four preference structures predict the same behavior. This means that the observed behavior—person *X* helps person *Y*—is thoroughly uninformative about whether the egoism or the altruism hypothesis is true. When self-interest and the welfare of others *conflict* (that is, the individual has to choose between upper-right and lower-left), Egoism and E-over-A Pluralism make one prediction while Pure Altruism and A-over-E Pluralism make another. But even here, if the agent avoids self-sacrifice, this result fails to distinguish between monistic Egoism and E-over-A Pluralism.[25] Perhaps it is possible to discern what preferences people have by observing their behavior, but it remains to be seen how this can be done. Those who think

that human behavior makes the egoism hypothesis obvious should think again.

Interacting Desires as Interacting Causes

When desires interact to produce a behavior, this is a special instance of causes interacting to produce an effect. If it is hard to tell what people's motives are by observing their behavior, this difficulty may trace to generic problems that pertain to inferring causes from effects.

Consider a farmer who grows two fields of corn. In the first, the corn plants are of identical genotype (G1) and they receive one unit of fertilizer (F1). In the second field, the corn plants also are genetically identical, but they have genotype G2; in this second field, the plants receive two units of fertilizer (F2). At the end of the growing season, the farmer sees that the plants in the first field are one unit tall on average, while those in the second field average four units of height. These observations may be summarized in two cells of a 2-by-2 table:

Environment	Genes G2	Genes G1
F2	4	–
F1	–	1

Suppose the farmer wants to answer a question about the importance of *nature* and *nurture:* Do the corn plants in the two fields differ in height because they are genetically different, because they grew in different environments, or for both these reasons? With the data described so far, the farmer has no way to tell. The reason is that the genetic and the environmental factors are perfectly *correlated;* G1 individuals always inhabit F1 environments and G2 individuals always live in F2 environments.

The way for the farmer to make headway on this problem is to break the correlation. The farmer should plant a third field in which G1 plants receive two units of fertilizer and a fourth field in which G2 plants get one unit of fertilizer. The results can be entered in the other two cells of the 2-by-2 table just displayed. With observations about what happens in all four treatment cells, the farmer can make an inference concerning how genetic differences and differences in fertilizer treatment contribute to variation in plant height. Here are four outcomes that this experiment might produce:[26]

	G2	G1
F2	4	4
F1	1	1

(i)

	G2	G1
F2	4	3
F1	2	1

(ii)

	G2	G1
F2	4	2
F1	3	1

(iii)

	G2	G1
F2	4	1
F1	4	1

(iv)

In outcome (i), the genetic factor makes no difference. Whether the plants have genotype G1 or G2 does not affect their height; it is the environmental factor—the amount of fertilizer the plants receive—that explains all the observed variation. Outcome (iv) is the reverse of (i). In (iv), the fertilizer treatment makes no difference; genetic variation explains all the variation in height. Outcomes (i) and (iv) support *monistic* explanations of the variation in plant height; each suggests that only one of the factors considered made a difference in the observed outcome.

Outcomes (ii) and (iii), on the other hand, support *pluralistic* conclusions. Both suggest that genetic and environmental factors made a difference. However, they disagree about which factor mattered more. In outcome (ii), changing the fertilizer treatment yields two units of change in height, whereas changing from one genotype to the other produces only a single unit of change. In this case, the environmental factor makes more of a difference than the genetic factor. By the same reasoning, we can see that outcome (iii) suggests that genetic variation was more important than the environmental factor considered.[27]

We hope the analogy between the egoism-altruism problem and the puzzle faced by our farmer is clear. When self-interest and the welfare of others *coincide,* it will be impossible to say whether the resulting behavior was produced by egoistic motives, by altruistic motives, or by both. This impasse resembles the farmer's initial situation. When

genetic and environmental factors are perfectly *correlated,* it will be impossible to say whether the resulting variation in height was caused by genetic differences, by environmental differences, or by both. In the problem concerning egoism and altruism, the obvious experiment to perform is to put individuals in situations in which self-interest and the welfare of others *conflict.* The parallel procedure for the farmer is to plant two more fields of corn so that the upper-right and lower-left cells in the 2-by-2 table can be filled in; in this way, the initial, confounding correlation is *broken.*

Although this way of understanding the egoism-altruism debate is quite fundamental, it would be a mistake to exaggerate its resemblance with the farmer's problem. One disanalogy between the farmer's problem and the debate over motivation is that the farmer is trying to explain the variation that exists within a population of individuals, whereas arguments about egoism concern the different motives that exist within individuals themselves. A second difference between the two problems is that the egoism and altruism hypotheses are more abstract than the one that the farmer is considering. As noted earlier, egoism and altruism do not say which *specific* desires people have; rather, they specify the *types* of desires that people have as their ultimate aims. This makes the hypotheses of egoism and altruism harder to test.

But perhaps the most fundamental difference between the farmer's problem and the egoism-altruism debate is this: In the farmer's problem, the candidate causes can be identified in advance of knowing their effect on plant height. The farmer measures out the fertilizer; the local seed distributor passes along information about the genotypes of the seeds planted. No such independent access is readily available in the problem of discerning people's motives. We infer people's motives from their behavior; aside from this, we have little or no access to what their motives really are. This does not mean that the question of altruism versus egoism is insoluble; it does mean that we must tread carefully, since the inference problem is a difficult one.

40

8

Psychological Evidence

In this chapter and the next, we will review a number of scientific and philosophical arguments that have aimed to resolve the controversy concerning psychological egoism and altruism. These arguments are a motley assemblage. Some come from empirical findings in experimental social psychology. Others involve science fiction thought experiments. Still others appeal to methodological principles that describe how one should evaluate rival hypotheses when observation is indecisive. Although there is much to be learned from these arguments, we will conclude that none of them settles whether human beings ever have altruistic ultimate motives. This verdict of "not proven" is where this chapter and the next one end, but it will not be the conclusion of the book as a whole. In Chapter 10, we will bring evolutionary considerations to bear on the question of motivation.

