Bombs and Coconuts, or Rational Irrationality

DEREK PARFIT

In an early article, Gauthier argued that, to act rationally, we must act morally.[1] I tried to refute that argument.[2] Since Gauthier was not convinced, I shall try again.[3]

1

Gauthier assumes that, to be rational, we must maximize our own expected utility. Though he distinguishes between 'utility' and 'benefit', this distinction does not affect his main arguments. We can regard him as appealing to the *Self-interest Theory*.[4]

Many writers have argued that, in self-interested terms, it is always rational to act morally. According to most of these writers, morality and self-interest coincide. But that is not Gauthier's line. Gauthier concedes that acting morally may be, and be known to be, worse for us. He claims that, even in such cases, it is rational to act morally.

If we appeal to the Self-interest Theory, it may seem impossible to defend that claim. How can our acts be rational, in self-interested terms, if we know them to be worse for us? But Gauthier *revises* the Self-interest Theory. On the standard version of this theory, an act is rational if it will maximize our expected benefit – or be *expectably-best* for us.[5] On Gauthier's version, we should aim to benefit ourselves not with our *acts* but only with our *dispositions*. A disposition is rational if having it will be expectably-best for us. An act is rational if it results from such a disposition.

Besides revising the Self-interest Theory, Gauthier restricts the scope of morality. To act morally, Gauthier claims, we must honour our agreements. In the cases with which he is concerned, each of us promises that, at some cost to ourselves, we shall give a greater benefit to others. If we all kept such promises, we would all gain. The cost to each would be outweighed by the benefits received from others.

Though such agreements are mutually advantageous, it would often be better for each if she broke her promise. Either she could break it secretly, or the damage to her reputation would be outweighed by what she gains. We

may think that, in self-interested terms, it is rational to break such promises. But Gauthier argues that, if we do, we are fools.

Gauthier's argument starts with a prediction. If we were straightforwardly self-interested – or, for short, *prudent* – we would intend to break such promises. Other people, knowing this, would exclude us from these advantageous agreements. That would be worse for us. It would be better for us if we were trustworthy, since we would then be admitted to these agreements.

It would be even better for us, as I remarked, if we appeared to be trustworthy but were really prudent. We would still be admitted to these agreements, but we would break our promises whenever we could expect to benefit.[6] Gauthier replied that we are too *translucent* to be capable of such deceit. When we were negotiating such agreements, we would sometimes be unable to conceal our true intentions. He therefore claimed that, on balance, it would be better for us if we were really trustworthy.[7]

Gauthier then appealed to his variant of the Self-interest Theory – which I shall call *Gauthier's view*. On this view, since it is in our interests to be trustworthy, it is rational for us to act upon this disposition. It is rational to keep our promises, even when we know that what we are doing will be worse for us.

Should we accept this argument? I believe not. When applied to trustworthiness, it may seem plausible. But we should reject Gauthier's view. It could be in our interests to have some disposition, and rational to cause ourselves to have it, but be irrational to act upon it.

2

One problem for Gauthier's view is that, at different times, different dispositions can be in our interests. This makes it hard to state Gauthier's view in a way that will suit his purposes.

In his earliest statements of his view, Gauthier assumed

> (A) If we have acquired some disposition because we reasonably believed that, by doing so, we would make our lives go better, it is rational to act upon this disposition.[8]

I challenged (A) as follows.[9] Just as it could be in our interests to be trustworthy, it could be in our interests to be disposed to fulfil our threats, and to ignore threats made by others. As before, it would be best to appear to have these dispositions, while remaining really prudent. But, to test Gauthier's view, we should accept his claim that we are too translucent to be able to deceive others. It might then be better for us if we really had these dispositions. But that would not show that it must be rational to act upon them.[10]

I gave the following example, which I shall here call *Your Fatal Threat*. Suppose that you and I are on a desert island, and we are both transparent. You become a *threat-fulfiller*. By regularly threatening to explode some bomb, you aim to make me your slave. My only way to preserve my freedom is to become a *threat-ignorer*. Since I know that you know that I am translucent, I can reasonably expect that having this disposition would be best for me. I manage to acquire this disposition. But I have bad luck. In a momentary lapse, you threaten that, unless I give you a coconut, you will blow us both to pieces. According to (A), it would be rational for me to ignore your threat. This would be rational even though I know that, if I do, you will kill us both.

Gauthier once accepted this conclusion.[11] But he later revised his view, moving from (A) to

(B) If we have reason to believe that, in acquiring some disposition, we made our lives go better, it is rational to act upon this disposition.

According to (B), for it to be rational to act upon some disposition, it is not enough that we *did* have reason to believe that, by acquiring this disposition, we would make our lives go better. We must *still* have reason to believe that this past belief was true. We need not 'adhere to a disposition in [the] face of its known failure to make one's life go better'.[12]

Gauthier intended (B) to handle my example. When you make your fatal threat, I lose my reason to believe that, in becoming a threat-ignorer, I made my life go better. On Gauthier's revised view, I need not 'adhere' to my disposition.

