2

Rationality and The Rational Aim

David Gauthier

I

'Many of us want to know what we have most reason to do. Several theories answer this question' (p. 3). Wanting to know what one has most reason to do might be understood simply as wanting to know what to do. But if I want to know what to do, a theory, whether about rationality or about morality, will not answer my question. Parfit's question must be understood another way. He supposes that S, the Self-interest Theory, gives this answer: 'What each of us has most reason to do is whatever would be best for himself' (p. 8).[1] One might then understand the question as asking what considerations give one sufficient reason for acting. The Self-interest Theory answers: considerations about what would be best for oneself.

Or so Parfit may claim. But let us consider what he tells us about S, as a theory about rationality:

> We can describe all theories by saying what they tell us to try to achieve. According to all moral theories, we ought to try to act morally. According to all theories about rationality, we ought to try to act rationally. Call these our *formal* aims. Different moral theories, and different theories about rationality, give us different *substantive* aims.
>
> By 'aim', I shall mean 'substantive aim'. . . . S gives to each person this aim: the outcomes that would be best for himself, and that would make his life go, for him, as well as possible. (p. 3)

A person who achieves the formal aim given him by S acts rationally and is rational. A person who achieves the substantive aim given him by S has his life go for himself as well as possible. How are these related? A natural supposition would surely be to treat the substantive aim given by

a particular theory of rationality as its specification of the formal aim. Each theory of rationality provides an account of what it is to be rational; this account is formulated in the substantive aim that it gives each person. Thus, according to S, to be rational is to have one's life go for oneself as well as possible.

But this is not Parfit's view. He says: 'According to S, our formal aim is not a substantive aim' (p. 9). And he insists that S does not 'give to each person *another* substantive aim: to be rational, and to act rationally' (ibid.). Now of course, if the substantive aim were a specification of the formal aim, then it would be true that S did not give to each person *another* substantive aim, but in giving to each person the aim that her life go as well as possible, it would thereby give her the aim of acting and being rational. And Parfit denies this. Indeed, he says that 'In the case of some people, according to S, being rational would *not* be part of what makes their lives go better' (p. 10).

So what is Parfit's account of the relation between the formal aim, acting and being rational, and the substantive aim, having one's life go as well as possible? To answer this, we must know what, on Parfit's understanding of the Self-interest Theory, it is to be rational. But what can it be if not to act so that one's life goes as well for oneself as possible? And if this is what it is to be rational, then how can Parfit suppose that 'being rational would *not* be part of what makes their lives go better'? For if to be rational is to act so that one's life goes as well as possible (henceforth I shall take 'for oneself' as read), then surely if one is rational, one's life must go as well as possible. If one acts so that one's life goes as well as possible, then one's life goes as well as possible.

This last argument moves too quickly. But I shall defer discussing this; for the present we should note that Parfit does not give us a direct account of what it is to be rational. However, we might construct one for him, by taking what he says about rational actions, desires and dispositions, and supposing that, to be rational, one must always do what is rational, and have both 'the supremely rational desire' and 'the supremely rational disposition' (p. 8). And we may now apply this to S. Parfit says that according to S, 'What it would be rational for anyone to do is what will bring him the greatest *expected* benefit' (ibid.). He also says that according to S, 'The supremely rational desire is that one's life go as well as possible for oneself,' and 'The supremely rational disposition is that of someone who is never self-denying,' where to be never self-denying is never to do what one believes will be worse for one (ibid.).

On this reading, the formal aim given by S is that one always do what will bring one the greatest expected benefit, and desire that one's life go as well as possible, and be disposed never to do what one believes will be

worse for one. And the substantive aim is that one's life go as well as possible. Parfit's examples show that the substantive aim is not a specification of the formal aim, and indeed that someone who adopts the formal aim may well fail to achieve his or her substantive aim. Suppose I am never self-denying. Then, if I promise to do what at the time of keeping or breaking my promise would be worse for me, I shall break the promise. Suppose I know that I am never self-denying. Then I cannot promise sincerely to do what would be worse for me. Suppose I am transparent. Then I cannot convincingly purport to promise to do what would be worse for me. But as in Parfit's example of my car breaking down in the desert (p. 7; I shall henceforth refer to this as 'the desert breakdown case'), it may be greatly to my advantage to make such a promise. Only a convincing promise to pay you a large reward will induce you to drive me out of the desert, and I have no other way out. So my life will go better if I make such a promise. If I am trustworthy, 'disposed to keep my promises even when doing so will be worse for me' (p. 7), I can make a convincing promise. If I am never self-denying, I cannot. So my life will go better if I am trustworthy, rather than never self-denying. But the rational disposition is to be never self-denying. If I follow my substantive aim, I shall, if I can, make myself trustworthy. But if I follow my formal aim, I shall be never self-denying. Thus the substantive aim is not a specification of the formal aim, and someone who adopts the formal aim may fail to achieve his substantive aim.

