PHILOSOPHICAL TOPICS
Volume XIV, No. 2, Fall 1986

Rational Agent, Rational Act[1]

STEPHEN L. DARWALL
University of Michigan, Ann Arbor

It has long been well understood in moral philosophy that differences between competing theories of right conduct are often symptomatic of deeper and more systematic differences. A theory of right is rarely advanced in isolation from other ethical theories, theories, for example, of the good, moral character and the person. Thus adherence to a given view of moral conduct will frequently be animated by a larger vision of how that view fits within a whole moral philosophy.

Some consequentialists, for example, are inclined to advance their theory of right because it is best supported by an independent view about the value of states of affairs. Deontologists, on the other hand, may be attracted to a different theory of right because they hold a different view of how a theory of right should fit with other theories; they may hold the right to be prior to the good. And different views are possible even here. Some deontologists, like Ross, hold that propositions about what it is right to do, *prima facie* anyway, are utterly fundamental and underived. Others, such as Kant, can more fruitfully be understood as beginning with an account of the moral agent, of the morally good person, and proceeding to a theory of right from there.

A similar complexity is possible, and is in fact developing, in theorizing about rationality. Since the sort of theorizing I have in mind is critical or normative, this is not surprising. These theories often make claims about both the rationality of *persons* and the rationality of *actions*. So, as in moral philosophy, we should expect that conflicting views of rational conduct will sometimes be symptomatic of different systematic practical philosophies. What manifests itself as a relatively specific disagreement about what it is rational to do in certain circumstances may actually be rooted in deeper, more systematic disagreements about the relation between

rational action, rational agency, and the good.

In fact, this is the case. Different theories of rational conduct often do stem from more systematic differences. However, that this is so is insufficiently appreciated. There has been nothing like the investigation of different systematic approaches to the theory of rationality of the sort with which we are familiar in moral philosophy. My first task in this essay will be to demonstrate the necessity of conducting such an investigation in order fully to grasp the issues that divide different theories of rational action.

In Section I, I briefly describe a debate between Derek Parfit and David Gauthier regarding the adequacy of individual utility maximizing theories of rational conduct. It will become clear that this debate rests on the sort of difference in systematic view to which I have referred: how a theory of rational action should relate to a theory of the rational person.

Because different systematic approaches have been more thoroughly investigated in moral philosophy, I turn in Section II to a discussion of these. My particular interest is to exhibit how theories of the morally good person have been differently related to theories of right conduct in order better to illustrate by analogy the issue between Parfit and Gauthier. As we shall see, different systematic approaches in moral philosophy line up in interesting ways with analogues in philosophizing about rationality.

I return to the Parfit/Gauthier debate in Section III to demonstrate how the categories developed in Section II provide a deeper understanding of its root issue. Like value-based consequentialisms in moral philosophy, some theories of rationality can be seen as fundamentally end- or value-based.[2] Both their theories of rational conduct and of the rational person derive from an independent view about what rational conduct or being a rational person should bring about or accomplish. A different systematic approach is to begin with an ideal of the rational person and to work towards a theory of rational action from there. Conduct is then held to be rational if it is related in the right way to principles or motives that are characteristic of such a person. As I shall argue, this latter approach to rationality has affinities to the Kantian approach in moral philosophy.

Gauthier's attack on an individual utility maximizing theory of rational conduct, and Parfit's defense against the attack, partly reflect these two different approaches. Parfit's defense assumes an end-based systematic view, and Gauthier's attack assumes that propositions about what an ideally rational person would do are independent evidence, at least, of what it would be rational for him to do. In Section IV, however, I explore a tension in Gauthier's view; for there are both Kantian *and* value-based

elements in his position. What leads Gauthier in the direction of an individual utility maximizing theory in the first place threatens to undermine his attack on it.

Finally, in Section V, I briefly sketch one picture of a theory of rational conduct based on a conception of the rational person. Its underlying motivation is the Kantian one that principles of rational conduct are those by which an ideally self-determining rational being would be guided. Since Kantian theories provide a main alternative to value-based theories in ethics, it seems only prudent to investigate them seriously in the theory of rationality also.

I

In recent years David Gauthier has argued that the theory according to which it is always rational to maximize the agent's utility must be revised since the theory is, in a certain way, self-defeating.[3] Call this theory U. Crudely put, persons who are governed by U are likely to do worse in its own terms than they would if they accepted some other theory.[4] In Prisoner's Dilemma (henceforth PD) situations they will often do better if they accept, and act on, a theory requiring them to keep mutually beneficial agreements, since they will often only be able to make such agreements if they are so governed.

Because this is so, Gauthier argues that agents should accept the theory of "constrained maximization." Call this theory C. According to C, if a person can make a mutually beneficial agreement with others in a PD situation, he should keep it. And in any other situation a person should simply maximize individual utility.

C rather than U is the theory of rational conduct that agents should accept. And this means, Gauthier argues, not only that C is the theory that it is rational to dispose oneself, or to be disposed, to believe. It means that C is the correct theory.

Chapter I of Derek Parfit's *Reasons and Persons* is devoted to a study of theories of rational conduct that are, in his terms, "indirectly self-defeating."[5] The sort of self-defeat that figures in Gauthier's argument against U is an example of what Parfit calls "indirect individual self-defeat." He writes:

If we call some theory T, call the aims that it gives us *our T-given aims*. Call T

> *indirectly individually self-defeating* when it is true that, if someone tries to achieve his T-given aims, these aims will be, on the whole, worse achieved.[6]

If T is U, then its T(orU)-given aim is: the agent's utility being maximized. U is indirectly individually self-defeating, then, just in case if someone tries to maximize her own utility it will be less likely to be maximized.

Parfit distinguishes two subcases. If a person tries, on a given occasion, to achieve her T-given aim, she may fail actually to do what would best accomplish it. So suppose T is U. Someone trying to maximize utility may simply fail to perform acts that are utility maximizing. If a theory is indirectly individually self-defeating in this way, however, this can hardly be thought to count against the theory. The theory still points unambiguously to acts it delimits best even if specific agents on specific occasions cannot discern what that is, or, having discerned it, cannot act appropriately. Any fault here seems to be in the agents and not in the theory.

