From Parfit, Reasons and Persons

for them, which is a bad effect in S's terms. Is this an objection to S? It will be easier to answer this question after I have discussed other theories. My answer is in Section 18.

5. COULD IT BE RATIONAL TO CAUSE ONESELF TO ACT IRRATIONALLY?

I turn now to a new question. A theory may be unacceptable even though it does not fail in its own terms. It is true of many people that it would be worse for them if they were never self-denying. Does this give us independent grounds to reject S?

According to S, it would be rational for each of these people to cause himself to have, or to keep, one of the best possible sets of motives, in self-interested terms. Which these sets are is, in part, a factual question. And the details of the answer would be different for different people in different circumstances. But we know the following, about each of these people. Since it would be worse for him if he was never self-denying, it would be better for him if he was sometimes self-denying. It would be better for him if he was sometimes disposed to do what he believes will be worse for him. S claims that acting in this way is irrational. If such a person believes S, it tells him to cause himself to be disposed to act in a way that S claims to be irrational. Is this a damaging implication? Does it give us any reason to reject S?

Consider

> *Schelling's Answer to Armed Robbery*. A man breaks into my house. He hears me calling the police. But, since the nearest town is far away, the police cannot arrive in less then fifteen minutes. The man orders me to open the safe in which I hoard my gold. He threatens that, unless he gets the gold in the next five minutes, he will start shooting my children, one by one.
>
> What is it rational for me to do? I need the answer fast. I realize that it would not be rational to give this man the gold. The man knows that, if he simply takes the gold, either I or my children could tell the police the make and number of the car in which he drives away. So there is a great risk that, if he gets the gold, he will kill me and my children before he drives away.
>
> Since it would be irrational to give this man the gold, should I ignore his threat? This would also be irrational. There is a great risk that he will kill one of my children, to make me believe his threat that, unless he gets the gold, he will kill my other children.
>
> What should I do? It is very likely that, whether or not I give this man the gold, he will kill us all. I am in a desperate position. Fortunately, I remember reading Schelling's *The Strategy of Conflict*. [3] I also have a special drug, conveniently at hand. This drug causes one to be, for a brief period, very irrational. Before the man can stop me, I reach for the bottle and drink. Within a few seconds, it becomes apparent that I am crazy. Reeling about the room, I say to the man: 'Go ahead. I love my children. So please kill them.' The man tries to get the gold by torturing me. I cry out: 'This is agony. So please go on.'
>
> Given the state that I am in, the man is now powerless. He can do nothing that will induce me to open the safe. Threats and torture cannot force concessions from someone who is so irrational. The man can only flee, hoping to escape the police. And, since I am in this state, he is less likely to believe that I would record the number of his car. He therefore has less reason to kill me.
>
> While I am in this state, I shall act in irrational ways. There is a risk that, before the police arrive, I may harm myself or my children. But, since I have no gun, this risk is small. And making myself irrational is the best way to reduce the great risk that this man will kill us all.

On any plausible theory about rationality, it would be rational for me, in this case, to cause myself to become for a period irrational.[4a] This answers the question that I asked above. S might tell us to cause ourselves to be disposed to act in ways that S claims to be irrational. This is no objection to S. As the case just given shows, an acceptable theory about rationality *can* tell us to cause ourselves to do what, in its own terms, is irrational.

Consider next a general claim that is sometimes made:

> (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.

In the case just described, while this man is still in my house, it would be irrational for me to cause myself to cease to be irrational. During this period, I have a set of motives of which both (a) and (b) are true. But (c) is false. During this period, my acts are irrational. We should therefore reject (G1). We can claim instead that, since it was rational for me to cause myself to be like this, this is a case of *rational* irrationality.

6. HOW S IMPLIES THAT WE CANNOT AVOID ACTING IRRATIONALLY

Remember Kate, who accepts the Hedonistic Theory about self-interest. We may accept some other theory. But on these other theories there could be cases that, in the relevant respects, are like Kate's. And the claims that follow could be restated to cover these cases.