This argument may seem insufficient. In the desert breakdown case I do better to be trustworthy. But does this show that it is better for me to be trustworthy than to be never self-denying? Someone might object that even if being trustworthy is sometimes beneficial, at other times it is costly. Suppose you are gullible. You believe whatever you are told. Then, if I am never self-denying, I can falsely promise to reward you if you drive me out of the desert, knowing that I will not pay you. Whereas if I am trustworthy, I cannot avoid paying if I promise you a reward, and so do worse. The objector grants that if we restrict our attention to a particular case, we may think that the formal and substantive aims diverge, but he claims that from an overall standpoint the aims coincide.

A first response to this objection is that the overall benefits of being able to promise sincerely to do what will be worse for me may reasonably be expected to outweigh the overall costs of keeping promises when one could have got away with insincerity. A person need not make promises except when she expects to benefit thereby, and if she is rational (by the standard of S), she will not make promises except in such contexts. A person who frequently made foolish promises might of course suffer from his trustworthiness, but in making foolish promises, this person would already be acting in a way that was not best for him. Someone who could count on herself to make promises only when it would be best for her could expect to benefit from being trustworthy.

A further response rests on the fact that even generally trustworthy persons have been known to make false or insincere promises. There are occasions that call for trustworthiness; on other occasions one may do better not to be self-denying. Since our concern is with rationality and not morality, we need not hesitate here over the thought that selective trustworthiness may be a less than admirable characteristic. We need only note that a person must expect to do better overall if she is disposed to be selectively trustworthy rather than never self-denying, assuming of course that she is reasonably astute.

The objection fails. And so Parfit insists that although according to theory S never to be self-denying is the supremely rational disposition, yet being never self-denying is not part of the substantive aim that S gives to many, if not all, persons. Thus, according to Parfit, S may be indirectly individually self-defeating. 'It can be true that, if I try to do whatever will be best for me, this will be worse for me' (p. 5). And this is not because I will fail to do what is best for me. 'Even if I never do what, of the acts that are possible for me, will be worse for me, it may be worse for me if I am purely self-interested [i.e., never self-denying]. It may be better for me if I have some other disposition' (ibid.).

Parfit further claims that 'S implies that we cannot avoid acting irrationally' (p. 13). I shall not summarize his rather convoluted discussion, but reconstruct the argument in somewhat different terms. It is irrational for anyone to do what he believes will be worse for himself. It is therefore irrational for anyone to perform a self-denying act. But it is worse for one to be never self-denying. If one is never self-denying, then one has acted irrationally in not trying to acquire some other disposition, such as trustworthiness, that it would be better for one to have. If one has some other disposition, then one acts irrationally in performing the self-denying acts towards which one is sometimes disposed. Thus, whether or not one is never self-denying, sometimes one acts irrationally.

This argument is not conclusive. For it may be that one can do nothing about one's dispositions. If so, then one may be never self-denying without having acted irrationally. I shall therefore replace Parfit's claim by a weaker one: S implies that to the extent that the disposition to be never self-denying is in our control, we cannot avoid acting irrationally. According to the account of rationality that I have ascribed to Parfit, a rational person has the supremely rational disposition, and so the claim has this corollary: S implies that to the extent that the disposition to be never self-denying is in our control, it is rational to make oneself

irrational, and irrational to remain rational. For the supremely rational disposition is to be never self-denying; but it is rational to acquire another disposition, thereby becoming irrational, and irrational to remain never self-denying and rational.

Earlier I questioned the view that if one acts so that one's life goes as well as possible, then one's life goes as well as possible. We may now see why. When I act so that my life goes as well as possible, I am not being self-denying. But if I am never self-denying, then I must expect my life to go worse than if I am disposed to perform acts some of which are self-denying. In so far as it is within my power to affect my dispositions, then, my life will go as well as possible only if I bring it about that I sometimes act so that my life does not go as well as possible.