The more interesting subcase occurs when even though a person who tries to achieve her T-given aims never fails to perform an act recommended by T, nonetheless her T-given aims are achieved less well than they would have been if she had not been governed by T. This is the sort of case with which Gauthier is concerned. He argues that, if people's motivations are relatively "translucent" to others, someone disposed to choose acts conforming to U will sometimes find himself in PD situations where others will be unwilling to make mutually beneficial agreements with him.[7] This will occur whenever others can tell that he is an unconstrained maximizer. Thus anyone disposed to conform to U in such circumstances will be condemned to forego benefits he could have had if he had been disposed to constraint, if, that is, he had been governed by C. Even if each *act* he performs maximizes utility, being disposed to conform to U does not.

But is it an argument against a theory of rational conduct that it is individually indirectly self-defeating in this way? This is where Parfit and Gauthier part company. What is interesting is that divergence on this issue signals divergence on deep and systematic issues of practical philosophy.

To begin to see why, consider Parfit's response to Gauthier's charge that indirect individual self-defeat (of the second kind) counts against a theory of rational conduct. While Parfit does not actually discuss Gauthier's argument directly, he does consider an identical charge made against a

slightly different theory, one he calls the Self-Interest Theory, or S.

Parfit holds that theories of rationality are individuated by the "substantive aim" they give to agents. S's substantive aim is "the outcomes that would be best for [oneself], and that would make [one's] life go, for one, as well as possible."[8] So S's theory of rational conduct is this: an agent should perform all and only those acts whose performance would make his life go best for him. Depending on which theory of self-interest is adopted, S as applied to acts may differ in details from U, but we can ignore these details for the moment.[9]

Parfit concedes that in PD situations straightforward maximizers will often do worse than constrained maximizers. They will do worse not because they perform any act that is worse for them, but because they are *disposed* to maximize, and their being so disposed makes them ineligible partners in mutually beneficial agreements. Being disposed to follow S makes them achieve their S-given aim less well. But does this have any tendency to show that S is a mistaken theory of rational conduct?

Parfit argues that it does not. The essence of his defense of S, for present purposes, is this. A theory of rational conduct says which *acts* are rational. S says unambiguously, for any person in any situation, that the rational act is the one, of those available, that would maximize her utility, since that is always the one most highly recommended by the substantive aim. It may well be that a person who deliberately chooses only maximizing acts, or is disposed unreflectively to maximize, will do worse even though every act she performs is optimific. But S's theory of conduct does not tell a person to *decide* what to do by determining the optimific act. Nor does it tell a person to be disposed unreflectively to maximize. It tells her what to *do*.

If the facts are as Gauthier supposes, and the costs are acceptable, then S will presumably also direct the agent to acquire a disposition to constraint, to be governed by C rather than S (or U). If so, then S will recommend acquiring a disposition to do what is irrational by its own lights. But these recommendations are not inconsistent since they concern different acts: the acquisition of a disposition and its manifestation.

But while S is not self-contradictory, Gauthier argues that it offends nonetheless against a central intuition we have about rationality: *if a fully rational person's "dispositions to choose are rational then surely her choices are also rational."*[10] Because Gauthier uses this intuition to argue that the rationality of an ideal agent's dispositions to choose are inherited by acts that manifest them, we may call it the *Inheritance Principle*.[11]

Although S is not inconsistent, it does violate the Inheritance Principle. Assuming Gauthier's empirical claims to be correct, S entails that a fully rational person would be disposed sometimes to choose acts that are irrational. So, Gauthier argues, S must be rejected. From the fact that it is rational to be disposed to make choices in conformance with C (rather than S or U) and the truth of the Inheritance Principle, it follows that C is the correct theory of rational conduct.

To this line of argument Parfit has several replies. First, he argues that there are situations in which the Inheritance Principle seems plainly false. We shall consider some of these in Section III. He makes, however, two other replies that are more implicitly systematic.

For the Inheritance Principle to have force as an independent argument for a theory of rational conduct two things must hold. First, it must be possible to frame an ideal of the rational person (and his dispositions to choose) independently of a criterion of the rationality of acts that manifest his dispositions to choose: otherwise, the arguments will go in the wrong direction—from rationality of act to rationality of the person. And second, it must be possible to determine an ideal of rational character without depending on something which can also serve as a suitable standard for directly assessing conduct as rational. In that case, the necessary relation between the rational person's dispositions to choose and the rationality of his acts will not be assured.

Parfit raises questions about each of these, especially about the latter. Since S gives agents a substantive aim, this aim can be used to assess both the rationality of acts and the rationality of dispositions directly. Most obviously it can be used to assess the rationality of *acquiring* dispositions, since that, broadly speaking, is an action. But the aim can also be used to assess dispositional states, whether of choice or belief, directly. So, Parfit writes, "S tells each person to believe the theory belief in which would be best for him."[12]

The dispositions to choice it would be rational to have, the dispositions of a fully rational person, are determined, according to S, by the same substantive aim that determines what it is rational to do. Because the rationality of acts and of dispositional states are each determined by reference to some third thing, there is no reason to expect any particular relation between them.[13] A theory of rationality defined by a "substantive aim" will not necessarily satisfy the Inheritance Principle.[14]

Nonetheless, there is a sense, Parfit agrees, in which rational motives or dispositions to choice are not simply whichever would best advance the

substantive aim. Since S gives to agents the aim that their lives go as well as possible there is a sense in which *that* aim, desire, or motive is uniquely or "supremely" rational.[15] In this sense the ideally rational person just *is* one who is governed by S. That is the person whose dispositions to choice themselves embody the correct theory of rationality.

Now in this sense of 'rational motive' what we have (somewhat misleadingly) called the Inheritance Principle will be satisfied. But this is no help to Gauthier, for, as I noted above, the Principle taken by itself does not really express the idea of heritability of rationality in either direction. It expresses only a consistency constraint. Moreover, the sense of 'rational motive' in which Parfit accepts the Principle is one in which the rationality of a disposition to choose is *derivative* from the rationality of conduct that manifests it. The line of thought runs not from a standard of the rational person to a standard of rational conduct, but *vice versa*. Or more properly, it runs from a substantive aim to the dispositions to choice that embody the aim.

It should be beginning to become clear that what is at issue between Gauthier and Parfit is likely to be something much more systematic than a simple disagreement about what it is rational to do in certain circumstances. Parfit's response to Gauthier is informed, at least in part, by the idea that a theory of rational conduct can in principle radically diverge from a theory of rational person in the way Gauthier denies because both are appropriately framed relative to a guiding fundamental aim. And what is essential to Gauthier's argument is the notion that no fact exists about what it is rational for a person to *do* that is independent of what a fully rational person would be disposed to choose.