The present chapter explores three approaches to the egoism-altruism debate that empirical psychology requires us to consider.[1] The first concerns introspection. Can we tell what our ultimate motives are simply by gazing within our own minds? The second involves the law of effect. Does this psychological principle show that hedonism must be the motivational structure of an organism that is capable of learning from experience? The third approach will be to review the experimental literature in social psychology. What do these experiments teach us about psychological egoism and motivational pluralism?

Is Introspection the Answer?

Before we begin examining the details of psychological experiments and philosophical arguments that seek to determine what our ultimate motives are, we want to address a reaction that some readers may have to this problem. Can't people tell by introspection what they want as ends in themselves? If so, the debate concerning egoism and altruism is easy to resolve; we can simply gaze within our own minds and see whether we are egoists or pluralists.

We need to be clear about what problem introspection is being asked to solve here. The problem is not just to determine what people want, but to tell what their *ultimate* as opposed to *instrumental* desires are. Social psychologists have asked people who gave money to charity and who did volunteer work for philanthropic causes why they did so. Helpers often reply that they "wanted to do something useful" or to "do good deeds for others" (Reddy 1980). Even if these introspective reports were true, they would not tell us whether the reported desires are ultimate or instrumental. Psychological egoism can grant that people want to help others; it claims that these desires are merely instrumental. Asking "Why did you help?" is the wrong question, if the point is to assess this theory. However, even the direct question "What are the ultimate motives behind your helping?" may fail to produce the information we seek, if people lack introspective access to their ultimate motives.

Introspection has had a bad reputation in psychology for a long time. When psychology broke away from philosophy at the end of the nineteenth century and became an autonomous discipline, one of the ways in which it developed its credentials as an objective science was by rejecting introspective methods. Like other objective sciences, psychology was expected to attend exclusively to data that are publicly accessible. Behaviorism took this flight from introspection to its logical extreme. Not only were introspective reports thought to provide no evidence about the inner workings of the mind; in addition, behaviorism rejected the very goal of elucidating mental states. Instead, behaviorism sought to explain behavior solely on the basis of environmental stimuli. Many nonbehavioristic approaches to psychology shared behaviorism's rejection of introspective methods, even if they retained the goal of understanding inner mental states and processes.

For example, Freud and his school maintained that the unconscious conceals and systematically distorts mental contents; one of Freud's deepest influences on psychology was to cast doubt on the reliability of introspection.

Quite apart from the tradition of shunning introspection that has developed in psychology, it is important to realize that the reliability of introspection is, in the end, a contingent matter. A well-grounded opinion about this issue should be based on evidence; a kneejerk faith in the unreliability of introspection is no more defensible than a complacent confidence in its infallibility. Furthermore, we have to realize that introspection may be more reliable in some domains than in others. For example, even if introspection is unable to reveal why people commit verbal slips, it is a separate question whether introspection can answer the question posed by egoism and altruism.

If the reliability of introspective reports about our ultimate motives must be decided by evidence, how should we proceed? The most direct way to evaluate the reliability of a report requires that one have independent access to the state of affairs that the report is supposed to describe. For example, to determine directly whether a thermometer accurately reports temperature, you must know what an object's true temperature is. Similarly, to assess directly the reliability of introspective reports about motives, one must know independently what people's true motives are. Quite obviously, the reliability of introspective reports about ultimate motives cannot be decided directly, if we don't already know how to resolve the debate between egoism and altruism.

Is a more indirect strategy available? The reliability of a thermometer can be checked even when you don't know what the temperature of any object is. Suppose you know that a certain manipulation leaves an object's temperature unchanged. You don't know what the temperatures are of the objects on your desk, but you are prepared to say that whether an object is on the right or the left side of your desk does not influence its temperature. If so, you can randomly assign objects to the left and right sides of your desk, use the thermometer on each, and see if there is a significant difference between the two sets of measurements you obtain. If the thermometer is reliable, and if your assumption that location does not affect temperature is correct, then there should be no difference.

This type of experiment could be performed to help decide whether people have reliable introspective access to their own ultimate motives. For example, suppose we have the subjects in our experiment fill out a questionnaire, which we tell them measures how empathic they are. We then throw away these questionnaires without looking at them and randomly divide the subjects into two groups. We clearly explain to each what the hypotheses of psychological egoism and motivational pluralism assert. We then tell the subjects in group 1 that they scored high on the empathy measure; we ask them to determine introspectively whether they are egoists or pluralists. We tell the people in group 2 that they scored low on the test and ask them to figure out by introspection whether they are egoists or pluralists. On the assumption that one's ultimate motives are not affected by being told whether one scored high or low on an empathy test, there should be no difference between these groups in terms of their introspective reports, if introspection is highly reliable. On the other hand, if the groups differ in their reports, this suggests that people are *suggestible*. When people think they are introspecting, they in fact are applying to themselves a theory obtained from the outside.

This experimental design needs to be fine-tuned in several ways. For example, we need to control for the possibility that subjects do not tell the truth to others about what they introspect; perhaps they know their own minds by introspection but tailor their verbal reports to fit what they think the experimenter wants to hear. One way to address this problem might be to have subjects not put their names on the reports about introspection that they write, thus assuring their anonymity. Another strategy might be to further subdivide the two groups, telling half the people in group 1 and half the people in group 2 that the experimenter is trying to prove that psychological egoism is true and telling the other subjects that the experimenter is trying to prove that psychological pluralism is correct. Although some wrinkles need to be ironed out here, it seems reasonably clear that the reliability of introspective reports about ultimate motives is amenable to empirical study.[2]

To our knowledge, the type of experiment we have just described has not been performed. Even so, it is important to realize that skepticism about introspection is not an undefended prejudice on the part of psychologists; there is considerable evidence for thinking that people often have erroneous conceptions of what is going on in their own minds. We'll describe one such finding, drawn from the useful article by Nisbett and Wilson (1977), "Telling More Than We Can Know—Verbal Reports on Mental Processes." One of the most striking results of research on situational factors that influence helping behavior is the so-called bystander effect (reviewed by Latané, Nida, and Wilson 1981). A bystander's probability of helping another person who is perceived as being in need declines as the number of other bystanders increases. Psychologists often suggest that this lowering of the probability of helping is due to a diffusion of perceived responsibility; when there are more bystanders, an agent is more likely to think that someone else should do the helping. Whether or not this is the right explanation, the bystander effect has been confirmed when the needy other is a stranger, but also when the relationship is quite intimate. For example, people who need kidney transplants are more likely to find a sibling who agrees to donate if they have *fewer* siblings (Simmons, Klein, and Simmons 1977, p. 220).