II

Although S may be *indirectly* individually self-defeating, Parfit denies that it can be *directly* individually self-defeating. He says (p. 55) that if S were directly individually self-defeating, then it would be certain that if someone were successfully to follow it, he would thereby cause the substantive aim given him by S to be worse achieved than if he had not successfully followed it. But this is not so:

> S gives to me at different times one and the same *common* aim: that my life goes, for me, as well as possible. If my acts at different times cause my life to go as well as possible, I must in doing each act be successfully following S. I must be doing what, of the acts that are possible for me, will be best for me. So it cannot be certain that, if I always successfully follow S, I will thereby make the outcome worse for me. (p. 55)

Parfit's definition and argument may both seem puzzling. S may be indirectly individually self-defeating, yet it is not *certain* that if someone tries to achieve his S-given aim, that aim will be worse achieved than if he were disposed in some other way. Parfit shows that it is rational to expect this to occur, but nevertheless someone might never find himself in circumstances in which a self-denying disposition would benefit him. Why, then, does Parfit claim that for S to be directly individually self-defeating, it would have to be certain that a person who successfully follows it will cause his S-given aim to be worse achieved than if he failed in some particular way successfully to follow it? I agree with Parfit that success in following S does not guarantee that one will cause one's S-given aim to be worse achieved than it might otherwise be. But is it possible that someone might be successful in following S and yet thereby

cause his S-given aim to be worse achieved than it might be? And if it is possible, then surely S could be directly individually self-defeating.

Recall Parfit's claim that S implies that we cannot avoid acting irrationally. If this is true, then it is not possible always successfully to follow S. If the disposition never to be self-denying is in my control, then if I remain never self-denying, I do what I expect to be worse for me, thus failing to follow S, or I cease to be never self-denying, and must expect sometimes to do what is worse for me, thus failing to follow S. If I begin by successfully following S, then I bring it about that I do not always follow S. So if the disposition never to be self-denying is in my control, then I must sometimes do what is worse for me, and the question whether S is directly individually self-defeating does not arise.

What if the disposition never to be self-denying is not in my control? Then if I have this disposition, it may be that I always act rationally, and so successfully follow S. It is worse for me that I am never self-denying, but nothing I do makes the outcome worse for me than something else I might do. And so Parfit's claim seems to be true, in that, in so far as it is possible always successfully to follow S, it is not directly individually self-defeating.

This last argument may seem mistaken. If I am never self-denying, then surely I do what is worse for me. For example, in the desert breakdown case I do not make the sincere promise that would elicit your assistance. What I do is worse for me than making that promise. But does this show that I do what is worse for me? I *cannot* make the promise. Knowing myself to be never self-denying, I cannot sincerely promise to reward you for driving me out of the desert. What I can do is limited by my disposition. I do what is best for me given that I am never self-denying. That it would be better for me were I differently disposed does not show that I fail successfully to follow S.

Is this right? Is what a person can do limited by his or her dispositions? If I am never self-denying, does it follow that I can perform only non-self-denying actions? Surely this is wrong; a person with a cowardly disposition may on occasion perform a courageous action. But this analogy misses the real point. Suppose that I am *firmly* disposed to be never self-denying. Then in deliberating about what to do, I consider the various alternatives, choosing an action that affords me an expected benefit at least as great as that of any action that I believe possible for me. The criterion of possibility here cannot include the condition that the action not be self-denying. An action can be shown not to be self-denying only by comparing its expected outcome with that of the alternative possible actions. Thus I select from the members of the set of possible actions one that, relative to that set, will involve no self-denial. Suppose that I am considering whether to pay you a reward for driving me out of the desert.

I take both paying you and not paying you to be my possible actions, and I consider which affords me the greater expected benefit and so involves no self-denial. If I choose not to pay you, believing it to afford me the greater expected benefit, I do not suppose that paying you would be impossible simply because it would be self-denying.

Now consider promising to pay you a reward. The claim is that it is not possible for me to *sincerely* promise to pay you a reward, if I believe that I am never self-denying and that paying you the reward would involve self-denial. But why? I believe that it is possible for me to pay you the reward. Indeed, I believe that I shall pay you the reward if paying you the reward will lead to my life going as well as possible. So why is it not possible for me to promise, even though I am disposed to be never self-denying? If my disposition does not make it impossible for me to pay you, why should it make it impossible for me to promise to pay you?