II

Parfit's defense of S insists on a sharp distinction between a theory of conduct, on the one hand, and theories of how persons should decide what to do, how they should be motivated to act, and what sort of character they should have, on the other. To students of moral philosophy this will have a familiar ring. Writers in the consequentialist tradition have long insisted on a similar distinction between a theory of right and theories of motivation, moral decision, or character. So, for example, Mill wrote that critics of utilitarianism frequently "confound the rule of action with the motive of it."[16]

That consequentialist moral philosophers should cleave to such a distinction is not surprising. After all, act-consequentialist theories of right are subject to a line of criticism that is parallel to Gauthier's critique of U; they can also be indirectly individually self-defeating.[17] A person who tries to bring about the best states may be less likely to succeed even if every act he performs has the best consequences.

This may be for reasons similar to those Gauthier mentions in connection with U, or for quite different reasons. There may be intrinsically valuable states that can only be realized, because they are partly constituted, by motives other than the desire to maximize value. Relations of love and friendship might be an example. If so, someone whose deliberations are typically guided on the spot by an allegiance to consequentialism would be incapable of achieving certain valuable states and, perhaps, of maximizing value.

The standard consequentialist rejoinder is to insist that such facts have no tendency to impugn consequentialism as a theory of right. A principle of right says what a person should do and not how he should be motivated or how he should decide what to do. There is no reason in principle why it cannot be true that the right thing to do in certain circumstances is an act that it would also be right to dispose oneself not to do. And it is consistent, moreover, that a morally good person would be so disposed.

Value-based consequentialists believe they have a particularly good reason for thinking that the latter can happen. A *value-based* consequentialist, again, is someone who holds a consequentialist theory of right to be justified by a view of the good that is independent of and prior to the right.[18] She takes the same view of morality that Parfit takes of rationality: theories of morality, of right conduct, of moral character, the good person, and so on, must have a "substantive aim" to which they answer.

Value-based consequentialists approach the questions, What makes for good character?, and What motives are morally good?, in the same way they approach the question, What makes an act the right thing to do?—by appeal to the good. And because what drives their answer in each case is a relation to some further thing, there is no reason to expect any particular relation between them.[19]

This, again, is fully analogous to the tack Parfit takes in defending S. S is, in effect, a value-based theory of rationality, and so its theory of decision making and of the person will also be dictated by its defining aim.

The analogy between S and a value-based consequentialism in moral theory suggests the possibility that other options in the theory of rationality may line up in interesting ways with familiar systematic views in moral philosophy. Because much of what is at issue in the Gauthier/Parfit debate concerns the relation between rational conduct and the rational person, in what remains of this section I shall explore different systematic approaches in moral philosophy to the relation between right action and moral character. With this framework in hand we shall be in a better position to reconsider the issue between Gauthier and Parfit, and between theories of rationality more generally.

One way of distinguishing systematic approaches in moral philosophy is by how they relate the subject's central notions: the right, the good, and the morally good. Value-based views, as we have seen, begin with an independent conception of the good which is then taken to provide a fundamental justification for any moral theory, whether of the right or of moral character. On a value-based approach a consequentialist theory of right will seem almost inevitable. If the fundamental truths of ethics concern what is intrinsically valuable and worthy of existence for its own sake, then it will seem unavoidable that what one ought to do must be reckoned, in some way, in relation to such value.

Deontological theories of right, on the other hand, are frequently thought to make most sense on a right- or duty-based approach. Just as the value-based consequentialist takes as fundamental certain propositions about what outcomes are good, a deontologist, such as W.D. Ross, may similarly treat certain propositions about what it is right to do, if only *prima facie*.

Both value- and duty-based approaches will regard moral character as derivative. On a duty-based view, the morally good person will be held to be the person who is disposed to do right acts, independently specified, or, perhaps, one who has traits that are likeliest actually to issue in right action. An alternative to value- and to duty-based approaches is to make an account of moral character fundamental and to work towards an account of the right from there. This sort of approach will accept, in effect, the "inheritance" idea for ethics: what it is right to do depends in some way or other on what the morally good person would do.

The more familiar versions begin with a substantive theory of virtue such as Aristotle's. The virtuous person is conceived to have various specific virtues, and what it is right to do is then thought to turn on what the virtuous person would do. Another way of running this general line, however, is to begin as Kant does in Chapter I of *The Groundwork*, with

a relatively formal theory of the moral person. We may think of what is fundamental in Kant's view as an ideal of moral character as self-governing moral integrity. From this account of moral character, of the good will, Kant attempts to argue to the fundamental principle underlying his theory of right: the Categorical Imperative.

Since Kant's good will is itself characterized in terms of a governing concern to do what is right, it may be thought that this cannot be Kant's strategy. But Kant also writes that "the concept of duty ... includes that of a good will."[20] And the first statement of the Categorical Imperative occurs as the conclusion of a pattern of reasoning that, while startlingly telescoped, begins with his account of the good will.[21]

The question with which Kant begins this passage illuminates his strategy:

> But what kind of law can this be the thought of which, even without regard to the results expected from it, has to determine the will if this is to be called good absolutely and without qualification?[22]

Kant apparently thought it possible to base an account of right conduct on a fundamental theory of moral character as autonomous integrity in something like the following fashion. Principles of right must be capable of guiding the conduct of a morally good person, a self-governing moral agent. To play this role they must have, so to speak, the right form. They must be capable of being autonomously chosen, from a standpoint that is impartial between moral agents as such, as suitable to guide the conduct of all.

One way of motivating the sort of broadly contractarian theory of right that Rawls has termed "rightness as fairness," then, is to see it as rooted in something like this way in a fundamental account of the moral person. What is particularly interesting about this strategy is that it aims to motivate a deontological theory of right in some other than a duty-based way.[23] I mention it at some length now because, like Kant, I shall suggest that the same strategy is available in approaching the theory of rationality.

In addition to these different fundamental approaches in systematic moral philosophy, there are other, mixed strategies. A moral theorist may steadfastly refuse to treat any of the right, the good, or the morally good, as fundamental, or even as relatively more central. She may offer her theory simply as a coherent whole without even attempting to give some underlying rationale.[24] Or a theorist may be moved by different rationales in different directions.[25] Nonetheless, the canvassed alternatives—value-

based, duty-based, and both substantive and more formal character-based theories—represent the purest strategies for relating theories within a systematic moral philosophy.

III

With this framework in hand we are in a position to reconsider the debate between Parfit and Gauthier. Before we consider the systematic issues in more detail, however, there are some things we should note about Gauthier's argument.