We can revise the example. Suppose I know that, if I had not become a threat-ignorer, I would have died some time ago.[13] Gauthier's view again implies that I should ignore your threat. Since my disposition once saved my life, my acquiring of this disposition made my life go better. True, it will now kill me. But that is not what counts. According to (B), I should deny you the coconut, and be blown to pieces.[14]

As this example shows, even if some disposition has become disastrous, (B) can still imply that it is rational to act upon it. This would be rational if this disposition brought past benefits that were greater than its future costs. Gauthier claims that we should 'adhere' to such dispositions. We should be true to our 'commitment'.

When applied to promises, such a view has some appeal. If we have gained from trustworthiness, we may think it rational to act upon this disposition, even if it becomes a burden. Talk of *commitment* here makes sense. But, in the case of threat-behaviour, it makes little sense. Why should I remain a threat-ignorer, at the cost of death, merely because this disposition once saved my life?[15]

If my alternative was to be your slave, death would hardly be a cost. But we can add a further detail to the case. Suppose that a rescue party has just landed on the beach. I know that, if I give you the coconut, I shall soon be freed.

To handle this version of the case, Gauthier must again change his view. It may have been rational for me to become a threat-ignorer. But, as Gauthier must agree, it would now be rational for me to try to lose this disposition.[16] If I could lose this disposition, it would be irrational to keep it. Since that is so, Gauthier cannot claim that it must still be rational to act upon it. Now that I could soon be free, it would be irrational for me knowingly to bring about my death.[17]

How should Gauthier revise his view? (B) could be restated so that it covered temporary dispositions. But there is a simpler formulation. Gauthier could turn to

(C) If we have reason to believe that, in having some disposition, we are making our lives go better, it is rational for us to act upon this disposition.

If he appealed to (C), Gauthier would cease to be embarrassed by my example. When I see that my disposition has become disastrous, (C) does not imply that it must still be rational for me to act upon it.[18]

I gave another example, which we can here call *Schelling's Case*. A robber threatens that, unless I unlock my safe, he will start to kill my children. It would be irrational for me to ignore this robber's threat. But, even if I gave in to his threat, there is a risk that he will kill us all, to reduce his chance of being caught. I claimed that, in this case, it would be rational for me to take a drug that would make me very irrational. The robber would then see that it was pointless to threaten me; and, since he could not commit his crime, and I would not be capable of calling the police, he would also be less likely to kill either me or my children.

When Gauthier considered this example, he seemed to accept (C). He agreed that it would be rational for me to make myself, for a brief period, insane; and he claimed that it would be rational for me to act upon this disposition.[19]

If he turned to (C), however, Gauthier would pay a price. In his defence of contractual morality, Gauthier compared only permanent dispositions. He thought it enough to show that, if we are trustworthy, this will on the whole make our lives go better.[20] But, if he appealed to (C), he would need to show more than this. According to (C), for it to be rational to act upon a disposition, it is not enough that it was in our interests to acquire it. We must have reason to believe that, *at the time of acting*, it is in our interests to have it. Gauthier

must therefore show that, if we are trustworthy, this disposition is in our interests when we are *keeping* our agreements.

He does not, I believe, show this. What he shows is, at most, that trustworthiness is in our interests when we are negotiating our agreements. In some cases, when the time comes to keep one agreement, we are negotiating some new agreement. Gauthier's argument may then apply. But in other cases there is no such overlap. There are some promises that we could secretly and swiftly break, to our own advantage. When this is possible, it would be worse for us if we were trustworthy. It would be better for us if we lost that disposition, and became self-interested, even if only for just long enough to break our promise.[21]

To defend his view that it is always rational to act morally, Gauthier must claim that it would be rational to keep such promises. If he appealed to (C), however, he would lose his argument for that claim. (C) implies that it would be rational to break such promises, since we would then be acting on the disposition that we could reasonably believe to be, at the time, best for us.

Gauthier might try a different reply. He might claim that, if we are trustworthy, we would be unable to lose, or to overcome, this disposition. In the sense that is relevant here, this claim may not be true.[22] But suppose that it were true. Suppose that, because I am trustworthy, I would find it impossible to break some promise. Gauthier might appeal to the claim that 'ought' implies 'can'. He might say that, since I cannot break my promise, it cannot be true that it would be rational for me to do so. And he might say that, given the strength of my disposition, it would be rational for me to act upon it.[23]

Is this an adequate reply? Return to the case in which I am disposed to ignore your fatal threat. If I overcome my disposition, and thereby manage to remain alive until I can be rescued, Gauthier must agree that my act is rational. But suppose that my disposition proves too strong. I find that I cannot bring myself to give you the coconut. Could Gauthier claim that, since I cannot overcome my disposition, it cannot be true that it would be rational for me to do so? Could he claim that, since it is causally impossible for me to act differently, it is rational for me to bring about my death?