The answer, I think, is that promising requires that one suppose not only that it is possible to perform the promised act, but that one will perform it. It is not possible for me to promise – sincerely – to pay you, while holding the belief that I shall almost certainly not pay you. But if I am aware that I am disposed to be never self-denying, and that paying you would involve self-denial, then in all likelihood I hold the belief that I shall almost certainly not pay you. And then it is not possible for me to promise sincerely to pay you. Holding the belief that I shall not pay you excludes promising to pay you from the set of possible actions over which I deliberate. I do not decide against promising to pay you by assuming it to be possible and then finding that, relative to its alternatives, it would involve self-denial. Rather, I rule it out by realizing that given my beliefs about my dispositions and about what other actions would involve self-denial – beliefs that I hold prior to comparing promising to pay you with the alternatives – I cannot form the requisite intention. A person's possible actions are not directly limited by what she may be disposed to do, but they are indirectly limited by what, given her knowledge or beliefs about her dispositions, she can form the intention to do.

We may now return to the main point. Parfit claims that S is not directly individually self-defeating. As I have said, he seems to be right. A person who is able to affect whether she has the disposition never to be self-denying cannot always successfully follow S, so that the question of being worse off should she follow it does not arise. And a person who has the disposition but is unable to affect having it can successfully follow S without thereby causing herself to be worse off, but she is worse off in virtue of having the disposition. This latter person may be rational; according to S, as Parfit understands it, she does have the supremely rational disposition. But she is *cursed* by her rationality.

III

This conclusion should make us suspect Parfit's account of rationality. Even if we are willing to admit that rationality may not be an unmixed blessing, we should, I think, admit this with reluctance, and only when we are satisfied that our admission does not rest on misunderstanding rationality. I believe that there is a better understanding than Parfit offers, and one that avoids at least some of the unwelcome consequences of his account. I cannot fully develop such an understanding here, but taking comfort in the fact that Parfit's own discussion is sketchy, I shall offer an equally sketchy alternative.

My starting-point is the relation between the formal and the substantive aim given each agent by a theory of rationality. I have no objection to Parfit's idea that theories of rationality may be characterized primarily in terms of such aims; rather, I want to insist, as I suggested at the outset, that the substantive aim given by a theory of rationality is its particular way of giving content or substance to the formal aim that all such theories share. Thus I propose that we interpret S, the Self-interest Theory, as giving to each agent the aim that his life go as well for himself as possible, this being its way of specifying the formal aim of being rational that it must give him in so far as it is a theory of rationality.

I shall suppose that to be rational one must always do what is rational, and have both the supremely rational desire and rational dispositions. Although this is somewhat similar to the account of rationality that I ascribed to Parfit, I give a different content to what it is rational to do, and I do not speak of any disposition as supremely rational. Let us consider the question of rational dispositions first. Parfit supposes that for theory S it is possible to specify a disposition as rational without regard to the agent's circumstances, by relating it straightforwardly to the substantive aim given by the theory. Thus he supposes that, according to S, the supremely rational disposition is 'that of someone who is never self-denying' (p. 8). I propose, however, that a disposition is rational if and only if having it is most conducive to one's substantive aim. S gives one the aim that one's life go as well as possible; it therefore claims that a disposition is rational if, among those humanly possible, having it will lead to one's life going as well as having any other. Since, as Parfit has shown, to be never self-denying is self-defeating in terms of this aim, to be never self-denying is not always a rational disposition. If a selectively trustworthy person may expect to do better than someone who is never self-denying, and if there is no alternative better still, then selective trustworthiness is a rational disposition, and being selectively trustworthy is a necessary condition of being rational. But there need be

no one disposition that, independently of an agent's circumstances, is sufficient to ensure that his life will go as well as possible, and thus I do not suppose that there need be a single supremely rational disposition.

The supremely rational desire, on Parfit's account of S, 'is that one's life go as well as possible for oneself' (p. 8). He says little about how this is to be interpreted. It would, I think, be clearly mistaken to suppose that a person is rational only in so far as she is directly motivated by the supremely rational desire. However, if we understand the supremely rational desire as that which in effect governs or regulates one's other desires, ensuring that a person's particular desires, at least in so far as they are motivationally effective, are compatible with her life going as well as possible, then I need not object to Parfit's account. If, as I have suggested, selective trustworthiness is a rational disposition, then on particular occasions a person should be moved by the desire to keep her promise, even though she may realize that she would do better to break it. But such a desire is fully compatible with her desiring that her life go as well as possible, in so far as she recognizes that did she not desire to keep her promises, she would expect to lose out overall; since she would be unable to make convincing promises in situations in which she could profit by doing so, or to receive benefits that depended on her fellows believing her to be trustworthy or reliable. Acting on those desires that make one trustworthy may be in itself a cost, but having those desires that make one trustworthy, at least in some relationships and with some persons, is a much greater benefit.