The argument, recall, is that U, the theory that holds the rational act always to be the maximizing act, must be rejected in favor of C, Gauthier's theory of constrained maximization, because in PD situations where people's motivations are relatively translucent, one will do better if guided by C rather than U. This is so even though one will then be disposed by acceptance of C to perform acts that are less than optimal from one's own point of view. In PD situations each does better if each performs a less than optimal act than each would do if each performed an optimal act. So when there are others who are prepared to cooperate to secure jointly optimal outcomes, but only on the condition that each act in the agreed upon individually suboptimal way, then, if the motivations of each are relatively translucent to the other, each does better to be disposed to cooperate and keep the agreement. Since C differs from U only for such cases, it will be better overall to be disposed to follow C rather than U.

There are a number of things to note about this argument. First, it plainly rests on an important empirical premise, viz., the relative translucency of motivation. And this may be thought to be implausible in enough cases so that when the increased benefits of being able to get away with seeming guidance by C, but actual guidance by U, are added in, the argument is undermined. At least, the argument may be undermined for specific individuals, given the situations they are likely to face, their capacity for dissembling, the costs to them of doing so, and so on. Of course, there may well be other benefits of being guided by C on the other side. Since, however, the issue between Parfit and Gauthier does not turn on these matters of empirical detail, I propose to ignore them.

Second, Gauthier's argument is not made against S, the theory that in any circumstance the rational act is whatever available act would make the agent's life go best for him on the whole. The theory that Gauthier

has in his sights is U, and U makes no distinction between the agent's preferences for his own life and his preferences for anything else. According to U, the rational act is the one that maximizes the satisfaction of the agent's preferences whatever they may be. Moreover, unlike S, U looks only to the agent's preferences, or perhaps to his informed preferences, at the time of choice. It is an example of what Parfit calls a Present Aim Theory. Where S looks to maximize the agent's good throughout his life, U looks to maximize the satisfaction of the agent's present preferences.

Now both of these facts are relevant to the validity of Gauthier's argument. Since what rational acts are supposed to maximize according to U is satisfaction of the agent's preferences, why not say that the disposition it is rational to have in Prisoner's Dilemma situations is simply a *preference*, of the appropriate strength, for cooperating and for keeping one's agreements. There seems no reason why a person may not simply prefer most to keep agreements in all and only those cases where C would require her to do so.[26] Indeed, on some views about preference a person who is guided by C, *will* have such a preference simply by virtue of her dispositions to choose as she does.[27] And even if that view of preference is rejected it seems clear that a person can care about cooperating in such a way that if motivation is translucent, others will be willing to cooperate. That is, whatever dispositions to constraint are necessary to generate cooperation can apparently as well be achieved by a preference for cooperation as by the acceptance of an alternative theory of rational conduct. So it is unclear why the phenomenon Gauthier discusses requires the rejection of U.

This objection to Gauthier's argument rests on the two features that distinguish U from S. By that I mean that the objection could not be raised to a Gauthier-like argument against S. Even if what I dominantly prefer now is to cooperate, S will recommend cooperation only if that is the choice that is likeliest to make *my* life go best *through time*. So if others know that I guide my choices by S, knowing that I have even a very strong present preference to cooperate may not be sufficient to secure their cooperation.

As with possible objections to translucency, I propose largely to ignore these matters also. The issue between Gauthier and Parfit on which I wish to concentrate does not concern them. It concerns rather the question of how rationality of conduct is related to rational agency and motivation. Let us turn to that.

It will help to begin with something of a caricature of Gauthier's argument. Gauthier apparently concludes from the fact that it would be

rational for a person to be motivated to constrain utility maximizing by mutually beneficial agreements that it would be rational for him actually to keep them.[28] Against this inference, Parfit points out that there are many situations in which "there is some motive that it would be both (a) rational for someone to cause himself to have, and (b) irrational for him to cause himself to lose," but also "irrational for this person to act upon...."[29]

One particularly clear case, he points out, is an example of Thomas Schelling's. Suppose you are about to be confronted by a person who will try to gain something you value greatly by threat. The person is only likely to threaten if he believes you to be sufficiently rational to be able to respond to threats, to be able, that is, to make decisions about what to do that take the threat into account. Suppose, moreover, that you have available a drug that induces random behavior for an acceptably short time. By taking the drug motives will be induced that lead you to act irrationally. But it will be rational to become so motivated since you will then be immune to the person's threats. Here there will be motives it is rational to acquire, and irrational to lose, even though they do not motivate rational acts. What one does while under the drug is not made rational by the rationality of being motivated to do it. On the contrary, the motives are rational to acquire precisely because they motivate acts that bear no particular relation to rationality. So if Gauthier's argument depends on an inference from rationality of motive to rationality of conduct so motivated, it must be mistaken.

But recall now exactly what Gauthier's claim is. It is not simply that there is *some* way of being motivated that it is rational to induce. And so he does not infer the rationality of conduct from the rationality of inducing just any motivation that will produce it. His claim, rather, is that there is a disposition to choose acts in conformance with a specific principle of action, C, that will maximize utility.[30] The relevant inference is from the rationality of being disposed to choose in accordance with a principle to the rationality of conduct conforming to the principle. So Gauthier can agree with Parfit about Schelling's case. Indeed, he can offer an explanation of the actions' irrationality, though their motivations be rationally induced: the actions would not result from governance by any recognizable principle of conduct.

Now as I mentioned in passing above, Parfit appreciates this aspect of Gauthier's argument. The principle Parfit actually considers to be at its core is one he calls (G2):

> If it is rational for someone to make himself believe that it is rational for him to act in some way, it is rational for him to act in this way.[31]

At first glance, Gauthier does seem to rely on (G2). Actually, however, Gauthier rejects (G2) in the presentation of the argument in *Morals by Agreement*.[32] There he makes clear that he thinks there to be cases where (G2) is not satisfied, and that this does not affect his argument for C.

The cases where Gauthier believes (G2) to fail all involve various "weakness[es] or imperfection[s] in the reasoning of the actor." (186) For example, a tendency to wishful thinking may lead a person to "confuse true expectations with hopes." (185) In this case it will be rational for the person to guide choice by principles that discount for this tendency to epistemic error. But that does not mean that the rational act is the one that actually accords with the principle that it is rational for the wishful thinker to be governed by. Far from it, she should be guided by that principle because her acts will then be likelier to conform to the undiscounted principle.