In their work on the bystander effect, Latané and Darley (1970) asked experimental subjects whether their inclination to help was influenced by how many bystanders were present. Subjects consistently denied that this was so, and also denied that other people are influenced by this consideration. If behavior is determined by the agent's beliefs and desires, then the fact that people behave differently in two situations must mean that they have different beliefs or different desires in those situations. If subjects are not aware that their behavior when bystanders are present would differ from their behavior when bystanders are absent, then they presumably are not aware that they would have different beliefs or different desires in those two circumstances. It doesn't follow from this that people are unaware of what their *desires* are, or that they are not aware of what their *ultimate desires* are. Still, this last possibility cannot be dismissed. If the mind is not an open book, then why think that one chapter in that book—the one in which one's ultimate desires are inscribed—can be read infallibly by introspection?

In the culture we inhabit, some people are sincerely convinced of psychological egoism, while others are convinced that they and others have altruistic ultimate motives. Defenders of egoism often think that

pluralists are trapped by a comforting illusion. Pluralists sometimes entertain a reciprocal hypothesis—that egoists embrace a darker view of human motivation because they enjoy thinking that they have the fortitude to do without comforting illusions. Of course, neither of these suggestions tells us what our ultimate motives in fact are. However, both indicate that sincere introspection is not enough.

Proponents of both positions believe that their pet theories apply to others and to themselves as well. It is conceivable that defenders of egoism are right about themselves *and* that defenders of pluralism are right about themselves. It is even conceivable that people who change their minds about whether they are egoists or pluralists have true views about themselves both before *and* after. Yet the possibility remains that one side or the other is mistaken in the claims they make about their own motives. Sincere avowals, by both parties, must be set to one side. Introspective claims should be regarded as just that—as claims, whose accuracy must be judged on other grounds.

The Law of Effect

Hedonism is sometimes defended by saying that the theory describes what an organism must be like if it is to be capable of learning from experience. The idea is that learning requires organisms to experience positive and negative sensations; experiencing the former and avoiding the latter must constitute its ultimate goals in behavior. According to this proposal, learning takes the form of a conditioning process. If a behavior is followed by a positive sensation, this raises the probability that the organism will repeat the behavior. If the behavior is followed by a negative sensation, this lowers that probability. This is the law of effect, first proposed by E. L. Thorndike (see Dennett 1975 for discussion). Without this feedback loop through the experiential consequences of behavior, there is no way for the organism to change the way it acts.

It is a curious historical fact that the law of effect, which adverts to the positive and negative experiences that accompany behavior, was embraced as a central principle by behaviorists, who also demanded that psychology stop trying to talk about inner mental states. Be that as it may, the law of effect is of interest to our present inquiry because of its connection with hedonism; the claim we need to assess is that

hedonism must be the true theory about human motivation if organisms that learn must obey the law of effect.

The law of effect does not say that every behavior occurs because the organism was conditioned earlier; that would mean that no behavior ever occurs for the first time. This is not only absurd on its own terms; it also conflicts with the very idea of conditioning. A conditioning process requires that the organism perform the target behavior at least once *before* it receives the conditioning rewards and punishments. Rather, what the law of effect says is suggested by a simple example of operant conditioning. Consider a pigeon in the controlled environment that has come to be called "a Skinner box." At first, the pigeon's pecking is unrelated to whether a light in the box is on. If the pigeon is rewarded for pecking when the light is on, however, the pigeon's behavioral pattern will change. As the pigeon is repeatedly rewarded, the probability increases of its pecking when the light is on. Eventually, it pecks precisely when the light is on. Before the conditioning process, the pigeon's probability of pecking is *independent* of whether the light is on; after the conditioning process is over, the probability of pecking if the light is on is close to 1, and the probability of pecking if the light is off is close to 0. Understood in this way, the law of effect does not rule out the occurrence of behaviors that were never conditioned; rather, what it rules out is the existence of probabilistic dependencies between behavior and environment that were not caused by a conditioning process. It also rules out the possibility that a conditioning process could fail to induce such dependencies.

Both these implications are problematic in the context of behaviors that are strongly influenced by "innate" or "instinctual" factors.[3] Consider Konrad Lorenz's famous example of imprinting in greylag geese. A gosling will follow the first adult goose or human it sees that gives calls in response to the gosling's calls. However, goslings will not treat a rock as "Mom," nor will they imprint on a model chicken that emits prerecorded calls, if the calls are not produced in response to the goslings' calls (Lorenz 1965). That is, the probability of the imprinting behavior differs according to the environmental cue, but this is not because a conditioning process occurred.

Just as a behavior can depend on an environmental stimulus without its having been rewarded earlier, so it can fail to occur even

though it was rewarded before. Garcia and Koelling (1966) exposed rats to a complex stimulus consisting of flashing lights, noise, and saccharin-flavored water. Afterwards, the rats were exposed to X-rays, which made them sick. The rats thereby acquired an aversion to foods flavored with saccharin, but not to food that was accompanied by noise or by flashing lights. In another experiment, the same compound stimulus was used, but this time the rats received an electric shock to their feet. As a result, they developed an aversion to the noise and the flashing lights, but not to saccharin. The same pattern has been recorded for a variety of species, our own included (Breland and Breland 1961; Hineline and Rachlin 1969; Sevenster 1973; Gallistel 1980). For example, it is easier to condition human infants to be afraid of snakes, caterpillars, and dogs than of opera glasses or cloth curtains (Rachman 1990, pp. 157–158), and it is easier to condition a physiological response to angry faces than to happy ones (Ohmman and Dimberg 1978).