Let us turn to the claims of theory S about what it is rational to do. In taking the rational action to be what maximizes the agent's expected benefit (or utility, in the standard parlance), Parfit follows what in my view is the orthodox position advocated by the theory of rational choice. But, as we have seen, at least some agents – those who are capable of affecting their dispositions and who would do best to be selectively trustworthy, or at least not to be never self-denying – *cannot* always do what on this account is rational. On Parfit's view, not only are some persons cursed by rationality; others are condemned to irrationality.

Parfit's account connects rational action with rational motivation. A person who has the supremely rational disposition never to be self-denying and the supremely rational desire that his life go as well as possible will be moved to perform those actions that will bring him the greatest expected benefit. I propose to retain the connection between rational action and rational motivation, but of course in terms of what I have claimed to be rational dispositions, which are those that, given the agent's circumstances, will lead to his life going as well as possible. Thus I interpret theory S as claiming that what it would be rational for one to do is whatever one would be rationally motivated to do. In so far as

selective trustworthiness is a rational disposition, it is rational to keep at least some of one's promises, even though doing so may be self-denying and not maximize one's expected benefit.

How does rational action relate to the substantive aim given by theory S? Parfit's account (p. 8) makes this relation simple and direct. One's aim is that one's life go as well as possible. Therefore, in any situation in which one has a choice among actions, an action is rational if and only if it may reasonably be expected to lead to one's life going as well as possible. But this simple relation is sacrificed by my alternative account. A particular action may be rational even though the agent does not expect it to lead to her life going as well as possible.

There is of course an indirect link between the rationality of an action and the agent's life going as well as possible. This link is provided by the connection between rational actions and rational dispositions. I have claimed that according to theory S, an action is rational if and only if it would be motivated by a rational disposition. This implies that if an action is rational, then it must be a member of a *set* of actions that are collectively performable, could be motivated by a particular disposition or set of compatible dispositions, and that would lead to the agent's life going as well as would the actions belonging to any set that would be motivated by some alternative possible disposition or set of compatible dispositions. Let us call such a set of actions an optimal dispositionally coherent set.

Consider once again the desert breakdown case. Suppose that I am disposed to prudent trustworthiness; I am prudent in making promises when I can expect to benefit from sincere promising, and I keep the promises I make. I promise to reward you if you drive me out of the desert; you drive me out, and I reward you. Rewarding you does not lead to my life going as well as would not rewarding you. However, the actions belonging to the set motivated by prudent trustworthiness – promising to reward you and rewarding you – lead to my life going at least as well as the actions belonging to any set that would be motivated by any alternative possible disposition, such as being never self-denying. There is, of course, a set of actions whose members would lead to my life going better than the set motivated by prudent trustworthiness – the set whose members are promising (sincerely) to reward you and not rewarding you – but no single coherent disposition can motivate both members of this set.

Of course an action may belong to an optimal dispositionally coherent set and yet not be rational. Not rewarding you belongs to such a set – namely, the set of which it is the sole member. Being never self-denying provides adequate motivational basis for not rewarding you; thus dispositional coherence is readily satisfied. And optimality is evident: if

I do not reward you, my life will go better than if I perform any alternative action. But not rewarding you is not rational, since it is not motivated by a rational disposition. Furthermore, an action may be rational and yet belong to a sub-optimal dispositionally coherent set. Rewarding you belongs to such a set – namely, the set of which it is the sole member, since the alternative set of which not rewarding you is the sole member is an optimal dispositionally coherent set. But if an action is rational and belongs to a sub-optimal dispositionally coherent set, then this must be a subset of an optimal set given coherence by the same disposition. And if an action is not rational and belongs to an optimal dispositionally coherent set, then this must be a subset of a sub-optimal set given coherence by the same disposition. For a rational action is one motivated by a rational disposition, and a disposition is rational if and only if all of the actions it motivates collectively lead to the agent's life going at least as well as it would were she motivated by any other possible disposition. Thus an action is rational if and only if it belongs to a maximal set of actions that is given coherence by some disposition, and is optimal among dispositionally coherent sets.