However, from the fact that (G2) fails in these cases, it does not follow that a suitably revised principle restricted to "ideal," "perfect," or "fully rational" agents does not hold.[33] The principle that evidently underlies Gauthier's argument for C, then, is neither (G1) nor (G2), but the Inheritance Principle:

> If a *fully rational agent's* "dispositions to choose [to follow a principle of conduct] are rational, then her choices [the acts that accord with that principle] are also rational." (186)

It is utterly crucial to Gauthier's argument against U, then, both that a fully rational person would govern conduct by C, and that if a fully rational person would govern conduct by C, then action that accords with C is thereby rational. From these two premises he can conclude that C, rather than U, correctly describes which acts are rational.

Now since, as a matter of fact, Parfit's objections to (G2) do not turn in any simple way on any "weaknesses" or "imperfections" of the agent, it is appropriate to consider them also as directed at the Inheritance Principle. Parfit raises objections of two sorts.

One is a proposed counterexample to the Inheritance Principle that concerns the rationality of actually carrying out threats as required by principles of conduct that it is rational to dispose oneself to accept. Parfit argues that often carrying out such threats will be irrational, however rational it may

be to dispose oneself to carry them out. In reply, Gauthier simply rejects the counterexample.

Parfit's other objection is more general. It can be no argument for C, he points out, that U (or S) tells a person to dispose herself to be governed by C, for the following reason. Either S(U) is a correct theory of rationality or it is not. If it is, then even though S(U) tells one to become disposed to follow C, C cannot be a correct theory of rational conduct, for by hypothesis S(U) is, and C is incompatible with S(U). If S is not a correct theory, however, then although C may be, the fact that S(U) recommends acquiring a disposition to be governed by C can provide no support for C, since S(U), by hypothesis, is false.[34]

Now if the basis on which C is concluded is that U(S) recommends to any, including a fully rational, agent that he acquire a disposition to govern choice by C, then Parfit's objection seems sound. And Gauthier certainly sometimes presents the argument, even in *Morals by Agreement*, in a way that appears to invite this objection. Thus he writes:

> To demonstrate the rationality of suitably constrained maximization we solve a problem of rational choice. We consider what a rational individual would choose, given the alternatives of adopting straightforward maximization, and of adopting constrained maximization, as his disposition for strategic behavior. (170)

Not surprisingly, the standard of rational choice employed is U. And so Parfit's problem arises: how can C (which is inconsistent with U) be supported by the fact that U would recommend its choice as a principle to guide rational choice?

Fortunately, Gauthier need not confront this problem since there is an alternative construal of his argument that avoids it. What is interesting is that this alternative construal commits him to a wholly different practical philosophy than that embraced by Parfit's defense of S.

"We identify," Gauthier writes,

> rationality with utility-maximization at the level of dispositions to choose. A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition. (183)

What is rational in the first instance for Gauthier are neither acts nor choices, but dispositions to choose: how a rational person *is* rather than what she *does*—her rational character, if you like. The rationality of an

act is derivative from the rationality of a disposition to choose it.[35] Acts are rational, therefore, only in the second instance.

The argument for the rationality of acts conforming to C, then, is not that since the (acts of) acquiring a disposition to govern choice by C is rational (by U or S), then so are the acts that conform to C. That argument would be vulnerable to Parfit's objection. The premise of the argument is rather that a disposition to govern choice by C is itself rational. Such a disposition, under certain circumstances anyway, partly characterizes a fully rational person; it is part of fully rational *character*.

But what about the strategy of showing that C would be the rational principle to *choose* that oneself be governed by? That, Gauthier writes, is only a "heuristic device to express the underlying requirement, that a rational disposition to choose be utility-maximizing." (183)

What is fundamental to the argument, then, is not a principle of *conduct*, U, that recommends acquiring a disposition to choose by C. Any such principle would then be at odds with C. Rather, two distinct ideas are fundamental to the argument. The first is an ideal of the rational person, of rational character. As Gauthier puts it in another place: "[a]n ideally rational actor is one who chooses in such a way that he maximizes the satisfaction of his desires."[36] And the second idea is that rationality of conduct is inherited from the rationality of character: conduct is rational only if it accords with a principle of choice by which an ideally rational person would be guided. Since an ideally rational agent would be governed by C, it follows that C is the criterion of rational action.

This use of the Inheritance Principle signals a different fundamental approach to the theory of rationality than underlies Parfit's defense of S. And so the debate between Gauthier and Parfit is hardly a local dispute within the theory of rational conduct. It concerns a more systematic issue of fundamental approach, and in this it is similar to systematic disagreements in moral philosophy.

Parfit's defense of S, on the one hand, is fully analogous to the way value- or end-based consequentialists traditionally defend their theory of right in the face of similar criticisms. Both may consistently be defended as end-based theories.

For Gauthier, however, no fact exists regarding the rationality of action that is independent of a fully rational person's dispositions to choose. Rational character is prior to rational action. This suggests that Gauthier's approach to the theory of rationality is analogous to character-based approaches in moral philosophy.[37]

But if Gauthier's approach is character-based, what sort of character-based approach is it? It plainly is not Aristotelian. Gauthier offers no substantive conception of the "rational virtues" and certainly no account of the chief good as essentially including them.[38] Might it be a sort of Kantianism?

There are indeed remarks that point in this direction. In stressing that the self-critical character of fully rational agency extends to critical assessment of even the most fundamental principles of conduct, Gauthier explicitly alludes to and embraces Kant's identification of full rationality with autonomy:

> Far from supposing that the choice of a conception of rationality is unintelligible, I want to argue that the capacity to make such a choice is itself a necessary part of full rationality. A person who is unable to submit his conception of rationality to critical assessment ... is rational in only a restricted and mechanical sense. He is a conscious agent, but not fully a self-conscious agent, for he lacks the freedom to make, not only his situation, but himself in his situation, his practical object. Although we began by agreeing with Hume, that reason is the slave of the passions, we must agree, with Kant, that in a deeper sense reason is freedom.[39]

This suggests an analogue to a Kantian character-based approach in moral theory. A fully rational agent is not simply *governed by* principles of any particular sort. Rather, she is self-governing. She *governs herself* by principles. Think of principles of rational conduct, then, as those by which a fully self-critical and autonomous rational agent could govern her conduct. To meet this test a principle must be one such an agent could accept on the basis of her own critical reflection. Fix, then, the features of the relevant reflection and acceptance. For Kant: the principle must be choiceworthy as action-guiding for all from a standpoint that is partial to none. Rational conduct is whatever accords with such principles.