These results cast doubt on the law of effect's commitment to what learning theorists call "the equipotentiality thesis"—the idea that conditioning can successfully pair any stimulus with any behavior. This is not to deny that *some* behaviors can be explained by the law of effect. For example, Moss and Page (1972) ran an experiment in which people on a busy street were asked for directions to a well-known local store. Most complied. Of those who provided directions, some were thanked with a smile while others were abruptly interrupted and told that their directions were incomprehensible. A short time later, the people who provided directions encountered a person who had just dropped a small bag. More than 93 percent of those who had been graciously thanked when they provided directions helped the person who had dropped the bag, whereas only 40 percent of those who had been rebuffed and scolded offered help. Moss and Page also found that helping in a control group—people who had not been positively or negatively reinforced—occurred at a rate of 85 percent. These are patterns one would expect, given the law of effect.

It is important to remember that the law of effect is a *general* principle; the question is whether it is true of *all* behavior, not just of *some*. We suggest that it is not—being rewarded does not *always* raise the probability of a behavior's being repeated, and probabilistic dependencies between behavior and environment do not *always* stem

from this type of conditioning process. In addition, even circumstances that conform to the law—such as the experiment that Moss and Page performed—provide no evidence for hedonism. Moss and Page's observations were consistent with the hypothesis that pleasure and pain are motivators. Their experiment does not show that people care *only* about pleasure and pain.

The law of effect describes one possible mechanism whereby an organism can modify its behavior in the light of experience. However, it is not the only one that is conceivable. Consider the fact that means-end deliberation can generate actions without conforming to the dictates of hedonism. Deliberation leads us to revise an instrumental desire by using more ultimate desires as leverage. Any more ultimate desire will do the trick. Consider Arnold, who suddenly acquires the desire to get into his car and drive to the bakery. He acquires this new desire because he wants to buy bread and he believes that the bakery is the best place for him to do this. Deliberation is able to produce new instrumental desires because the agent regards old desires as ends for which means must be sought. The same is true when an individual abandons a desire already held. If Arnold is about to drive to the bakery when his friend brings him a loaf of bread, Arnold may lose his desire to drive to the bakery. When deliberation leads us to acquire a new instrumental desire or to abandon an old one, this is something that our *other* desires accomplish for us. There is no requirement that our ultimate goals must include the desire to attain pleasure and avoid pain; still less is it required that attaining pleasure and avoiding pain must be the only ultimate desires we have.

This point about learning is an important one, for it allows us to identify an evolutionary question that otherwise might escape our notice. Pleasurable and aversive sensations constitute one mechanism that allows an organism to learn from experience. In principle, there are others. Why did evolution assign to pleasure and pain the roles they now play in learning? This question deserves a substantive answer; we render the question invisible if we reply that learning, by definition, is a conditioning process mediated solely by pleasure and pain.

We conclude that psychological egoism cannot be defended by appealing to the law of effect. That "law" is sometimes false. And even when behaviors recur *because* they were rewarded earlier, it does

not follow that they do so *only because* they were rewarded. Even if people are motivated by the prospect of attaining pleasure and avoiding pain, it does not follow that this goal is the only thing that people ultimately care about.

Experiments in Social Psychology

Egoism is a (relatively) monistic theory, whereas the altruism hypothesis, as we understand it, is part of a more pluralistic view of human motivation. This logical difference between the two theories has implications for how each of them may be tested. An experiment demonstrating that people in fact possess a particular egoistic ultimate motive *does not* disconfirm the altruism hypothesis, but a hypothesis demonstrating that people possess a particular altruistic ultimate motive *does* disconfirm the egoism hypothesis. It does no good in this debate to show that people are motivated by egoistic concerns. Ideally, an experiment should put people in a situation in which they will behave one way if they have altruistic ultimate motives and behave another way if they do not. But this is not so easy to do, since, as we have noted, the egoism hypothesis is a flexible instrument that comes in a variety of forms.

The research in experimental social psychology that does the best job of coming to grips with the problem of testing egoism and altruism is that of Daniel Batson and his associates. This work was synthesized by Batson (1991) in his important book *The Altruism Question;* Batson and Shaw (1991) provides a useful summary. Batson is admirably alert to the risk of confusing real altruism with pseudo-altruism. His experiments are designed to track down and test different varieties of egoism; Batson is well aware that refuting one or two forms of egoism is not the same as refuting egoism *per se.* In discussing Batson's work, we also will examine the provocative work of Robert Cialdini and his associates, as well as some other experimental research.

Batson wishes to test a conjecture that he terms the *empathy-altruism hypothesis.* As noted in the previous chapter, Batson was the first person in experimental social psychology to delineate the distinction between empathy and personal distress. Whereas personal distress typically leads people to want to improve their own situations, the empathy-altruism hypothesis asserts that empathy causes people to

have altruistic desires. Since this claim is supposed to be incompatible with egoism, the altruistic desires that it says are triggered by empathy must be *ultimate.* Batson's methodology is to test different versions of egoism against the empathy-altruism hypothesis. In each case, he argues that the version of egoism considered is disconfirmed by the data, but that the data support the empathy-altruism hypothesis.

The first version of egoism that Batson considers is the *aversive-arousal reduction* hypothesis. This is the idea that bystanders who see a needy other have unpleasant experiences that they wish to expunge. They help for the same reason you turn down the thermostat when the room is too hot. When you lower the thermostat, this is not because you care about the room. Helping others has precisely the same type of motivation—it is merely a means of achieving a better level of personal comfort.