At the end of section II I said that a person who was unable to affect her disposition never to be self-denying might be rational in having (according to Parfit's version of S) the supremely rational disposition, but that she would be cursed by her rationality. On the reformulation of S that I have been sketching, such a person is not fully rational, since *never* to be self-denying is not a rational disposition. To be fully rational, a person must be able to have those dispositions that lead to her life going as well as possible. And the lifetime optimal dispositionally coherent set of actions that are motivated by these dispositions may, and normally will, include some that do not in themselves lead to her life going as well as possible. Being disposed to perform such actions leads to her life going better than if she were disposed to perform only those actions each of which at the time of performance would lead to her life going as well as possible. The desert breakdown case illustrates this in terms of the effects of the agent's dispositions on what she is able to do; if she is, and knows herself to be, never self-denying, then advantageous acts of promising prove unavailable to her. But it may be helpful to conclude this part of my argument with a rather different example.

We are farmers; my crops are ready for harvesting now, and yours will be ready next week. If we harvest together, each of us will do better than if we harvest alone. You are therefore willing to help me with my harvesting now if you can count on me to reciprocate next week. But for whatever reason (perhaps I don't much care for you, and am selling my farm and moving away after the crops are in) we both know that next week I shall do better not to help you. If I am never self-denying, then

next week I shall not help you; on the other hand, if I am disposed to reciprocity, then I shall. If I expect others to be fairly good judges of my dispositions, then I do better to be disposed to mutually beneficial reciprocity than to be never self-denying. The set of actions I am motivated to perform if I am disposed to mutually beneficial reciprocity makes my life go better, but not simply because this disposition enables me to perform advantageous acts, such as making sincere promises, that would not be possible for me were I never self-denying. Rather, my disposition gives rise to expectations by other persons that motivate them to act in ways that afford me opportunities that, even given my self-denying behaviour, make me better off than the opportunities I should have were I never self-denying.

This completes my sketchy reformulation of the Self-interest Theory of rationality. As I noted earlier, I do not want to endorse S as the correct theory of rationality, but only to offer an alternative and, I think, better account of what S claims than that given by Parfit. (It may not be the best account.) And in giving such an alternative I am of course intending to suggest how in general one should relate the formal and substantive aims of a theory of rationality, and determine what dispositions, desires and actions are rational according to that theory.

IV

Suppose that there is a very powerful demon whose rewards and punishments outweigh all other considerations in determining how well one's life goes. This demon issues rules to govern behaviour. But he does not reward obedience to his rules or punish disobedience. Rather, he rewards the disposition to obedience, and punishes all other dispositions. He does this by affecting one's opportunities, so that each person will enjoy more favourable opportunities if he or she is disposed to obey the demon's commands than if he or she is disposed in any other way. In this world the disposition to obey the demon's commands is rational – indeed, supremely rational. Since the demon does not reward actual obedience or punish disobedience, there may well be particular situations in which a person makes her life go less well by obeying than by disobeying. But on my interpretation of S this does not affect the rationality of the disposition to obedience, or of the particular obedient actions that it motivates.

Some persons think that there actually is a demon whose rewards and punishments are all-important in determining their fate; they regard this demon as god. Some of these persons may think that god directly rewards obedience and punishes disobedience, if not in this world then in

the next one. They are rational if they are disposed to be never self-denying, for they are then led straightforwardly to obedience by considering how their lives, including their lives after earthly death, will go if they obey or if they disobey. Others, however, may think that god rewards those who are sincerely disposed to obedience without thought of the benefits and costs thereof. According to theory S as I have interpreted it, they are rational not if they are disposed to be never self-denying, but only if they are disposed to be obedient.

But other still may have a different belief about god's concerns. They may think that god rewards, not the disposition to obedience, but the belief that obedience is the supremely rational aim. They may think that those who consider that to be rational is to have one's life go as well as possible will experience opportunities inferior to those who consider that to be rational is to conform to the will of god. If they are right, what does theory S claim? Notwithstanding divine rewards, theory S claims that having one's life go as well as possible is the supremely rational aim. But it also claims that, in so far as it is within one's power, it is irrational to believe this, and rational to believe instead that obedience to god's will is the supremely rational aim. It claims that being disposed to obey god is rational. A disposition is rational if it makes one's life go as well as possible; the disposition that makes one's life go as well as possible is the disposition that would be most conducive to attaining the supremely rational aim, were that aim to be obedience to god's will. But it claims that one should believe that being disposed to obey god is rational simply because it is most conducive to the aim of being obedient, and not because it will make one's life go as well as possible. Indeed, if one believes that being disposed to obey god is rational because it will make one's life go as well as possible, then one's life will not go as well as possible. In effect, the belief that being disposed to obey god is rational is true only if it is held on false grounds. And theory S claims that an action is rational in so far as it conforms to god's will. But again, the real reason that this is true is not the reason why one should believe it. The real reason relates the action to a disposition that makes one's life go as well as possible, but one should believe it because the action is related to a disposition that is most conducive to obedience to god's will.