I believe not. As Gauthier elsewhere claims, what it is rational for us to do does not depend, in this way, on what is causally possible. We could have acted otherwise, in the relevant sense, if nothing stopped us from doing so except our desires or dispositions. If it would have been rational for me to have acted differently, it is irrelevant that, given my desires and dispositions, acting differently would have been causally impossible. Nor could I defend my act by appealing to the strength of my disposition. That may exempt *me* from certain kinds of criticism. But it cannot show that my *act* is rational.[24]

Gauthier admits as much in retreating from claim (A). Suppose that, though

it was rational for me to acquire some disposition, I have learnt that doing so was a terrible mistake. Gauthier no longer claims that it must still be rational to act upon such dispositions. He agrees that, from the fact that I rationally acquired some disposition, and that I cannot overcome it, we cannot infer that it is rational to act upon it.

3

I have described one problem for Gauthier's view. Since it can be in our interests to have temporary dispositions, it is hard to state his view in a way that suits his purposes. Let us now ignore this problem, and turn to the central question. Should we accept Gauthier's view? Should we believe that, if it is in our interests to have some disposition, or rational to cause ourselves to have it, it is rational to act upon it?

In the cases with which we are concerned, though it is in our interests to have some disposition, it is against our interests to act upon it. Only here does Gauthier's view make a difference.

Reconsider Schelling's Case. Because I am temporarily insane, the robber knows that, even if he starts to kill my children, he will not induce me to unlock my safe. He will therefore soon make his getaway. This is greatly to my advantage.[25] But, while I am in my drug-induced state, and before the robber leaves, I act in damaging and self-defeating ways. I beat my children because I love them. I burn my manuscripts because I want to preserve them.

Gauthier objects that my crazy acts are, in fact, better for me. They are what persuades this man that I am immune to his threats. Since these acts are better for me, they are, on any view, rational. So this is not, as I claimed, a case of rational irrationality.[26]

To answer this objection, we can add one feature to the case. We can suppose that, to convince this man that I am crazy, I don't need to act in crazy ways. He sees me take this drug, and he knows that it produces temporary madness. Since the robber already knows that I am in this state, my destructive acts have no good effects.

Though my acts have only bad effects, they result from an advantageous disposition. That is enough, on Gauthier's view, to make these acts rational.[27]

We should note the extremity of this view. Hume at least required that, for our acts to be rational, we must be trying to achieve our aims. On Gauthier's view, we could be trying to frustrate our aims. When I burn my manuscript, or beat my children, I might be doing what I believe to be irrational, and *because* I believe it to be irrational. My acts could be as crazy as we can imagine. They could still, on Gauthier's view, be rational.[28] That is hard to believe.

4

Of Gauthier's arguments for his view, one appeals to the claim that, if we accept his view, this will be better for us. We can first ask whether that is true.

Gauthier assumes that, to be rational, we should maximize our own expected utility. He compares two versions of this view. According to the standard version of the Self-interest Theory, which I called S, we should maximize at the level of our acts. An act is rational if it maximizes the expected benefit to us. According to Gauthier's view, we should maximize only at the level of our dispositions. An act is rational if it results from a maximizing disposition. This view we can now call G.[29]

In the cases with which we are concerned, we cannot always maximize at both levels. If we try to maximize with all our acts, we cannot have maximizing dispositions. Thus, if we break our promises whenever we can expect this to be better for us, we cannot be trustworthy, which will be bad for us.[30]

When we cannot maximize at both levels, it will be better for us if we have maximizing dispositions. The good effects of these dispositions will outweigh the bad effects of our acts.[31]

Gauthier claims that, given this fact, it will be better for us if we accept not S but G.[32] In making this claim, Gauthier assumes that, if we accept S, we would maximize with our acts *rather than* our dispositions.

This assumption may be incorrect. Since it would be better for us if we had maximizing dispositions, S would tell us, if we could, to acquire them. S agrees with G that we should try to *have* these dispositions.[33] What S denies is only that it must be rational to act upon them.

Gauthier may think that, if we accept S, we would always do what S claims to be rational.[34] Or he may think that, in judging any theory about rationality, we should ask what would happen if we always successfully followed it. This may be why he assumes that we would always maximize with our acts. But, if we can change our dispositions, we cannot always do what S claims to be rational. Acquiring these dispositions would itself be a maximizing act. If we maximize with all our other acts, we shall have acted irrationally in failing to acquire these dispositions. If instead we acquire these dispositions, we cannot always maximize with our other acts.[35]

Since we cannot always do what S claims to be rational, we must do the best we can. And S implies that, rather than maximizing with our other acts, we should acquire maximizing dispositions. This is the way of acting that we can expect to be best for us. The disagreement between S and G is not over the question of whether we should *acquire* maximizing dispositions. S claims this as much as G. The disagreement is only about whether, when we act on such dispositions, what we are doing is rational.[36]

Gauthier might now say that, if we accept S, we would be *unable* to acquire these dispositions. We would believe that, in some cases, acting on these dispositions would be irrational. And we might be unable to make ourselves disposed to do what we believe to be irrational. Perhaps, to acquire these dispositions, we must accept Gauthier's view, and believe that it is rational to act upon them.