It is best for Kate that her strongest desire is that her books be as good as possible. But, because this is true, she often works very hard, making herself, for a period, exhausted and depressed. Because Kate is a Hedonist, she believes that, when she acts in this way, she is doing what is worse for her. Because she also accepts S, Kate believes that, in these cases, she is acting irrationally. Moreover, these irrational acts are quite voluntary. She acts as she does because, though she cares about her own interests, this is not her strongest desire. She has an even stronger desire that her books be as good as possible. It would be worse for her if this desire became weaker. She is acting on a set of motives that, according to S, it would be irrational for her to cause herself to lose.

It might be claimed that, because Kate is acting on such motives, she cannot be acting irrationally. But this claim assumes (G1), the claim that was shown to be false by the case I called Schelling's Answer to Armed Robbery.

If we share Kate's belief that she is acting irrationally, in a quite voluntary way, we might claim that *she* is irrational. But Kate can deny this. Since she believes S, she can claim: 'When I do what I believe will be worse for me, my *act* is irrational. But, because I am acting on a set of motives that it would be irrational for me to cause myself to lose, *I* am *not* irrational. More precisely, I am *rationally irrational*.'

She can add: 'In acting on my desire to make my books better, I am doing what will be worse for me. This is a bad effect, in self-interested terms. But it is part of a set of effects that is one of the best possible sets. Though I sometimes suffer, because this is my strongest desire, I also benefit. And the benefits are greater than the losses. That I sometimes act irrationally, doing what I know will be worse for me, is the price I have to pay if I am to get these greater benefits. This is a price worth paying.'

It may be objected: 'You do not *have* to pay this price. You *could* work less hard. You could do what would be better for you. You are not compelled to do what you believe to be irrational.'

She could answer: 'This is true. I *could* work less hard. But I only *would* do this if my desire to make my books better was much weaker. And this would be, on the whole, worse for me. It would make my work boring. How could I bring it about that I shall not in the future freely choose, in such cases, to do what I believe to be irrational? I could bring this about only by changing my desires in a way that would be worse for me. This is the sense in which I cannot have the greater benefits without paying the lesser price. I cannot have the desires that are best for me without sometimes freely choosing to act in ways that will be worse for me. This is why, when I act irrationally in these ways, I need not regard *myself* as irrational.'

This reply assumes one view about voluntary acts: *Psychological Determinism*. On this view, our acts are always caused by our desires, beliefs, and other dispositions. Given our actual desires and dispositions, it is not causally possible that we act differently. It may be objected: 'If it is not causally possible that Kate act differently, she should not believe that, to act rationally, she *ought* to act differently. We only *ought* to do what we *can* do.'

A similar objection will arise later when I discuss what we ought morally to do. It will save words if Kate answers both objections. She can say: 'In the doctrine that *ought* implies *can*, the sense of 'can' is compatible with Psychological Determinism. When my act is irrational or wrong, I ought to have acted in some other way. On the doctrine, I ought to have acted in this other way only if I could have done so. If I could *not* have acted in this other way, it cannot be claimed that this is what I ought to have done. The claim (1) that I could not have acted in this other way is not the claim (2) that acting in this way would have been impossible, given my actual desires and dispositions. The claim is rather (3) that acting in this way would have been impossible, even if my desires and dispositions had been different. Acting in this way would have been impossible, whatever my desires and dispositions might have been. If claim (1) was claim (2), Determinists would have to conclude that it is not possible for anyone ever to act wrongly or irrationally. They can justifiably reject this conclusion. They can insist that claim (1) is claim (3).'

Kate could add: 'I am not claiming that *Free Will* is compatible with Determinism. The sense of 'can' required for Free Will may be different from the sense of 'can' in the doctrine that ought implies can. These senses are held to be different by most of those Determinists who believe that Free Will is *not* compatible with Determinism. This is why, though these Determinists do not believe that anyone deserves punishment, they continue to believe that it is possible to act wrongly or irrationally.'