The philosopher C. D. Broad (1952, pp. 218–231) argued against this version of egoism by describing a physician who travels to Asia to open a clinic for people suffering from leprosy. The physician knows that this line of work will bombard him with distressing experiences. Broad thought that this example straightforwardly refutes the egoistic hypothesis that helping is motivated just by the desire to escape from the unpleasant experiences occasioned by exposure to those in need. If one's only ultimate desire is to avoid or reduce aversive feelings, one would *avoid* any situation like the one the physician chose.

We suspect that many readers, and most philosophers, will find Broad's argument sufficient to refute the aversive-arousal reduction hypothesis. Broad's claim that people sometimes act the way the physician does in his example rings true; and the logic of his argument—that it refutes this egoistic hypothesis—sounds right as well. If this is correct, then there seems to be no need for further experiment or natural observation. Readers who feel this way may be surprised that Batson and his associates ran several experiments to test the aversive-arousal reduction hypothesis against the empathy-altruism hypothesis. Why did they do so? Psychologists often run experiments in which commonsense tells one what to expect; since common sense expectations are not always borne out, there is merit in testing propositions that people think are intuitively obvious. Let's examine how Batson and his colleagues proceeded.

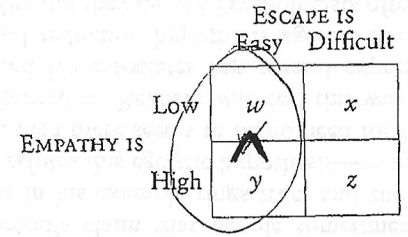
In one of these experiments, subjects were told that they would be part of a study in which they would watch over closed-circuit television while a student receives ten electric shocks. Actually, each subject watched a videotape of an actress ("Elaine") who pretends to find the first two shocks quite distressing. The videotape then shows one of the experimenters tell Elaine that he is concerned about her discomfort; he suggests to her that she stop taking the shocks if the subject agrees to substitute for her. Elaine gratefully consents to this arrangement, the television screen goes blank, and a confederate then enters the room where the subject is and asks whether the subject would be willing to take Elaine's place.

Subjects in the "easy-escape" treatment of the experiment had been told before they started to monitor Elaine that they would be required to watch just two shocks; the confederate reminds them of this and offers them the opportunity of exiting the experiment if they do not want to substitute for Elaine. Subjects in the "difficult-escape" treatment had agreed at the outset to watch the entire sequence of ten shocks; the confederate reminds them of this promise and says that they will have to watch Elaine experience eight more shocks if they do not take her place.

Not only did subjects vary in terms of whether escape was easier or more difficult. They also were manipulated in ways designed to influence how much empathy they would feel for Elaine. This was achieved by a variety of means. One technique exploited the fact that people tend to empathize more with individuals whom they take to be similar to themselves (Stotland 1969; Krebs 1975). Subjects in the high-empathy treatment received a description of Elaine that closely matched what they had reported about themselves in an inventory of personal values and interests that they filled out before they saw Elaine. So-called low-empathy subjects were given a description of Elaine that failed to match. The idea behind this manipulation was not that all people in the high-empathy treatment would have high empathy for Elaine and that all in the low-empathy treatment would have little empathy, but that the average level of empathy would be higher in the former group than in the latter.[4]

Each subject in the experiment was given either the easy-escape or the difficult-escape treatment, and each was given the high-empathy or the low-empathy treatment. This means that each subject was

placed in one of four circumstances. The experiment discovered how often people in each of these four treatment cells volunteered to help Elaine. Let's represent the frequencies of offers to help in the four circumstances as w, x, y, and z:



Before describing what the results of the experiment were, we need to consider what the two hypotheses predicted. To do this, we must consider carefully what each hypothesis says.

Suppose we interpret the empathy-altruism hypothesis as saying just that a higher level of empathy makes people more inclined to help, at least sometimes. Construed in this way, the hypothesis predicts merely that $y > w$ or that $z > x$ or both. In parallel, suppose we interpret the aversive-arousal reduction hypothesis as saying just that people are sometimes more likely to help if it is difficult for them to escape from the needy other. Under this construal, the hypothesis predicts merely that $x > w$ or $z > y$ or both. Notice that if we interpret the two hypotheses in this modest way, their predictions do not conflict. The empathy-altruism hypothesis makes "vertical" predictions about the 2-by-2 table, whereas the aversive-arousal reduction hypothesis makes "horizontal" predictions. If this is all the hypotheses say, then they do not disagree; in fact, it would be wrong to view the aversive-arousal reduction hypothesis as a version of egoism, since it does not rule out the possibility that the empathy-altruism hypothesis is true. Batson avoids this problem by interpreting both hypotheses as making both "horizontal" and "vertical" predictions about the frequencies of helping. Construed in *this* way, the hypotheses come into conflict with each other.

Batson interprets the two hypotheses as disagreeing over whether empathy level will make a difference among subjects who are in the easy-escape treatment (i.e., over how w and y are related).[5] He reads the empathy-altruism hypothesis as predicting that empathy will aug-

ment helping when escape is easy; he interprets the aversive-arousal hypothesis as predicting that empathy will make no difference in this circumstance. The result of the experiment was that high-empathy subjects offer to help more than low-empathy subjects when escape is easy.[6] This disconfirms the aversive-arousal reduction hypothesis as an *exclusive* explanation of helping. Even if this motive sometimes plays a role, it cannot be the only one at work.

It does not follow, as Batson realizes, that the empathy-altruism hypothesis is correct; there may be *other* egoistic motives besides aversive-arousal reduction that can explain the experimental results. For example, perhaps high-empathy subjects in the easy-escape treatment realized that they would retain painful memories of the needy other if they declined to help, whereas low-empathy subjects in the easy escape treatment were less plagued by this worry. If so, we have an egoistic explanation for why high-empathy subjects more often offered to help. This point was made separately by three commentators on the Batson and Shaw (1991) article (Hoffman 1991; Hornstein 1991; Wallach and Wallach 1991).