When he discusses nuclear deterrence, Gauthier does make such a claim.[37] He supposes that it would be in our interests to form an intention to retaliate, if we are attacked. Forming this intention might be what protects us from attack. Gauthier then claims that, if we believed that such retaliation would be irrational, we would be unable to form this intention.[38]

It would be implausible to claim that we could *never* acquire some disposition if we believed that acting upon it would be irrational. Schelling's Case is one exception, and there are many others. But Gauthier would not need so strong a claim. He might say that it would often be impossible to acquire such dispositions. Or he might say that, if we believe that it would be irrational to act in some way, it would be more difficult for us to become disposed to act in this way. We might have to use some indirect method, such as taking drugs, or hypnosis, both of which have disadvantages. Things might be easier if we believed that it would be rational to act in this way. We might then be able simply to decide to do so.[39]

This may only shift the problem. How could we acquire this belief? Suppose that, as Gauthier claims, we could not intend to retaliate unless we believed that retaliation would be rational. If retaliation would be both pointless and suicidal, as Gauthier concedes, how could we persuade ourselves that, as Gauthier also claims, such retaliation would be rational? How could we make ourselves believe Gauthier's view? It is not easy to acquire some belief if our only ground for doing so is that this belief would be in our interests. Here too, we might need some costly indirect method. Let us, however, ignore this problem. It might be impossible for us to acquire some useful disposition unless we can somehow manage to believe that it would be rational to act upon it. It might then be in our interests to acquire this belief.[40]

Suppose that, for these or other reasons, it would be worse for us if we accepted the standard version of the Self-interest Theory. It would be better for us if we accepted Gauthier's view. That would not yet show that Gauthier's view is true, or is the best view. To reach that conclusion, Gauthier needs another premise.

In the original version of his argument, Gauthier's other premise was – surprisingly – the standard version of the Self-interest Theory. He assumed that we should start by accepting S. We should believe that an act is rational if it will be expectably-best for us. He then claimed that it would be better for us if we changed our own conception of rationality, by moving from S to

G. Since it would be better for us if we made this change, S implies that it would be rational to do so. S tells us to believe that the true theory is not S but G. Gauthier concluded that the true theory *is* G.[41]

Shelly Kagan suggested the following objection.[42] If S is true, G must be false, since G is incompatible with S. If S is false, G might be true, but G would not be supported by the fact that S tells us to believe G. It is irrelevant what a false theory tells us to believe. Either way, Gauthier's argument cannot support his conclusion.

Gauthier later revised his argument. He no longer claimed that we should first accept S, and then move to his view. He argued directly that we should accept his view.[43]

In this version of his argument, Gauthier's main claim still seems to be that, if we accept his view, this will be better for us. What should his other premise be?

Though he no longer appeals to S, Gauthier might still say that, if it is in our interests to accept some belief, it is rational to do so. He could then keep his claim that it is rational for us to accept G.

As before, such a claim does not imply that G is true. It could be rational to accept a false theory. But Gauthier might think it enough to show that it would be rational to accept his view. He might say that, even in the sciences, we cannot prove our theories to be true. We can at most show that it is rational to believe them.

Such an argument, however, would conflate two kinds of rationality. When we claim that it would be rational to have some belief, we usually mean that this belief would be *theoretically* or *epistemically* rational, since we have epistemic reasons to have it. Such reasons *support* this belief, since they are provided by facts which either entail this belief, or make it likely that this belief is true. But Gauthier's argument does not appeal to epistemic reasons. His claim would be that, since it is in our interests to believe his view, this belief would be *practically* rational. When we have practical reasons to cause ourselves to have some belief, these reasons do not support this belief, since they are not related, in relevant ways, to this belief's truth.

The point could be put like this. Gauthier claims that it is in our interests to believe that certain acts are rational. He concludes that such acts *are* rational. This argument assumes

(D) If it is in our interests to believe that certain acts are rational, this belief is true.

Gauthier, however, rightly rejects (D). He imagines a demon who rewards various beliefs about rationality. He then claims that, if there were such a demon, it would be 'rational to hold false beliefs about rationality'.[44] Gauthier here concedes that, though it would be in our interests to hold these beliefs,

they would still be false. The fact that they would be in our interests could not make them true.

Could Gauthier withdraw this claim, and appeal to (D)?[45] It seems clear that he could not. Suppose that Gauthier's demon rewarded the belief that, for our acts to be rational, we must be called Bertie, and be wearing a pink bow tie. Gauthier could not claim that, if there were such a demon, this belief would be true. Nor do we need fantastic cases to refute (D). It might be in the interests of some people to have one belief about rationality, and in the interests of others to have some contradictory belief. Gauthier could not claim that these beliefs would both be true.

Since we should reject (D), we should reject this argument for Gauthier's view. Even if it were in our interests to believe his view, or rational to cause ourselves to believe it, this would not show that Gauthier's view was true.