Kate may be wrong to assume Psychological Determinism. I claimed earlier that our beliefs about rationality may affect our acts, because we may want to act rationally. It may objected:

> This misdescribes how these beliefs affect our acts. We do not *explain* why someone has acted rationally by citing his desire to do so. Whenever someone acts rationally, it may be trivially true that he wanted to do so. But he acted as he did because he had a belief, not a belief *and* a desire. He acted as he did simply because he believed that he had a reason to do so. And it is often causally possible for him to act rationally whatever his desires and dispositions are.[4]

Note that this objector cannot claim that it is *always* possible for someone to act rationally, whatever his desires and dispositions are. Even if he denies Determinism, this objector cannot claim that there is *no* connection between our acts and our dispositions.

This objector must also admit that our desires and dispositions may make it *harder* for us to do what we believe to be rational. Suppose that I am suffering from intense thirst, and am given a glass of iced water. And suppose I believe that I have a reason to drink this water slowly, since this would increase my enjoyment. I also have a reason not to spill this water. It

is much easier to act upon this second reason than it is, given my intense thirst, to drink this water slowly.

If the objector's claims are true, Kate's reply must be revised. She might say: 'It would be worse for me if my strongest desire was to avoid doing what I believe to be irrational. It is better for me that my strongest desire is that my books be as good as possible. Since this is my strongest desire, I sometimes do what I believe to be irrational. I act in this way because my desire to make my books better is much stronger than my desire not to act irrationally. You claim that I could often avoid acting in this way. By an act of will, I could often avoid doing what I most want to do. If I could avoid acting in this way, I cannot claim that I am in no sense irrational. But, given the strength of my desire to make my books better, it would be *very hard* for me to avoid acting in this way. And it would be irrational for me to change my desires so that it would be easier for me to avoid acting in this way. Given these facts, I am irrational only in a very weak sense.'

Kate might add: 'It is not possible *both* that I have one of the best possible sets of motives, in self-interested terms, *and* that I never do what I believe to be irrational. This is not possible in the relevant sense: it is not possible *whatever* my desires and dispositions are. If I was never self-denying, my ordinary acts would never be irrational. But I would have acted irrationally in causing myself to become, or allowing myself to remain, never self-denying. If instead I cause myself to have one of the best possible sets of motives, I shall sometimes do what I believe to be irrational. If I do not have the *disposition* of someone who is never self-denying, it is not possible that I *always act* like someone with this disposition. Since this is not possible, and it would be irrational for me to cause myself to be never self-denying, I cannot be criticised for sometimes doing what I believe to be irrational.'

It may now be said that, as described by Kate, S lacks one of the essential features of any theory. It may be objected: 'No theory can demand what is impossible. Since Kate cannot always avoid doing what S claims to be irrational, she cannot always do what S claims that she ought to do. We should therefore reject S. As before, *ought* implies *can*.'

Even if we deny Determinism, this objection still applies. As I have claimed, we must admit that, since Kate does not have the disposition of someone who is never self-denying, she cannot *always* act like such a person.

Is it a good objection to S that Kate cannot always avoid doing what S claims to be irrational? Remember Schelling's Answer to Armed Robbery. In this case, on any plausible theory about rationality, it would be irrational for me not to make myself very irrational. But, if I do make myself very irrational, I cannot avoid acting irrationally. On both alternatives, at least one of my acts would be irrational. It is therefore true that, in this case, I cannot avoid acting irrationally. Since there can be such cases, an acceptable theory can imply that we cannot avoid acting irrationally. It is no objection to S that it has this implication.

We may believe that these claims do not fully answer this objection. A similar objection will be raised later against certain moral theories. To save words, I discuss these objections together, in Section 15.