Batson and Shaw (1991, pp. 167–168) formulate a reply to this suggestion. They argue that the egoistic hypothesis just described—that people offer to help because they know that refusing to do so will leave them with painful memories of the needy other—makes a prediction about what will happen in another experiment. The prediction is that high-empathy individuals will choose to receive news in the future about the situation of needy others only to the extent that they expect the news to be good. A third experiment, which we'll describe shortly, provides evidence that this isn't so. We agree that this argument supports Batson and Shaw's claim that individuals who choose to help rather than to escape aren't motivated solely by the desire to avoid having the belief that the needy other is doing poorly. However, this point about information seeking does not undercut a different egoistic explanation that appeals to guilt feelings. Just as exiting without helping can produce guilt feelings, so can refusing information about the situations of needy others on the grounds that the news might be bad. More on this soon.

Batson's second experiment focuses on a different egoistic hypothesis, which he calls the hypothesis of *empathy-specific punishment.* This conjecture comes in two forms; it says that empathically aroused

individuals help because they want to avoid the censure of others or because they want to avoid self-censure. We'll focus on the second formulation. Batson reasons that if this hypothesis were correct, people would be less inclined to help when they are provided with a strong justification for not helping. In the experiment, subjects were asked whether they would help a needy other. Some were told that many others had declined to help when placed in the same situation; these subjects thus received a *high* justification for not helping. Other subjects were told that few people had refused to help, thus receiving a *low* justification for not helping. Subjects also differed in whether they had low or high-empathy for the individual who needed help. The 2-by-2 experiment thus placed subjects in four treatment cells, and the frequency of volunteering to help among subjects in each cell was tabulated:

Empathy is	Justification provided for not helping is: High	Low
Low	w	x
High	y	z

Batson (1991, p. 136) interprets the empathy-specific punishment hypothesis and the empathy-altruism hypothesis as agreeing about a lot: both predict that empathy level makes a difference, regardless of how much justification the subject receives for not helping; both predict that $y > w$ and $z > x$. He also says that the two hypotheses predict that low-empathy subjects will help less when they have a strong justification for not helping ($w < x$). Where, then, do the hypotheses disagree? The nub of the matter, as Batson sees it, concerns how the level of justification for not helping affects high-empathy subjects (Batson and Shaw 1991, p. 116). The empathy-specific punishment hypothesis predicts that high-empathy subjects will help more when they have little justification for not helping than they will when they have a lot of justification for not helping ($z > y$). The empathy-altruism hypothesis, in contrast, is said to predict that high-empathy subjects will not be influenced by this; they should provide

the same amount of help regardless of whether they receive high or low justification for not helping ($z = y$).[7]

Thus construed, the empathy-altruism hypothesis scores a victory; it turns out that the frequency of offers to help made by high-empathy subjects is not affected by whether they receive high or low justification for not helping. Yet, a central interpretive question remains: Does empathy promote helping by causing subjects to have an altruistic ultimate motive? Even if the wish to avoid disapproval cannot explain the outcome of this experiment, this question remains open.

The third egoistic hypothesis that Batson examines is the hypothesis of *empathy-specific reward,* which comes in two forms. The first says that we help only in order to receive a mood-enhancing reward, either from self or from others; a special case of this hypothesis asserts that we help only to secure the mood-enhancing news that the needy other is better off. The second version of the empathy-specific reward hypothesis says that empathy causes sadness, which we want relieved, so we seek some mood-enhancing experience that will do the trick.

Batson, Dyck, Brandt, Batson, Powell, McMaster, and Griffitt (1988) tested the first of these hypotheses by seeing how the mood (as determined from self-reports) of high and low-empathy individuals was affected by depriving them of the opportunity to help. As before, we can understand the study in terms of the predictions that the hypotheses make about what will happen in different treatment cells. However, now there are six treatments and the effects recorded in the cells are not frequencies of offers to help but of (self-reported) levels of mood:

Empathy is	Help administered by: No one	Subject	Third party
Low	*a*	*b*	*c*
High	*d*	*e*	*f*

The empathy-altruism hypothesis predicts that the mood of a high-empathy subject will depend on whether the needy other receives help, not on whether that help comes from the subject or from some third party; the prediction is that $e = f > d$.[8] The empathy-specific reward

hypothesis, on the other hand, predicts that high-empathy subjects have their moods boosted only if they themselves provide the help—i.e., that $e > f = d$. The experiment came out as the empathy-altruism hypothesis predicted (Batson 1991, p. 150; Batson and Shaw 1991, p. 117). High-empathy subjects have their moods improve when they learn that the needy other has received help; it doesn't matter to them who the helper was.

One way to try to explain this result within the framework of egoism is provided by the *empathic joy hypothesis* (proposed by Smith, Keating, and Stotland 1989). This hypothesis says that people help, not to receive the rewards that come from helping, but to gain the good feelings that derive from sharing vicariously in the needy person's relief. Good news about the needy person's improved situation provides a boost in mood; it doesn't matter why the needy person now is better off. Batson, Batson, Slingsby, Harrell, Peekna, and Todd (1991) tested this hypothesis by seeing if the proportion of subjects choosing to have a second interview with a needy other would be influenced by whether they were told that the person has a 20 percent, a 50 percent, or an 80 percent chance of improvement before that second conversation. The authors reasoned that the empathic joy hypothesis and the empathy-altruism hypothesis make different predictions about how often individuals in six different treatment cells will elect to have a second interview:

Empathy is	Probability that needy other will show improvement: 20%	50%	80%
Low	*a*	*b*	*c*
High	*d*	*e*	*f*

Both hypotheses predict that high-empathy subjects should request the second interview more often than low-empathy subjects, and the experiment bore this out.