The argument might show something. Gauthier might still claim that it would be practically rational to believe his view. But, unless he claimed that his view was true, Gauthier would have to abandon his main aim. He could not argue that it *is* rational to act morally. He could only argue that this belief is a useful illusion.[46]

5

In his discussion of nuclear deterrence, Gauthier gave a second argument for his view. Gauthier assumed that it could be rational to form the intention to retaliate, if one is attacked. He then claimed that, since it would be rational to form this intention, it would be rational, if deterrence failed, to act upon it.

David Lewis rejected this inference. While agreeing that it could be rational to intend to retaliate, Lewis denied that retaliation would itself be rational.[47]

In his reply, Gauthier denied 'that actions necessary to a rational policy may themselves be irrational'. If we accept deterrent policies, he wrote, we 'cannot consistently reject the actions they require'. Since we 'cannot claim that such actions should not be performed', we cannot call them irrational. 'To assess an action as irrational is . . . to claim that it should not be . . . performed'.[48]

These retaliatory acts cannot be *necessary* to deterrent policies since, if these policies succeed, these acts won't even be performed. But this is a special feature of deterrence, which we can set aside. In most of the cases with which we are concerned, the relevant acts will be performed. Thus, if I become trustworthy, because this disposition will be in my interests, I must expect that I shall keep my promises. Similarly, in Schelling's Case, I must expect my drug-induced state to affect my acts. In both cases, if I adopt the policy that will be good for me, I must expect to act in ways that will be bad for me.

Note next that, even in these cases, my acts aren't *required* by my policy. They aren't necessary to my policy's success. If they were, and my policy was good for me, my acts could not be bad for me. What is necessary to my policy is not my acts, but only my intention, or my disposition. My acts are merely the unwelcome side-effects.

This distinction, I believe, undermines Gauthier's reply to Lewis. If some policy is justified despite having bad effects, we may agree that, in one sense, these effects 'should occur'. But this only means, 'Things should be such that they occur'. And, in accepting that claim, we need not endorse, or welcome, these effects. The same applies to the acts that result from an advantageous disposition. We can agree that, in one sense, these acts should be performed. Things should be such that these acts will be performed. But we can still, consistently, believe these acts to be irrational.

6

Gauthier suggests another argument in favour of his view. This view avoids, he claims, 'some of the unwelcome consequences' of the Self-interest Theory. The chief such consequence is that, on that theory, it could be a curse to be rational.[49]

This argument does not, I believe, support Gauthier's view. One way to show that is this. Gauthier says that, even on his view, it might be a curse to be *cognitively* rational. This would be so if cognitive irrationality were directly rewarded. But this unwelcome consequence could not, he claims, be avoided by any theory.[50]

That is not so. Gauthier might extend his view. He might claim that our reasoning is cognitively rational if and only if it is in our interests. On this version of Gauthier's view, cognitive rationality could never be a curse. This revision would not, however, improve Gauthier's view. When crazy reasoning would be in our interests, that does not make it rational.

Cognitive irrationality could be in our interests, as any good theory should agree. So could practical irrationality. Both kinds of irrationality could be rewarded. It is no objection to the Self-interest Theory that it assumes or accepts these facts.

Gauthier makes one other claim in support of his view. He admits that, when his view is applied to Schelling's Case, it may seem counterintuitive. We may hesitate to claim that my crazy acts are rational. But Gauthier suggests that this is no objection, since 'whatever we might intuitively be inclined to say . . . "rationality" is a technical term in both Parfit's enquiry and my critique'.[51]

That is not so. I was asking what, in the ordinary sense, it is rational to want and do. And Gauthier claims that Schelling's Case 'shows that our

ordinary ideas about rationality . . . are sometimes mistaken'. Since Gauthier wishes to reject our ordinary ideas, he cannot defend his use of 'rational' by making it a mere stipulation. And that, in any case, would make his view trivial.

On Gauthier's view, acts are rational if they result from an advantageous disposition. Such acts are rational even if they are merely the regretted side-effects of this disposition, and are as crazy as we can imagine. That is very hard to believe. I have discussed what seem to me all of Gauthier's arguments for this view. None, I suggest, succeed. I conclude that we should reject this view. It could be in our interests to have some disposition, and be rational to cause ourselves to have it, but be irrational to act upon it.

If Gauthier drops these claims about rationality, he would need, I believe, to revise some other parts of his moral theory. But given the range and subtlety of Gauthier's theory, I cannot try to defend that claim here.