I shall now summarize my other conclusions. In the case of many and perhaps most people, the Self-interest Theory is indirectly self-defeating. It is true, of each of these people, that it would be worse for him if he was never self-denying—disposed never to do what he believes would be worse for him. It would be better for him if he had some other set of motives. I have claimed that such cases do not provide an objection to S. Since S does not tell these people to be never self-denying, and tells them, if they can, not to be, S is not failing in its own terms. Nor do these cases provide an independent objection to S.

Though they do not refute S, for those who accept S these cases are of great importance. In these cases S must cover, not just ordinary acts, but also the acts that bring about changes in our motives. According to S, it would be rational to cause ourselves to have, or to keep, one of the best possible sets of motives, in self-interested terms. If we believe that we could act in either of these ways, it would be irrational not to do so. In the case of most people, any of the best possible sets would cause these people sometimes to do, in a quite voluntary way, what they know will be worse for them. If these people believe S, they will believe that these acts are irrational. But they need not believe *themselves* to be irrational. This is because, according to S, it would be irrational for them to change their motives so that they would cease to act irrationally in this way. They will in part regret the *consequences* of these irrational acts. But the *irrationality* of these acts they can regard with complacency. This is *rational* irrationality.

It may be objected, to these claims, that they falsely assume Psychological Determinism. It may sometimes be possible for these people to do what they believe to be rational, whatever their desires and dispositions are. If this objection is correct, these claims need to be revised. When these people do what they believe to be irrational, they cannot claim that they are in no sense irrational. But they can claim that, given their actual motives, it would be very hard for them to avoid acting in this way. And it would be irrational for them, on their theory, to change their motives so that it would be easier to avoid acting in this way. They can therefore claim that they are irrational only in a very weak sense. Having explained once how these claims could be revised, I shall not mention this objection whenever, in what follows, it would be relevant. It would be easy to make the needed revisions to any similar claims.

7. AN ARGUMENT FOR REJECTING S WHEN IT CONFLICTS WITH MORALITY

It has been argued that the Self-interest Theory might tell us to believe, not itself, but some other theory. This is clearly possible. According to S, it

would be rational for each of us to cause himself to believe some other theory, if this would be better for him.

I have already mentioned one way in which this might be true. It might not be possible for us to do what we believe to be irrational. S would then tell us, in the cases I have been discussing, to try to believe a different theory. There are also other ways in which this might be true. Let us return, for an example, to the keeping of our promises.

One kind of mutual agreement has great practical importance. In these agreements, each person in some group makes a conditional promise. Each person promises to act in a certain way, provided that all the others promise to act in certain ways. It can be true both (1) that it will be better for each of these people if all rather than none of them keep their promises, and (2) that, whatever the others do, it will be worse for each person if he himself keeps his promise. What each person loses if he keeps his promise is less than what he gains if all the others keep their promises. This is how (1) and (2) are both true. Such agreements are *mutually advantageous, though requiring self-denial.*

If I am known to be never self-denying, I shall be excluded from such agreements. Others will know that I cannot be trusted to keep my promise. It has been claimed that, since this is true, it would be better for me if I ceased to be never self-denying and became trustworthy.[5]

This claim overlooks one possibility. It may be best for me if I *appear* to be trustworthy but remain really never self-denying. Since I appear to be trustworthy, others will admit me to these mutually advantageous agreements. Because I am really never self-denying, I shall get the benefits of breaking my promises whenever this would be better for me. Since it is better for me to appear trustworthy, it will often be better for me to keep my promise so as to preserve this appearance. But there will be some promises that I can break secretly. And my gain from breaking some promises may outweigh my loss in ceasing to appear trustworthy.

Suppose, however, that I am transparent, unable to lie convincingly. This is true of many people. And it might become more widely true if we develop cheap and accurate lie-detector tests. Let us assume that this has happened, so that we are all transparent—unable to deceive others. Since we are to some degree transparent, my conclusions may apply to our actual situation. But it will simplify the argument to assume that all direct deception has become impossible. It is worth seeing what such an argument might show. We should therefore help the argument, by granting this assumption.