But if there are Kantian elements in Gauthier's approach there are also other elements of his basic strategy that are profoundly unKantian. While he makes use of the Kantian ideal of a self-governing rational person to motivate the heritability of rationality from character to conduct, there are other ways in which his account is ultimately end-based, ways, I shall argue in the next section, that create a disturbing tension in his overall view.

On a Kantian character-based approach in moral philosophy, an ideal of moral character, of moral autonomy or self-governance, is specified independently of any particular theory of value or right. So on the analogous approach to rationality, an ideal of the rational person would be specified independently of any theory of rational conduct or of any independent end that rational personality or conduct is conceived to serve as such. But while Gauthier's ideal of the rational person is of course independent of a theory of rational conduct, it is not independent of any substantive end.

Recall what makes a disposition part of rational character for Gauthier. "A disposition is rational if and only if an actor holding it can expect his choices to yield no less utility than the choices he would make were he to hold any alternative disposition." (182) And more generally: "An ideally rational actor is one who chooses in such a way that he maximizes the satisfaction of his desires ...."[40] So what makes a given disposition to choose part of rational character is that it serves a specific end, an end that can in principle be specified quite independently of any ideal of rational conduct or personality.

So Gauthier's approach is not really character-based, after all. Its account of character is more basic than its account of conduct, but there is something more basic still. And what is most basic is precisely the same as in an end-based defense of U: the agent's utility being maximized.

Gauthier's approach actually combines two quite distinct elements. As against Parfit, he holds that whether an act is rational is not determined by whether some objective relation obtains between the act and good outcomes. Rational acts are supported by reasons. And whether something is a reason depends on whether it would register as a reason in the deliberations of an ideally rational agent. This aspect of Gauthier's thought stresses a connection between rational conduct and rationally self-critical self-government. It is apparently analogous to the Kantian approach in ethics.

But there is also a crucial disanalogy with the Kantian approach. It is essential to Kantianism that its ideal of the person be specified independently of any aim that persons are supposed actually to achieve as such. After all, if what makes a person rational is that he conducts himself in a way that achieves a given end, then why not also say that what makes an act rational is that it achieves the end? That is, why not have a fully end-based theory of rationality, such as Parfit's S or an end-based U?

But Gauthier's ideal of the rational person is not an independent ideal. It is based on the aim of maximal agent utility. Gauthier's view is ultimately end-based.

These distinct elements create a deep tension in Gauthier's view, for they pull in quite different directions. If what is crucial to a rational act is not that it achieve any particular outcome, but that it be supportable by principles that would figure in the deliberations of an ideally rational agent, then why should an ideal of the rational person, or of rational deliberation, turn on whether *they* achieve a given outcome? On the other hand, if it is determinative of an ideally rational person or of rational deliberation that they achieve a given outcome, then why is it not equally determinative of whether an act is rational?[41]

Of course, it is open to someone simply to *define* 'rational conduct' as what accords with principles that a fully rational person would be governed by, and then take an end-based approach to rational personality. But even if we could no longer sensibly ask whether it would be *rational* to act on dispositions that are utility-maximizing, we could still use some other term to raise the old question: Is that the act we really should perform? And it seems that if what governs our approach at the most basic level is a specific end, then the natural answer to that question will be: only if it serves the end. That is, we shall be led to U as a theory of a conduct and not C.

The situation here is fully analogous to that of end-based consequentialism. Rule-consequentialism will seem a less well-motivated theory of right than act-consequentialism on an end-based approach. Again, one may simply define 'right' in such a way that rule-consequentialism will seem the appropriate account or the defined concept on an end-based approach. For example, like Mill, we may define right conduct as that which accords with utility maximizing sanctioned rules. But we will still be able to ask: I know this is *right* (as stipulatively defined), but should I do it? And if our fundamental approach is end-based it is hard to see how we can sensibly give any other than the act-consequentialist response: only if it will have the best consequences.

V

Because Gauthier's theory has two quite different motivations, it threatens to come apart. Those who agree that a theory of rationality must be responsible to an independent aim will be likely to agree with Parfit's defense and accept U rather than C. And those who agree with Gauthier

against Parfit that the rationality of an act does not depend simply on its relation to any specific outcome, but on whether it would be recommended by reasons that would weigh with an ideally rational agent, are likely to think that an ideal of the rational person, or of rational deliberation, ought not to depend on its tendency to achieve some outcome either. They are likely to reject an end-based view of character in favor of some more fundamentally character-based view.

In this final section I shall briefly sketch the outlines of one such approach, that is, a strategy that is fully analogous to that employed by a Kantian character-based ethical theory.

The Kantian approach conceives of full rational agency as self-critical self-government by principles. A fully autonomous rational agent submits even his most ultimate principles of conduct to critical assessment. And conduct will be rational if, and only if, it accords with principles that are acceptable on reflection to a fully rational agent. But what does 'acceptable' mean here? And what sort of reflection is the appropriate kind?

On a Kantian character-based approach, no fact exists about what are correct principles of rational conduct that is independent of what would be acceptable on reflection to fully self-critical and self-governing rational agents. Nor is there any fact about what *external* ends rational personality pursues. In this way Kantianism is deeply internalist; it rejects the idea that the relevant sort of acceptability is like the recognition of some truth independent of suitably constrained autonomous choice—like, for example, accepting that nine is the number of the planets.

The relevant acceptability is *practical* acceptability. And, on the Kantian approach, this is a matter of whether a principle could be chosen as action-guiding from a suitable standpoint.[42] Since ultimate principles apply to all if, and only if, they apply to any, the relevant standpoint will be one that is impartial between rational persons as such. And the relevant question will be: What principles are choiceworthy to guide the conduct of all when a choice is made from a standpoint that is impartial between them?

But choiceworthy in virtue of what? Must not a Kantian theory, like any theory of rationality or morality, ultimately appeal to a theory of value or to some end at this point? Yes, but not in the way that end-based theories do. For what will be appealed to will be something like a theory of the interests of rational agents as such, and not an end or theory of value that is itself constituted independently of rational or moral conduct or agency.

The model I have in mind is Rawls's "original position" with a thicker veil so that it expresses the idea of a perspective that is impartial between

rational persons as such. Something must guide the choice of principles; a "thin" theory of the good is required. But the relevant interests guiding the choice are interests intrinsic to rational personality; the latter is not merely an instrument to serve the former.[43]

The rationality of conduct, on the present suggestion, will be a "purely procedural" matter in much the same way that the justice of institutions is on the Rawlsian approach. Whether conduct is rational will depend on whether it accords with principles that would be the outcome of an ideal procedure: the best choice from behind a veil of ignorance among alternative action-guiding principles in the light of interests that a person has *qua* rational agent.