Notes

1. 'Reason and Maximization' , *Canadian Journal of Philosophy*, 4:411–33, 1975. This argument's fullest statement is in Gauthier's *Morals by Agreement* (Oxford University Press, 1986), henceforth *MA*.
2. In an unpublished paper 'Rational Irrationality', and later in Sections 7–8 of my book *Reasons and Persons* (Oxford University Press, 1984).
3. This paper was completed in 1994, in response to Gauthier's contribution to *Reading Parfit*, edited by Jonathan Dancy (London: Routledge, 1997). I regret that, having since become obsessed with Kant's ethics, I have not tried to take into account Gauthier's most recent work.
4. Since Gauthier means by our *utility* the fulfilment of our *present* considered preferences, what he appeals to is, strictly, the *Deliberative Theory*. But, as Gauthier remarks (*MA*, p. 6), most of his claims apply equally to the Self-interest Theory. And Gauthier often uses words, like 'benefit' and 'advantage', that refer more naturally to our interests rather than our present preferences. So we can here ignore the differences – though they are often great – between the Deliberative and Self-interest Theories. We can suppose that, in all of the cases we discuss, our present considered preferences would coincide with what would be in our own interests.
5. What is expectably-best may not be the same as what we can expect to be best. Some acts are expectably-best for us though we can know, for certain, that they will not actually be best for us. Trying to do what is actually best may be, given the risks, irrational.
6. *Reasons and Persons*, Sections 7–8.
7. Gauthier gave this reply in *MA* (especially pp. 173–4). In his contribution to *Reading Parfit*, Gauthier later gave up the claim that we could not deceive others. He suggested that, if we remained self-interested, and merely appeared to be trustworthy, that would be worse for us. Thus he writes: 'the overall benefits of being able to promise sincerely . . . may reasonably be expected to outweigh the

overall costs of keeping promises when one could have gotten away with insincerity' (p. 26). But, if we could get away with insincerity, what are the benefits from being able to promise sincerely? Gauthier might appeal, like Hume, to the benefits of peace of mind, and a good conscience. But that seems insufficient for his purposes. Gauthier also claims that, even if we were generally trustworthy, we would be able to make some insincere promises. But this merely limits the costs of sincerity. It does not suggest that there is any gain. For Gauthier's distinctive argument to get off the ground, he needs, I believe, his earlier assumption that we could not rationally hope to deceive others.

8. See, for example, *MA,* Chapter VI.
9. In *Reasons and Persons*, Sections 7–8.
10. I also supposed that it might be rational to change our beliefs about rationality. This, too, was intended to help Gauthier's argument. If we did not change our beliefs, we would be doing what we believe to be irrational, and that might seem enough to make our acts irrational. But this element need not concern us here.
11. As he wrote (like Queen Victoria), 'We are unmoved' (*MA*, p. 185).
12. Gauthier asserted (B) – which he calls his 'second level of commitment' – in *Reading Parfit*, p. 40. I discussed a similar claim, which I called '(G1)', in *Reasons and Persons* (p. 13). On Gauthier's second level of commitment, it is rational to act on a disposition 'so long as one reasonably expects past and prospective adherence to the disposition to be maximally beneficial'. This claim may seem to mean 'if one both reasonably believes that adherence to this disposition in the past has been beneficial, and reasonably expects that adherence to it in the future will be beneficial'. But this cannot be what Gauthier intends, since it would remove the difference between his second level of commitment and his first level (discussed below). Gauthier must mean: 'if one can reasonably believe that acquiring it was beneficial in one's life as a whole, taking the past and future together'.

    Gauthier's move from (A) to (B), or from his third to his second level of commitment, hardly damages his defence of rational morality. On the view defended in *MA,* for morality's constraints to have rational force for us, accepting these constraints must have been expectably-best for us. On Gauthier's revised view, for these constraints to have rational force, they must also be known not to have been on the whole bad for us. Most of contractual morality's constraints would meet this second requirement.
13. Perhaps I would have obeyed some order that would have proved fatal.
14. It may be objected that I acquired too crude a disposition. Perhaps I should have become disposed to ignore threats, except in cases in which I believed that acting in this way would be disastrous. But, as Gauthier says, 'I may reasonably have believed that any qualification [to my disposition] would reduce its *ex ante* value, so that unqualified threat-ignoring offered me the best life prospects' (*Reading Parfit*, p. 39). We can add the assumption that only the unqualified disposition would in fact have been as good for me. (There is another reason not to allow this disposition to take this qualified form. If we did, we must allow similar qualifications to the disposition of trustworthiness. As we shall see, that would undermine Gauthier's argument.)
15. Gauthier endorses the action of a would-be deterrer who, when deterrence fails, disastrously carries out her threat. He writes 'Her reason for sticking to her guns . . . is simply that the expected utility . . . of her failed policy *depended* on her willingness to stick to her guns' ('Deterrence, Maximization, and Rationality',

Ethics, 94:489, 1984). So what? Her expectation may have depended on that willingness. But why should she remain faithful now?

16. Note that, in claiming this, I need not appeal to the Self-interest Theory, S. I need not assume that this attempt would be rational because it would be likely to be good for me. Since Gauthier rejects S, that would beg the question. But even on Gauthier's theory, it would be rational for me to try to lose this disposition. Suppose that I lose my dispositions whenever they become disastrous. It would be in my interests to have this meta-disposition. So, on Gauthier's theory, it would now be rational for me to act upon it.
17. Suppose first that, if I tried, I could cease to be a threat-ignorer. As I have just argued, it would then be irrational for me to keep my disposition. If Gauthier accepts this conclusion, could he still assert (B)? Could he claim that, even though it would now be irrational to *keep* my disposition, it must still be rational to act upon it?