If we were all transparent, it would be better for each of us if he became trustworthy: reliably disposed to keep his promises, even when he believes that doing so would be worse for him. It would therefore be rational, according to S, for each of us to make himself trustworthy.

Assume next that, to become trustworthy, we would have to change our beliefs about rationality. We would have to make ourselves believe that it is rational for each of us to keep his promises, even when he knows that this would be worse for him. I shall later describe two ways in which this assumption might be true.

It is hard to change our beliefs when our reason for doing so is merely that this change will be in our interests. We would have to use some form of self-deception. Suppose, for example, that I learn that I am fatally ill. Since I want to believe that I am healthy, I pay a hypnotist to give me this belief. I could not keep this belief if I remembered how I had acquired it. If I remembered this, I would know that the belief was false. The same would be true of our beliefs about rationality. If we pay hypnotists to change these beliefs, because this will be better for us, the hypnotists must make us forget why we have our new beliefs.

On the assumptions made above, S would tell us to change our beliefs. S would tell us to believe, not itself, but a revised form of S. On this revised theory, it is irrational for each of us to do what he believes will be worse for himself, *except when he is keeping a promise.*

If S told us to believe this revised theory, would this be an objection to S? Would it show that it *is* rational to keep such promises? We must focus clearly on this question. We may be right to believe that it is rational to keep our promises, even when we know that this will be worse for us. I am asking, 'Would this belief be supported if S itself told us to cause ourselves to have this belief?'

Some people answer Yes. They argue that, if S tells us to make ourselves have this belief, this shows that this belief is justified. And they apply this argument to many other kinds of act which, like keeping promises, they believe to be morally required. If this argument succeeded, it would have great importance. It would show that, in many kinds of case, it is rational to act morally, even when we believe that this will be worse for us. Moral reasons would be shown to be stronger than the reasons provided by self-interest. Many writers have tried, unsuccessfully, to justify this conclusion. If this conclusion could be justified in the way just mentioned, this would solve what Sidgwick called 'the profoundest problem of Ethics'.[6]

8. WHY THIS ARGUMENT FAILS

There is a simple objection to this argument. The argument appeals to the fact that S would tell us to make ourselves believe that it is rational to keep our promises, even when we know that this will be worse for us. Call this belief *B*. B is incompatible with S, since S claims that it is irrational to keep such promises. Either S is the true theory about rationality, or it is not. If S is true, B must be false, since it is incompatible with S. If S is not true, B might be true, but S cannot support B, since a theory that is not true cannot support any conclusion. In brief: if S is true, B must be false, and if S is not true, it cannot support B. B is either false, or not supported. So, even if S tells us to try to believe B, this fact cannot support B.

We may think that a theory about rationality cannot be true, but can at most be the best, or the best justified theory. The objection just given could be restated in these terms. There are two possibilities. If S is the best theory, we should reject B, since it is incompatible with S. If S is not the best theory, we should reject S. B cannot be supported by a theory that we should reject. Neither of these possibilities gives any support to B.[7]

This objection seems to me strong. But I know some people whom it does not convince. I shall therefore give two more objections. These will also support some wider conclusions.

I shall first distinguish threats from warnings. When I say that I shall do *X* unless you do *Y*, call this a *warning* if my doing X would be worse for you but not for me, and a *threat* if my doing X would be worse for both of us. Call me a *threat-fulfiller* if I would always fulfil my threats.

Suppose that, apart from being a threat-fulfiller, someone is never self-denying. Such a person would fulfil his threats even though he knows that this would be worse for him. But he would not *make* threats if he believed that doing so would be worse for him. This is because, apart from being a threat-fulfiller, this person is never self-denying. He never does what he believes will be worse for him, *except when he is fulfilling some threat*. This exception does not cover *making* threats.