So a Kantian approach to rationality will, not surprisingly, be the same as its approach to morality. Indeed, for the Kantian, rationality and morality will not be two distinct things. Since the Kantian conceives of morality as binding on rational persons as such, it will mark no distinction between the moral and practically rational 'ought', at least when the latter is not simply the 'ought' of relative rationality.[44]

This sketch is much too compressed, of course, to be convincing, but my purpose here is less to defend it than to place it within a framework of theoretical alternatives.[45] Like Gauthier's theory it conceives of rational action not as whatever would achieve a specific outcome, but as conduct that is recommended by reasons that would weigh in the deliberations of an ideally rational agent. But unlike Gauthier's theory, its ideal of the rational agent and of rational deliberation is not end-based either.

I remarked in Section IV that Gauthier's case against U is threatened by conflicting rationales. The theory I have sketched avoids this problem. It is, I believe, the theory that should be embraced by those who are attracted by the idea that to be a reason is to be a consideration that would weigh with an ideally rational agent.

NOTES

1. An earlier draft of this paper was presented at a conference held at Virginia Polytechnic and State University in Blacksburg in May of 1985. I am indebted to my commentator, James Klagge, and to other participants for helpful comments. I am also grateful to David Velleman, Don Regan, and Peter Railton for useful discussion.

2. By value-based consequentialism I refer only to consequentialist theories of right that are held to be justified by a theory of objective or intrinsic value—of the value that something, e.g., an event, or state of affairs, can have in itself and not simply for someone. Moore is, of course, the best example. Many consequentialists, for example R.M. Hare and Richard Brandt, are plainly not value-based consequentialists in this

sense. Others are more difficult to peg. What should be said about a theorist who abjures a notion of objective or intrinsic value in favor some person-relative notion such as welfare, or good-for-x, and who holds that acts are right insofar as they promote total good, so reckoned? Typically, unlike Moore, such a theorist will feel the need to provide some rationale over and above the mere fact that such acts will promote the good, and to that extent their view may not be value-based in the same simple sense that a Moorean view is. For Moore, propositions about objective or intrinsic value just are propositions about what "ought to exist for its own sake" and hence what is worth, to some extent or other, the efforts of agents. The problem is less whether anything's having value in that sense would be relevant to what it would be right to do, and more whether there is or can be such value. If one rejects such a notion and holds a welfarist or 'good-for-x' consequentialism, there will be less question about whether such value exists, but there will be a correspondingly larger question why it is right to maximize it and wrong not to. Typically more justification will seem necessary than that provided by propositions of welfare or person-relative value themselves.

3. First in "Reason and Maximization," *Canadian Journal of Philosophy* 4 (1975), 411-433, and most recently in *Morals by Agreement*, (Oxford: Oxford University Press, 1986), Chapter VI.
4. The formulation, 'being governed by a principle' (which is mine) is unhappily vague. Gauthier sometimes speaks simply of dispositions to choose (acts that accord with a given principle), sometimes of "seeking" to maximize or to constrain maximization, and sometimes of "accepting the rationality of" either straightforward or constrained maximization. There are differences among these, and between any of them and Parfit's formulation below: "trying to achieve" a given aim. With one major exception, I shall have largely to ignore these with the hope that they do not affect the main points at issue.

   There is one difference that will be important, however. We can mark this difference as that between a person who is simply governed by P and a person who *governs himself* by P. This is the difference, roughly, between someone who is simply disposed to choose acts that conform to P and someone who is disposed autonomously to conform to P because of his reflective acceptance of the principle. Gauthier's talk of "accepting the rationality of" maximization or constrained maximization, and acting accordingly, suggests this latter idea. Moreover, it looms large in the Kantian approach to rationality generally.
5. *Reasons and Persons* (Oxford: Oxford University Press, 1984), Chapter I.
6. *Ibid.*, p. 5.
7. The assumption of relative translucency of motivation was neither made explicit nor defended in the version of the argument in "Reason and Maximization," but is in *Morals by Agreement*.
8. *Reasons and Persons*, p. 3.
9. So for the present I will suppose (inaccurately) that both S and U direct the agent to maximize his utility. Two differences will become salient later. First, U directs the satisfaction of the agent's preferences, regardless of their objects, and regardless of the effect satisfying them has on the goodness of a person's life for her. Second, U directs the person to satisfy her *present* preferences (at the time of action). Preferences that she will have, but does not now have, have no intrinsic weight even if satisfying them will make her life go better for her.
10. *Morals by Agreement*, p. 186. Further references to this work will be placed parenthetically in the text. As will become clearer in Section III, the ideality of the agent is crucial to Gauthier's argument.

11. Actually it is mistaken to say that the principle alone expresses the idea that rationality is heritable by acts from the dispositions to choose of an ideally rational agent. Taken by itself, the principle only requires a kind of consistency between rationality of conduct and rationality of agency. Someone might accept the principle as a consistency constraint on theories of rationality without accepting the sort of dependence of rationality of conduct on rationality of character that, as we shall see in Part III, Gauthier proposes. For this point I am indebted to David Velleman.
12. *Reasons and Persons*, p. 45.
13. Note that there are three distinct things here that can be assessed for rationality: a disposition (say, to choose certain acts), the acquiring of the disposition, and acts that manifest the disposition. It will be important to keep these three separate. Whether, and how, they are to be related is precisely what is at issue. We may regard Parfit as proposing that any relation between these three is possible. It is possible, for example, that it is rational to do something, that it would be rational to be disposed not to do, but also that it would not be rational to acquire a disposition not to do (because, say, of the costs).
14. Thus it will be at odds with any thesis of dependence between rationality of dispositions to choose and rationality of conduct manifesting such dispositions, whether dependence in either direction or interdependence.
15. *Ibid.*, p. 8.
16. John Stuart Mill, *Utilitarianism*, ed. G. Sher (Indianapolis: Hackett, 1979), p. 17.
17. This sort of self-defeat is not the only, nor even the primary, sort with which consequentialism is typically charged.
18. See footnote 2.
19. This, of course, is oversimplified. An account of moral character might on a value-based approach be related to the good not immediately, but through activities of praise and blame that give propositions of moral goodness content, which activities are assessed by the value of their consequences.
20. Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. H.J. Paton (New York: Harper Torchbooks, 1964), p. 65; *Preussische Akademie* p. 397.
21. *Ibid.*, pp. 69-70, *Ak.* p. 402.
22. *Ibid.*
23. Rawls discusses "rightness as fairness" in *A Theory of Justice* (Cambridge: Harvard University Press, 1971), p. 111. For his discussion of the relation of his normative views to an underlying conception of the moral person, see "Kantian Constructivism in Moral Theory," *The Journal of Philosophy* 77 (1980), 515-572. I discuss this sort of approach to deontology in "Agent-Centered Restrictions from the Inside Out," *Philosophical Studies*, 50 (1986), 291-319.
24. I noted above in footnote 11 that someone might accept the Inheritance Principle as a sort of consistency constraint on theories of rationality, without viewing either rationality of dispositions to choose or rationality of acts as fundamental. One might simply believe them to be interdependent. The analogous remarks hold for moral theory. One might hold a consistency principle relating morally good dispositions to choose and right conduct without believing either to be fundamental. The same question will arise for either view: In virtue of what are these things interdependent?
25. So it is possible through one sort of mixed strategy, for example, to be led to a value-based theory of moral motives without adopting a wholly act-consequentialist theory of conduct. One may hold the sort of position Robert Adams calls "conscience utilitarianism": "we have a moral duty to do an act, if and only if it would be demanded of us by the most useful kind of conscience we could have." (See Adams, "Motive Utilitarianism," *Journal of Philosophy* 73 (1976), 479.) Rule-consequentialism may be regarded