If there is a sufficiently powerful demon who rewards the belief that obedience to his will is the supremely rational aim, then it is rational to hold false beliefs about rationality. To hold such beliefs, one may need to be irrational; if so, then it is rational to be in this way irrational. Such rational irrationality is not ruled out by my formulation of theory S. And it could not be ruled out by any formulation that would leave it possible for a person's fate to depend on her beliefs about rationality. But its

scope is significantly reduced. Rationality is a curse only if irrationality is directly rewarded. This is, I think, a significant improvement on Parfit's account.

V

Parfit supposes that it can be rational to cause oneself to act irrationally, and to do this in circumstances in which irrationality itself is not the direct object of reward, or rationality the direct object of punishment. To support this view he considers an example adapted from Schelling (pp. 12–13), which I shall refer to simply as 'Schelling's answer'. An armed robber orders me to open the safe where I keep my gold, threatening to kill my children, one by one, so long as I refuse. I fear that even if I comply, he will kill us anyway, to prevent our identifying him later to the police. Fortunately I have a drug at hand that 'causes one to be, for a brief period, very irrational' (ibid.). I take the drug; the robber now 'can do nothing that will induce me to open the safe. Threats and torture cannot force concessions from someone who is so irrational' (p. 13). And since I am irrational, I am less likely to be able to identify him later. Thus 'making myself irrational is the best way to reduce the great risk that this man will kill us all' (ibid.).

Parfit describes my behaviour under the influence of the drug thus: 'Reeling about the room, I say to the man: "Go ahead. I love my children. So please kill them"' (ibid.). Observing someone behave in this way, we might well consider him irrational. But should we? Suppose you fully understand the situation. If someone said to you, 'Look at Gauthier! He's totally irrational,' you might well respond, 'Not at all. Crazy as he seems, and crazy as his behaviour would be in other circumstances, actually what he's doing is perfectly rational. He's best off acting in a way that bears no predictable relation to what the armed robber does or says. And that's exactly what he's doing.'

We might hesitate, though, to allow that my behaviour is rational. But I think this is because what I do is not under my control, and more especially not under my *reasoned* control. It is true that I do best to act in a way that escapes my reasoned control, but does this make my uncontrolled actions rational? Indeed, we might want to deny that my behaviour under the influence of the drug should be treated as a set of actions at all. But of course we could not then say that I am acting irrationally. The only action that we could judge rational or irrational would be my taking the drug, and that is a perfectly rational action according to theory S, whether we interpret it as Parfit does or as I propose.

But whatever we might intuitively be inclined to say, we must recognize that 'rationality' is a technical term in both Parfit's enquiry and my critique. In Schelling's answer the disposition that will make one's life go as well as possible is to act in a quite random and uncontrolled manner. On my account of theory S, it is therefore the rational disposition in such situations, and the actions to which it gives rise are rational actions. In most situations it would of course not make one's life go best to be disposed to act in a random and uncontrolled way. It is therefore not surprising that we should pre-theoretically characterize all random and uncontrolled behaviour as irrational. Someone who accepts theory S interpreted as I propose will think that in the unusual circumstances of Schelling's answer, our pre-theoretical characterization is mistaken. I find this unsurprising.

Parfit rejects the claim that he labels G1:

> If 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, then (c) it cannot be irrational for this person to act upon this motive. (p. 13)

He thinks that Schelling's answer shows this claim to be false. I think it shows that our ordinary ideas about rationality and irrationality are sometimes mistaken.

Parfit also rejects the claim that he labels G2:

> If 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. (p. 23)

He supports this with an example that raises difficult issues for any theory of rationality, 'How I End My Slavery' (p. 22). He considers someone who rationally becomes a threat-ignorer – that is, someone who rationally comes to think it rational to ignore all threats. Faced with a threat-enforcer, who thinks it rational to execute all threats, and who has made an apocalyptic threat, he ignores it with the predictable apocalyptic result. He has rationally come to believe that it is rational for him to ignore the threat, but Parfit claims that his belief is false.

Given Parfit's interpretation of theory S, a person who comes to believe it rational to be a threat-ignorer no longer fully accepts the theory. But on my interpretation this is not so. If a person reasonably believes that his life will go best if he is a threat-ignorer, then according to theory S, threat-ignoring is for him a rational disposition, and ignoring a threat a rational action. It is compatible with theory S to believe it

rational to ignore threats, and this belief may be true (although a person might have reason to believe it even if it were not). Should we conclude that on my interpretation of S, G2 is true? Specifically, should we agree that if it is rational to dispose oneself to be a threat-ignorer, and so to believe that it is rational to ignore all threats, then it is rational to ignore an apocalyptic threat despite what one expects to be the consequences? And should we welcome this conclusion?