Suppose that we are all both transparent and never self-denying. If this was true, it would be better for me if I made myself a threat-fulfiller, and then announced to everyone else this change in my dispositions. Since I am transparent, everyone would believe my threats. And believed threats have many uses. Some of my threats could be defensive, intended to protect me from aggression by others. I might confine myself to defensive threats. But it would be tempting to use my known disposition in other ways. Suppose that the benefits of some co-operation are shared between us. And suppose that, without my co-operation, there would be no further benefits. I might say that, unless I get the largest share, I shall not co-operate. If others know me to be a threat-fulfiller, and they are never self-denying, they will give me the largest share. Failure to do so would be worse for them.

Other threat-fulfillers might act in worse ways. They could reduce us to slavery. They could threaten that, unless we become their slaves, they will bring about our mutual destruction. We would know that these people would fulfil their threats. We would therefore know that we can avoid destruction only by becoming their slaves.

The answer to threat-fulfillers, if we are all transparent, is to become a *threat-ignorer*. Such a person always ignores threats, even when he knows that doing so will be worse for him. A threat-fulfiller would not threaten a transparent threat-ignorer. He would know that, if he did, his threat would be ignored, and he would fulfil this threat, which would be worse for him.

If we were all both transparent and never self-denying, what changes in our dispositions would be better for each of us? I answer this question in Appendix A, since parts of the answer are not relevant to the question I am now discussing. What is relevant is this. If we were all transparent, it would probably be better for each of us if he became a trustworthy threat-ignorer. These two changes would involve certain risks; but these would be heavily outweighed by the probable benefits. What would be the benefits from becoming trustworthy? That we would not be excluded from those mutually advantageous agreements that require self-denial. What would be the benefits from becoming threat-ignorers? That we would avoid becoming the slaves of threat-fulfillers.

We can next assume that we could not become trustworthy threat-ignorers unless we changed our beliefs about rationality. Those who are trustworthy keep their promises even when they know that this will be worse for them. We can assume that we could not become disposed to act in this way unless we believed that it *is* rational to keep such promises. And we can assume that, unless we were known to have this belief, others would not trust us to keep such promises. On these assumptions, S tells us to make ourselves have this belief. Similar remarks apply to becoming threat-ignorers. We can assume that we could not become threat-ignorers unless we believed that it is always rational to ignore threats. And we can assume that, unless we have this belief, others would not be convinced that we are threat-ignorers. On these assumptions, S tells us to make ourselves have this belief. These conclusions can be combined. S tells us to make ourselves believe that it is always irrational to do what we believe will be worse for us, *except when we are keeping promises or ignoring threats*.

Does this fact support these beliefs? According to S, it would be rational for each of us to make himself believe that it is rational to ignore threats, even when he knows that this will be worse for him. Does this show this belief to be correct? Does it show that it *is* rational ignore such threats?

It will help to have an example. Consider

> *My Slavery*. You and I share a desert island. We are both transparent, and never self-denying. You now bring about one change in your dispositions, becoming a threat-fulfiller. And you have a bomb that could blow the island up. By regularly threatening to explode this bomb, you force me to toil on your behalf. The only limit on your power is that you must leave my life worth living. If my life became worse than that, it would cease to be better for me to give in to your threats.

How can I end my slavery? It would be no good killing you, since your bomb will automatically explode unless you regularly dial some secret

number. But suppose that I could make myself transparently a threat-ignorer. Foolishly, you have not threatened that you would ignore this change in my dispositions. So this change would end my slavery.

Would it be rational for me to make this change? There is the risk that you might make some new threat. But since doing so would be clearly worse for you, this risk would be small. And, by taking this small risk, I would almost certainly gain a very great benefit. I would almost certainly end my slavery. Given the wretchedness of my slavery, it would be rational for me, according to S, to cause myself to become a threat-ignorer. And, given our other assumptions, it would be rational for me to cause myself to believe that it is always rational to ignore threats. Though I cannot be wholly certain that this will be better for me, the great and nearly certain benefit would outweigh the small risk. (In the same way, it would never be wholly certain that it would be better for someone if he became trustworthy. Here too, all that could be true is that the probable benefits outweigh the risks.)