similarly. I express doubts about how well-motivated such mixed strategies can be at the end of Section IV and the beginning of Section V.

26. Note that it is important to say "a preference for keeping one's agreements" rather than simply a preference for "agreement-keeping" to avoid the objection that the latter preference may not be a sufficient security for cooperation, since others may have no reason to cooperate when it is believed that one can by not cooperating oneself bring about greater cooperation overall. Likewise there may be a similar intrapersonal problem with the preference for keeping one's agreements. Others may fear that I will be disposed to violate a present agreement with them, and so be unwilling to make it, if they believe that I can by breaking the present agreement make it the case that I will break fewer agreements in the long run. But it is not clear that there is no preference, knowledge of which will do as good an assurance job as the knowledge that one accepts C. That will face similar problems.

    I am grateful to Michael Bratman for raising this issue.

27. Gauthier makes it clear that he does not accept this "revealed preference" view. See *Morals by Agreement*, pp. 26f. Preferences are "attitudinal" and our choices may or may not reveal them in any given case.
28. I say "apparently" because this is not precisely Gauthier's argument, as will become clear below.
29. This is a denial of a principle that Parfit labels (G1). (Does Parfit intend 'G' here to stand for "Gauthier"? Not necessarily. Parfit discusses Principle GI before he discusses Gauthier's argument [pp. 17-24]. He recognizes the argument to be more complex than we have so far presented it. So when I say that Parfit objects to "this inference" I do not mean to imply that he takes it to be Gauthier's inference.)
30. As I noted above in footnote 4 there is some ambiguity about exactly what sort of disposition Gauthier has in mind. At the very least he seems to mean: a disposition to choose acts that the agent believes to have features that, as a matter of fact, are those that make an act rational according to P. It is apparently not sufficient that an agent be disposed to choose acts that conform with P but where the mechanism of conformance is something other than the agent's desires and beliefs with respect to "P-conforming features." The formulation in the text should be understood to include this.
31. *Reasons and Persons*, p. 23.
32. Since this was published after *Reasons and Persons* Parfit could not, of course, be said to have misunderstood Gauthier's argument.
33. Gauthier uses each of these terms in different places.
34. Parfit credits Shelly Kagan with this objection.
35. "We shall consider whether particular choices are rational if and only if they express a rational disposition to choose." (183).
36. "The Unity of Reason: A Subversive Reinterpretation of Kant," *Ethics* 96 (1985), p. 85.
37. Note that in suggesting that the disagreement between Parfit and Gauthier over the Inheritance Principle signals a systematic disagreement analogous to that between value-based consequentialist and character-based nonconsequentialists, I am not suggesting that it is analogous simply to the debate between consequentialism and nonconsequentialism. For one thing, a duty-based nonconsequentialist may reject the Inheritance Principle.

    In his comments on this paper at the Blacksburg conference, James Klagge suggested that the deep difference between Gauthier and Parfit is their different attitudes towards realism—Parfit's defense of S being committed to realism, Gauthier rejecting it. I think there is something to this, but baldly stated it is misleading. First, the difference, as I

see it, is not simply between realism and irrealism. There is nothing about irrealism per se that would incline one towards the Inheritance Principle. Now it is true that what Rawls calls *Kantian constructivism* does lead in the direction of the Inheritance Principle, but that is not primarily because it is not a realism, but because it is a character- or person-based approach. See "Kantian Constructivism in Moral Theory." And it is possible to hold that there are indeed facts of the matter in ethics (and, analogously, about what is rational), but that these fundamentally concern the moral person, and that principles of right are derivative.

38. This aspect of Aristotelianism, its mutually dependent accounts of character and the chief good, blurs the line between end-based and character-based theories.
39. "Reason and Maximization," 431. There are similar remarks in *Morals by Agreement*: "At the core of our rational capacity is the ability to engage in self-critical reflection. The fully rational being is able to reflect on his standard of deliberation, and to change that standard in the light of reflection." (183)
40. See note 36.
41. Especially since, unlike an Aristotelian account, the end is not itself inclusive of actions guided by any particular principle. Both conduct and personality are instrumental with respect to it.
42. Unlike conduct-based and end-based strategies, since a Kantian approach takes as fundamental an ideal of the rational person as self-governing, it must conceive of principles of rational conduct as action-guiding.
43. Here I take it that we can think of the Rawlsian notion of primary goods as interests of autonomous rational persons as such. For Rawls's discussion of this element of his view see "Social Unity and Primary Goods," in A. Sen and B. Williams, *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982).
44. By 'relative rationality' I mean the notion involved in hypothetical imperatives: that taking the means is rational relative to the (nonrelative) rationality of achieving an end. As far as any hypothetical imperative is concerned, it is equally rational to abandon the end as it is to take the means. Note that S, U, and C purport not simply to be theories of relative rationality.
45. I say more, but still not nearly enough, in *Impartial Reason* (Ithaca: Cornell University Press, 1983).