    There may be certain cases in which, though it would be irrational to keep some disposition, it would still be rational to act upon it. Suppose, for example, that it would be irrational for me to remain prudent. If I did, irrationally, keep this disposition, it might still be rational to act upon it, doing whatever would be best for me. (B), however, is a much stronger claim. According to (B), even if it would now be irrational to keep some disposition, it *must* still be rational to act upon it, simply because it *once* brought benefits that were greater than its present costs. This claim, I believe, cannot be true. If it is irrational to keep this disposition, why must it be rational, if I do keep it, to act upon it?

    If I have irrationally remained prudent, there is a different explanation of why it can be rational to act upon this disposition. Doing so will be better for me. The rationality of this act need not be defended by an appeal to the rationality of the disposition, or of my having kept the disposition, upon which I act. Things are quite different with ignoring your threat, in a way that I know will be disastrous for me. If this act is to be claimed to be rational, that can only be by an appeal to the rationality of the disposition on which I am acting. And if it is now irrational for me to keep this disposition, there seems no reason to conclude that, if I keep it, it must be rational for me to act upon it.

    Suppose, next, that I could *not* lose my disposition, even if I tried. Gauthier might say that, if that is true, it is not irrational for me to keep this disposition. This is not something that I *do*. But it *would* be irrational for me to keep it, if I *could* lose it. This seems enough to undermine the claim that it must still be rational to act upon it.
18. (C) is one interpretation of what Gauthier calls the 'weakest' version of his view, or what he calls his first level of commitment. On this view, he writes, one should act upon some disposition, even though one's actions are 'costly . . . only so long as one reasonably expects adherence to the disposition to be prospectively maximally beneficial' (*Reading Parfit,* p. 39).

    When Gauthier talks of 'adherence' to this disposition being beneficial, he must mean continuing to *have* this disposition. *Acting* on this disposition may be, as he agrees, costly. I shall also take 'adherence' to mean 'present adherence'. Though Gauthier might mean 'adherence now *and in the future*', that would make his claim less plausible. It would not cover cases where it would be advantageous first to acquire and then to lose some disposition. (Suppose that, while it was indeed better to acquire some permanent disposition than not to acquire it at all, it would have been expectably-best to acquire it simply for a

time. Acquiring this permanent disposition was not then, as Gauthier requires, 'maximally beneficial'.)

19. My drug-induced insanity, Gauthier claims, is 'the rational disposition in such situations, and the actions to which it gives rise are rational actions' (*Reading Parfit*, p. 38). Gauthier means only that it is in my interests to have this disposition *now*. He is not here concerned with a choice between two permanent dispositions. If I had to choose my disposition, not just until the police arrive, but for the rest of my life, it would be better to remain sane and give the man my gold.
20. *MA* (*passim)*.
21. Gauthier might extend his claim about translucency. He might say that we could not have reason to believe that, if we broke our promises, we could keep this fact secret. But this reply would jettison what is novel in Gauthier's view, since it would revert to the ancient claim that honesty is always the best policy.
22. There is one reading on which this claim must be true. It may be said that, if we are able to suspend our disposition, we were not *truly* trustworthy. But this reading is irrelevant since, for Gauthier's purposes, all that matters is whether we *appeared* trustworthy. It would be quite implausible to claim that, if we break some agreement, we cannot have earlier appeared to be trustworthy, even if, at the time, we sincerely intended to keep this agreement.

    If this claim is to help Gauthier's case, he must make other revisions in his view. He writes: 'a disposition is rational if, among those humanly possible, having it will lead to one's life going as well as having any other' (*Reading Parfit*, p. 31). This appeal to *human* possibility seems at odds with other parts of Gauthier's view. He claims elsewhere that we should not ask which dispositions are in general rational, since the answer may depend on a particular person's circumstances. Thus he writes, '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 need to suppose that there need be a single supremely rational disposition' (*Reading Parfit*, pp. 31–2). A person's circumstances can surely include what is possible for this person.

    This appeal to human possibility also raises a problem for Gauthier's argument. Trustworthiness is *not* the disposition that, among those *humanly* possible, is most advantageous. It would be more advantageous to appear to be trustworthy but to be really prudent; and that is surely possible for some human beings. If Gauthier appeals to what is humanly possible, he would have to judge trustworthiness to be an irrational disposition, even when it is had by people for whom, since they could not deceive others, it is the most advantageous possible disposition.
23. At one point, Gauthier may make this move. While honesty is the best policy, Hume writes, there may be some exceptions. According to Hume's 'sensible knave', he is wisest 'who observes the general rule, and takes advantage of all the exceptions'. Gauthier replies that, to be rational, we must be disposed to keep our promises, since this disposition will be best for us. He then writes, 'such a person is not able, given her disposition, to take advantage of the "exceptions"; she rightly judges such conduct irrational' (*MA*, p. 182).
24. In the doctrine that 'ought' implies 'can', the sense of 'can' is compatible with determinism. If that were denied, and we assumed determinism, we would have to claim that *every* act is rational.
25. It would of course be better if I merely appeared to be insane. But we can suppose that this is not possible, since if I had not taken the drug, the robber would know

this. (Perhaps one of the drug's effects is a characteristic look in the eyes; or perhaps I can convince the robber only if he sees me drink this drug.) Being actually in this state is then the disposition that is best for me.