Assume that I have now made these changes. I have become transparently a threat-ignorer, and have made myself believe that it is always rational to ignore threats. According to S, it was rational for me to cause myself to have this belief. Does this show this belief to be correct?

Let us continue the story.

How I End My Slavery. We both have bad luck. For a moment, you forget that I have become a threat-ignorer. To gain some trivial end—such as the coconut that I have just picked—you repeat your standard threat. You say, that, unless I give you the coconut, you will blow us both to pieces. I know that, if I refuse, this will certainly be worse for me. I know that you are reliably a threat-fulfiller, who will carry out your threats even when you know that this will be worse for you. But, like you, I do not now believe in the pure Self-interest Theory. I now believe that it is rational to ignore threats, even when I know that this will be worse for me. I act on my belief. As I foresaw, you blow us both up.

Is my act rational? It is not. As before, we might concede that, since I am acting on a belief that it was rational for me to acquire, *I* am not irrational. More precisely, I am *rationally* irrational. But what I am doing is not rational. It is irrational to ignore some threat when I know that, if I do, this will be disastrous for me and better for no one. S told me here that it was rational to make myself believe that it is rational to ignore threats, even when I know that this will be worse for me. But this does not show this belief to be correct. It does not show that, in such a case, it *is* rational to ignore threats.

We can draw a wider conclusion. This case shows that we should reject

(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.

Return now to B, the belief that it is rational to keep our promises even when we know that this will be worse for us. On the assumptions made above, S implies that it is rational for us to make ourselves believe B. Some people claim that this fact supports B, showing that it *is* rational to keep such promises. But this claim seems to assume (G2), which we have just rejected.

There is another objection to what these people claim. Even though S tells us to try to believe B, S implies that B is false. So, if B is true, S must be false. Since these people believe B, they should believe that S is false. Their claim would then assume

(G3) If some false theory about rationality tells us to make ourselves have a particular belief, this shows this belief to be true.

But we should obviously reject (G3). If some false theory told us to make ourselves believe that the Earth was flat, this would not show this to be so.

S told us to try to believe that it is rational to ignore threats, even when we know that this will be worse for us. As my example shows, this does not support this belief. We should therefore make the same claim about keeping promises. There may be *other* grounds for believing that it is rational to keep our promises, even when we know that doing so will be worse for us. But this would not be shown to be rational by the fact that the Self-interest Theory itself told us to make ourselves believe that it was rational. It has been argued that, by appealing to such facts, we can solve an ancient problem: we can show that, when it conflicts with self-interest, morality provides the stronger reasons for acting. This argument fails. The most that it might show is something less. In a world where we are all transparent—unable to deceive each other—it might be rational to deceive ourselves about rationality.[8]

9. HOW S MIGHT BE SELF-EFFACING

If S told us to believe some other theory, this would not support this other theory. But would it be an objection to S? Once again, S would not be failing in its own terms. S is a theory about practical not theoretical rationality. S may tell us to make ourselves have false beliefs. If it would be better for us to have false beliefs, having true beliefs, even about rationality, would not be part of the ultimate aim given to us by S.

The arguments given above might be strengthened and extended. This would be easier if, as I supposed, the technology of lie-detection made us all wholly transparent. If we could never deceive each other, there might be an argument that showed that, according to S, it would be rational for everyone to cause himself not to believe S.

Suppose that this was true. Suppose that S told everyone to cause himself to believe some other theory. S would then be *self-effacing*. If we all believed S, but could also change our beliefs, S would remove itself from the scene. It would become a theory that no one believed. But to be self-effacing is not to be self-defeating. It is not the aim of a theory to be believed. If we personify theories, and pretend that they have aims, the aim of a theory is not to be believed, but to be true, or to be the best theory. That a theory is self-effacing does not show that it is not the best theory.