The experimenters reasoned that the hypotheses make different predictions about how high-empathy subjects differ among themselves. They interpret the empathic joy hypothesis as predicting that

high-empathy subjects are more likely to request a second interview when the probability is higher that an interview will provide good news about the needy other; the prediction is that $d < e < f$ (Batson 1991, pp. 161–162). In contrast, the empathy-altruism hypothesis is said to predict that high-empathy individuals should either not be influenced by how probable it is that they will receive good news, or that they should be most interested in receiving news when they are maximally uncertain. That is, the empathy-altruism hypothesis predicts that either $d = e = f$ or $d < e > f$. The result of the experiment matched this prediction of the empathy-altruism hypothesis—the frequency of requests for a second interview among high-empathy subjects is not an increasing function of the probability that the news will turn out to be good.

Even so, it is not difficult to invent an egoistic explanation of this outcome. Uncertainty can be a torment; this is a familiar experience when the question mark concerns our own welfare, and also when the uncertainty involves the well-being of those we care about. Of course, we'd rather receive good news than bad, but people also prefer receiving information over remaining in the dark. We may apply this idea to Batson's experiment by hypothesizing that high-empathy subjects choose to receive news because they want to reduce the disagreeable feelings that accompany uncertainty. In addition, declining the offer of information might make high-empathy subjects feel guilty. Apparently, the results of this experiment can be accommodated within the framework of egoism.

The second version of the empathy-specific reward hypothesis that Batson studied is due to the work of Cialdini and colleagues (Cialdini, Schaller, Houlihan, Arps, Fultz, and Beaman 1987; Schaller and Cialdini 1988). This is the *negative-state relief hypothesis,* which says that empathic individuals become sad when they witness a needy other; they then help in order to lift themselves out of their sadness. One of Cialdini's experiments is striking and unexpected in its results. The experiment begins with all subjects taking a "drug" (actually, a placebo). Some subjects are given perspective-taking instructions designed to produce high-empathy with a fellow student; others are placed in a low-empathy treatment. They then are told that the student needs help in going over her class notes. Before giving subjects the opportunity to volunteer to help, some of the subjects are told that

the drug they had taken earlier will have the effect of freezing their mood for a half-hour or so. These students are thus in the "fixed-mood treatment." The other students are given no such story and thus are said to occupy the "labile-mood treatment." The experiment tabulated how often subjects volunteered to help in four circumstances:

		Mood is: Labile	Mood is: Fixed
Empathy is	Low	w	x
	High	y	z

Cialdini et al. (1987) reasoned that the negative-state relief model predicts that high-empathy subjects should help more than low-empathy subjects when their mood is labile, but that empathy should make no difference in helping when subjects believe that their mood is fixed; that is, the negative-state relief model predicts that $y > w$ and $z = x$. In contrast, Cialdini and his coauthors suggest that the empathy-altruism hypothesis predicts that high-empathy should produce more helping, regardless of whether subjects think their mood is fixed or labile ($y > w$ and $z > x$). The point of difference between the two hypotheses, therefore, concerns whether subjects in the fixed-mood treatment volunteer to help more when they feel a high degree of empathy.

Cialdini et al. (1987) tabulated both the *proportion* of individuals who volunteered to help in each treatment cell and the *amount of time* they volunteered to spend. The data on amount of time supported the negative-state relief model, whereas data on proportions volunteering to help did not. Although Cialdini et al. viewed this and another experiment as favoring the egoistic hypothesis under test, Batson (1991, p. 166; Batson and Shaw 1991, p. 118) regards the results as more equivocal. Despite this difference in interpretation, both parties note that the results might have been due to *distraction*—after empathy was induced, the subjects were told for the first time that the drug they had taken earlier would freeze their moods. Perhaps this jarring information diminished the amount of empathy they experienced.

This possibility was confirmed by an experiment performed by Schroeder, Dovidio, Sibicky, Matthews, and Allen (1988), which closely resembled the Cialdini experiment, except that subjects were told about the mood-fixing effect of the drug when it was given to them and then were simply *reminded* of this fact just before they were asked whether they would help. Schroeder et al. found that high-empathy individuals volunteered to provide more time helping than did low-empathy individuals, both when their moods were labile *and* when they were fixed (though these observed differences were not statistically significant). The pattern of data on how often individuals volunteered to help was not terribly strong and was not well explained by either hypothesis. Schroeder et al. and Batson (1991, p. 168) drew the conclusion that these experiments do not favor Cialdini's negative-state relief model.

Batson, Batson, Griffitt, Barrientos, Brandt, Sprengelmeyer, and Bayly (1989) tested Cialdini's hypothesis in another way. As usual, subjects were assigned to a high-empathy or a low-empathy treatment. In addition, some individuals were told that they would receive a mood-enhancing experience (such as listening to music) regardless of whether they chose to help a needy other; other individuals, who also were provided with the opportunity to help, were not offered the chance to receive the mood-enhancing experience. The experiment produced data about the proportion of individuals choosing to help in these four treatments:

		Mood-enhancing experience was	
		Promised	Not promised
Empathy is	Low	w	x
	High	y	z

The negative-state relief hypothesis predicts that empathy will not affect helping when a mood-enhancing experience has been promised ($y = w$), while the empathy-altruism hypothesis predicts that empathy will make a difference ($y > w$). The data favored the empathy-altruism hypothesis (Batson 1991, p. 172; Batson and Shaw 1991, p. 119).

Here again, an egoistic explanation is not far to seek. If empathizing with a needy other makes a subject sad, why expect the subject to think that listening to music will be a completely satisfactory mood corrective? When we are sad, we usually are sad about something in particular. It is not surprising that the pain we experience in empathizing with the suffering of others is not completely assuaged by any old pleasant experience; however, this presents no difficulty for the egoism hypothesis.