Suppose that I may reasonably expect my life to go better if I am a threat-ignorer. I ignore various threats; on some occasions I find this costly, but this need give me no reason to reconsider or regret my disposition. Although I may be unaware of the specific situations in which being a known threat-ignorer has saved me from being the victim of a threat, I may reasonably believe that I have benefited more from such situations than I have lost from ignoring threats that have been subsequently carried out. But now suppose that you, a known threat-enforcer, issue an apocalyptic threat. You will blow us all up – you, me and our respective families – unless I give you the last piece in my box of chocolate fudge. I don't much want the fudge, but I am a threat-ignorer; I refuse you the fudge, and, as I expect, you blow us all up. This may be irrational on your part, but rational or irrational, given that I could and did expect it, do I not act irrationally in denying you the fudge? For now my disposition to threat-ignoring does not make my life go better – and not just in terms of my future expectations, but on balance. Although in the past I reasonably expected my life to go better overall if I were a threat-ignorer, I realize that if I now ignore your threat, that expectation will prove false, and my life will have gone worse overall. One might suppose that I should have qualified my disposition to be a threat-ignorer, so that I should not have extended it to apocalyptic threats. But I may reasonably have believed that any qualification would reduce its *ex ante* value, so that unqualified threat-ignoring offered me the best life prospects.

I have no easy resolution of the problem implicit in this example. There are self-denying dispositions that may reasonably be expected to make one's life go better overall. If one holds any of these dispositions, one is committed to perform particular actions that will not make one's life go as well as possible. But there are several different forms of commitment possible. The weakest is to be committed to particular costly actions only so long as one reasonably expects adherence to the disposition to be prospectively maximally beneficial. In defending the rationality of dispositions requiring this level of commitment, and of the actions they motivate, I accept the claim G1. Stronger is to be committed to particular costly actions so long as one reasonably expects

past and prospective adherence to the disposition to be maximally beneficial in comparison to what one would have expected from past and prospective adherence to any other disposition. I also defend the rationality of dispositions requiring this level of commitment, and of the actions that they motivate. To this extent I accept the claim G2. Strongest is to be committed to particular costly actions even if one recognizes that past and prospective adherence to the disposition will not be maximally beneficial in comparison with what one would have expected from past and prospective adherence to some other disposition. This third level of commitment requires that one adhere to a disposition in the face of its known failure to make one's life go better. Can this be rational?

With this question my enquiry into rationality moves into an area not charted by Parfit in *Reasons and Persons*. Such an area must be explored elsewhere. I conclude by repeating what I have claimed. We should treat the aim given us by a theory of rationality as explicating what it is to be rational, and then consider the dispositions, desires and acts that it recommends in the pursuit of the rational aim as themselves rational. This is not Parfit's view. On his account dispositions, desires and acts are rational if they are directed *at* the aim, not if they are directed *by* it. For him the rational person is disposed to pursue the aim; for me the rational person is disposed in whatever way will best lead her to achieve the aim. Parfit supposes that theory S tells me that my reasons for acting are considerations about what would be best for myself. I suppose that S tells me that my reasons for acting are considerations determined by my being disposed in the ways that are best for me. More generally, Parfit's view implies that any theory of rationality tells me that my reasons for acting are considerations about what would be conducive to the aim it gives me. And I hold that any theory of rationality tells me that my reasons for acting are considerations determined by my being disposed in the ways that are most conducive to the aim it gives me. Were Parfit right, then rationality would all too frequently be a curse from which, rationally, I should seek to free myself, adopting irrational dispositions that would motivate me to perform irrational acts so that I might better achieve the rational aim. But perhaps it is only his view of rationality that proves a curse, from which I have been seeking to free myself.

Note

1 Parfit does not accept S. Neither do I. But I believe that his account of S shows how he supposes the various parts of a theory of rationality – the specification of the formal aim and the substantive aim, the characterization of rational dispositions, desires and actions – are related. My concern in this

essay is with these relations, and I use theory S to illustrate my differences with Parfit on these matters. I should note my gratitude to Parfit for reading an earlier draft of this paper and suggesting various clarifications and emendations, many – but as he will recognize, not all – of which I have accepted.