S would be self-effacing when, if we believed S, this would be worse for us. But S need not tell us to believe itself. When it would be better for us if we believed some other theory, S would tell us to try to believe this theory. If we succeeded in doing what S told us to do, this would again be better for us. Though S would remove itself from the scene, causing no one to believe itself, it would still not be failing in its own terms. It would still be true that, because each of us has followed S—done what S told him to do—each has thereby made the outcome better for himself.

Though S would not be failing in its own terms, it might be claimed that an acceptable theory cannot be self-effacing. I deny this claim. It may seem plausible for what, when examined, is a bad reason. It would be natural to *want* the best theory about rationality not to be self-effacing. If the best theory was self-effacing, telling us to believe some other theory, the truth about rationality would be depressingly convoluted. It is natural to hope that the truth is simpler: that the best theory would tell us to believe itself. But can this be more than a hope? Can we assume that the truth *must* be simpler? We cannot.

10. HOW CONSEQUENTIALISM IS INDIRECTLY SELF-DEFEATING

Most of my claims could, with little change, cover one group of moral theories. These are the different versions of *Consequentialism*, or *C*. C's *central claim* is

(C1) There is one ultimate moral aim: that outcomes be as good as possible.

C applies to everything. Applied to acts, C claims both

(C2) What each of us ought to do is whatever would make the outcome best, and

(C3) If someone does what he believes will make the outcome worse, he is acting wrongly.

I distinguished between what we have most reason to do, and what it would be rational for us to do, given what we believe, or ought to believe. We must now distinguish between what is *objectively* and *subjectively* right or wrong. This distinction has nothing to do with whether moral theories can be objectively true. The distinction is between what some theory implies, given (i) what are or would have been the effects of what some person does or could have done, and (ii) what this person believes, or ought to believe, about these effects.

It may help to mention a similar distinction. The medical treatment that is objectively right is the one that would in fact be best for the patient. The treatment that is subjectively right is the one that, given the medical evidence, it would be most rational for the doctor to prescribe. As this example shows, what it would be best to know is what is objectively right. The central part of a moral theory answers this question. We need an account of subjective rightness for two reasons. We often do not know what the effects of our acts would be. And we ought to be blamed for doing what is subjectively wrong. We ought to be blamed for such acts even if they are objectively right. A doctor should be blamed for doing what was very likely to kill his patient, even if his act in fact saves this patient's life.

In most of what follows, I shall use *right*, *ought*, *good*, and *bad* in the objective sense. But *wrong* will usually mean *subjectively* wrong, or *blameworthy*. Which sense I mean will often be obvious given the context. Thus it is clear that, of the claims given above, (C2) is about what we ought objectively to do, and (C3) is about what is subjectively wrong.

To cover risky cases, C claims

(C4) What we ought subjectively to do is the act whose outcome has the greatest *expected* goodness.

In calculating the expected goodness of an act's outcome, the value of each possible good effect is multiplied by the chance that the act will produce it. The same is done with the disvalue of each possible bad effect. The expected goodness of the outcome is the sum of these values minus these disvalues. Suppose, for example, that if I go West I have a chance of 1 in 4 of saving 100 lives, and a chance of 3 in 4 of saving 20 lives. The expected goodness of my going West, valued in terms of the number of lives saved, is $100 \times 1/4 + 20 \times 3/4$, or $25 + 15$, or 40. Suppose next that, if I go East, I shall certainly save 30 lives. The expected goodness of my going East is 30×1, or 30. According to (C4), I ought to go West, since the expected number of lives saved would be greater.

Consequentialism covers, not just acts and outcomes, but also desires, dispositions, beliefs, emotions, the colour of our eyes, the climate, and everything else. More exactly, C covers anything that could make outcomes better or worse. According to C, the best possible climate is the one that would make outcomes best. I shall again use 'motives' to cover both desires and dispositions. C claims