The strategy behind Batson's research program is to show that each of the versions of egoism he has formulated encounters observations that it is unable to explain. How do these findings bear on the question of whether there is a set of observations that no version of egoism will be able to explain? Of course, there is no deductive entailment here; from the fact that everyone has a birthday, it does not follow that there is a single day on which everyone was born. The failures of simple forms of egoism don't *prove* that more complex formulations also must fail. Even so, it might be suggested that Batson's experiments *raise the probability* that no version of egoism can be observationally adequate. This may be so. Nonetheless, when we survey the ingenious experiments that social psychologists have constructed, we feel compelled to conclude that this experimental work has not resolved the question of what our ultimate motives are. The psychological literature has performed the valuable service of organizing the problem and demonstrating that certain simple egoistic explanations are inadequate. However, there is more to egoism than the hypotheses tested so far. What we find here is a standoff.

One reason the methodology that Batson uses may not be able to resolve the egoism-altruism debate becomes apparent when we compare his experimental design with the experiment we discussed at the end of the previous chapter. There we described a farmer who wants to know whether the difference in height observed between two fields of corn is due to a genetic difference, an environmental difference, or both. The farmer plants two additional fields of corn and thereby obtains data that allow this question to be answered. If the farmer has so little difficulty in disentangling nature and nurture in this experiment, why are Batson's experiments more equivocal?

The farmer experimentally manipulates genes and environment, just as Batson manipulates empathy level and some other factor (e.g., ease

of escape). What can be discovered from this method, in the first instance, is that empathy makes a difference in the effect variable measured. However, this finding leaves unanswered the crucial question of whether empathy has its effect by triggering an altruistic ultimate motive. The same problem could arise for our farmer. Suppose the farmer isn't interested just in whether genes make a difference in plant height, but wants to test a more detailed hypothesis. This hypothesis has two parts; it says that there are genes that cause corn plants to grow tall *and* that these genes bring about this effect by producing enzyme *X*. If the farmer ran the experiment we described before and discovered that genes make a difference in plant height, one-half of the gene-enzyme hypothesis would be confirmed. But what about the other half? The farmer can tell from the observations that genes make a difference, but he can't tell whether they do so by activating enzyme *X*.

Of course, even if one type of experiment is incapable of disentangling psychological egoism and motivational pluralism, another design might be able to do the job. Nonetheless, it is tempting to claim that *any* behavior elicited in a psychological experiment will be explicable both by the egoism hypothesis and by the pluralism in which the altruism hypothesis is embedded (Wallach and Wallach 1991). We take no stand on this stronger thesis; what experimental psychology has been unable to do so far, new methods may yet be able to achieve. For now, however, the conclusion we draw is a discouraging one. Observation and experiment to date have not decided the question, nor is it easy to see how new experiments of the type already deployed will be able to break through the impasse.

It is essential to understand that just as the psychological literature has not established that the egoism hypothesis is false, it also has not established that pluralism is false. Why, then, has egoism struck so many as a theory one should believe unless one is forced to change one's mind? Perhaps there are compelling arguments, not based on observed behavior, that tip the scale in egoism's favor. Alternatively, the possibility needs to be faced that there is no good reason to regard egoism as a theory that is innocent until proven guilty. Historically, the egoism and altruism hypotheses have competed on an uneven playing field. In the next chapter, we will inquire further into the question of whether the privileged status assigned to egoistic explanations can be justified.

The Significance of the Psychological Question

The reason it is difficult to obtain experimental evidence that discriminates between egoism and motivational pluralism is that we have allowed egoism to appeal to *internal* rewards. If the view were more restrictive—if egoism claimed that we care only about *external* rewards such as money—the problem would be easier. We have already explained our reasons for choosing a definitional framework in which warm feelings are just as much a self-benefit as cold cash. The question we now want to address is, *why does it matter, given this definitional framework, what our ultimate motives really are?* The thought behind this question is that what matters, in human conduct, is how people treat each other; the motives behind helping really aren't important. If an altruist would be nice to you and an egoist would be nasty, you would want to know which type of person you were about to meet. But if altruists and egoists will treat you the same way, why should you be interested?

We have two replies. First, we think the question of whether people ever care about others as ends in themselves is theoretically significant. Of course, some people may find this question uninteresting, just as some may not be enchanted by theoretical issues in astronomy. However, if psychology is in the business of trying to explain behavior by elucidating mental mechanisms, it is hard to see how the structure of motivation can fail to be an important psychological problem.

Our second reply to people who think that what matters is whether people help, not why they do so, is practical. Consider the conjecture that what people believe about their ultimate motives influences how much they help. The hypothesis is that when people believe that egoism is true, they are inclined to be less helpful (Batson, Fultz, Schoenrade, and Paduano 1987). There is evidence in favor of this hypothesis. Frank, Gilovich, and Regan (1993) discuss several studies that compared economists with people in other disciplines in terms of a variety of measures of cooperative behavior. The pattern is that economists tend to be less helpful. It might be thought that this is because people who are less inclined to be helpful self-select for careers in economics. To rule this out, Frank et al. did a before-and-after study on students enrolled in two introductory economics courses and also on students in an introductory astronomy course.

Students were asked at the beginning of the semester, and also at the end, whether they would return to its owner an envelope they found that contained $100. The students also were asked whether they would inform a store about a billing mistake if they had been sent ten computers but had been billed only for nine. At the *beginning* of the semester, the economics students and the astronomy students said they'd perform the honest action about equally often. The economics and the astronomy students differed in how they *changed* during the semester. The willingness to act dishonestly increased among students in the economics classes more than it did among those in the astronomy class. This is evidence that studying economics inhibits cooperation. Of course, it is a further question whether economics has this effect by encouraging people to believe that psychological egoism is true. We think that this is a plausible guess, since this motivational theory plays a more prominent role in economics than in any other discipline.

Even if believing psychological egoism makes people less helpful, it does not follow that this theory is false. Maybe egoism is true, and the perception of its truth causes people to become a little more self-centered. The point we are making here is that if psychological egoism is false, then that may be a fact worth knowing, not just because it is theoretically important but because recognizing the falsehood of egoism may influence conduct (Batson 1991).