26. *Reading Parfit* (p. 37).
27. Provided, of course, that these bad effects do not outweigh the good effects of my disposition. Gauthier need not claim that, if I killed myself or my children, that would be rational.
28. It may be said that, in one respect, Gauthier's view is less extreme than Hume's. Even if my act has bad effects, these must be outweighed by the good effects of having my disposition. But we can remember here that, on Gauthier's main view, I maximize my utility if I fulfil my present considered preferences, and these need not coincide with my interests. As on Hume's view, these preferences could be as crazy as we can imagine. The difference between these views is that, on Hume's view, for my act to be rational, I must at least be trying to fulfil my aims, while on Gauthier's view, my acts need only be the side-effects of a state the having of which will achieve these aims.
29. 'Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices' (*MA*, p. 187).
30. It may seem that, if that is true, breaking our promises cannot be better for us. But this may not be so. The bad effects come, not from our breaking of these promises, but from the fact that we are both translucent and disposed to break our promises whenever this will be better for us.
31. It is worth explaining why. In our assessment of the good or bad effects of our dispositions, we include the acts to which these dispositions would or might lead. If it is best for us to have some disposition, even though this will lead to acts which are bad for us, those effects must be outweighed. Since the assessment of our dispositions includes the assessment of our acts, but goes beyond it, this is the assessment that tells us what on balance will be best for us.
32. *MA* (p. 170).
33. It may be questioned whether G tells us, if we can, to acquire these dispositions. That does not follow from the fact that, if we do, that will be better for us. If G does not tell us to act in this way, that would be an objection to G, and would again undermine Gauthier's argument. But Gauthier might claim that, in trying to acquire these dispositions, we would be acting on an advantageous, or maximizing, meta-disposition.
34. He would admit that, in practice, few of us are always rational. But he might claim that, in assessing the plausibility of these theories, we should consider what would happen if we always did what they told us to do. He might then claim that, if we fully followed S, we would always maximize at the level of our acts.
35. It may be objected that, if we cannot always do what S claims to be rational, S cannot claim that we ought to do so. 'Ought' implies 'can'. But this confuses two questions. When I say that we cannot always do what S claims to be rational, I mean that this is not causally possible. This is the kind of possibility that is relevant when we are comparing the effects of our having different dispositions. The sense of 'can' that is implied by 'ought' does not, as Gauthier agrees, require such causal possibility, since this other sense of 'can' is compatible with determinism.
36. It may seem that, if we cannot always do what S tells us to do, there is no way of predicting when we shall follow S. That is not so. Suppose that we are now

always disposed to do what we believe to be rational. If we know that we can acquire maximizing dispositions, we shall then do so, even though we know that this will cause us later to act irrationally. Acquiring these dispositions is, according to S, the rational thing to do. It is only *after* acquiring these dispositions that we shall start acting in ways that S claims to be irrational.

37. In 'Deterrence, Maximization, and Rationality', and in *The Security Gamble*, ed. Douglas MacLean (Totowa, NJ: Rowman & Allanheld, 1984).
38. 'Afterthoughts', in *The Security Gamble* (pp. 159–61).
39. Cf. Edward McClennen, 'Constrained Maximization and Resolute Choice', *Social Philosophy and Public Policy*, 5:95–118, 1988.
40. Such a claim is fairly plausible in the case of trustworthiness, the disposition that is Gauthier's chief concern. If we could not conceal our intentions, as he assumes, it might be better for us if we intended to keep our promises, even when this way of acting would be worse for us. Unless we have this intention, others might exclude us from advantageous agreements. And, for us to be able to form this intention, we might have to believe that it is rational to keep such promises.
41. 'Constrained Maximization'.
42. In a letter to me.
43. See *MA* (p. 182) and *Reading Parfit* (p. 31). (But see also *MA*, pp. 170 and 158.)
44. *Reading Parfit* (p. 36).
45. At one point, Gauthier comes close to accepting (D). He cites my book's version of (D) – there called '(G2)' – and writes, 'to this extent I accept . . . (G2)' (*Reading Parfit*, p. 40).
46. It may seem that, in making these remarks, I have presupposed a naively realistic view. Gauthier might say that a normative theory could not be *true*. But this would not rescue Gauthier's argument. Even on a noncognitivist view, we must give some content to the notion of a normative belief. We must be able to claim that an act *is* rational, and be able to assert or deny different theories. My remarks could be restated in these terms.
47. In *The Security Gamble*.
48. 'Afterthoughts', in *The Security Gamble*, pp. 159–61.
49. *Reading Parfit*, p. 30.
50. *Reading Parfit*, p. 36.
51. *Reading Parfit*, p. 38.