RATIONAL INTENTION, RATIONAL ACTION

by

Joseph Mintoff

BMath(Hons), GDipPhil, MA(Hons), W'gong

A thesis submitted for the Degree of
Doctor of Philosophy
of the Australian National University
Division of Philosophy and Law
Research School of Social Sciences

30 April 1993
PART III

RATIONAL DILEMMAS
Preliminaries

An Annoying Complication

My argument against the Self-Interest Theory, and in defence of the possible rationality of toxin-drinking and cooperation, depends on the possibility of certain types of situations. One objection to my argument is that these situations are actually not possible. However, I argued in Chapters Three and Six, respectively, that this objection is mistaken. My argument depends, in addition, on the existence of certain relations between what one rationally ought to do and what one can do:

\[(OP_n)\] If an agent rationally ought to $S_1$, rationally ought to $S_2$, ..., and rationally ought to $S_n$, then it is logically possible that he $S_1$'s, he $S_2$'s, ..., and he $S_n$'s. ($n = 1, 2, 3, ...$)

\[(OC_n)\] If an agent rationally ought to $S_1$, rationally ought to $S_2$, ..., and rationally ought to $S_n$, then he can be such that he $S_1$'s, he $S_2$'s, ..., and he $S_n$'s. ($n = 1, 2, 3, ...$)

My opponents may grant that the type of situations I have in mind are possible, but may nevertheless deny the implication I drew – that the Self-Interest Theory is false. They may claim I have just shown principles \((OP_n)\) and \((OC_n)\) are false. In light of this possible objection, I will spend the final part of the thesis examining whether there are independent arguments against these principles.

[1] Before doing so, however, I need – for the sake of correctness – to note an annoying complication. It comes in the form of two counterexamples to principles \((OP_n)\) and \((OC_n)\). These are: (a) The Good Samaritan Paradox.\(^1\) I have every reason not to rob the bank to get the

---

\(^1\) This type of example is usually put in terms of the (moral) obligation of the so-called good samaritan to help someone who has been the victim (by someone else) of robbery; I put the example in terms of a (rational) obligation for one to correct a past (rational) misdeed. See A. N. Prior, 'The Paradoxes of Derived Obligation,' *Mind* 63 (1954), pp. 64-5; R. Chisholm, 'Contary-to-Duty Imperatives and Deontic
money I need for my new car, and since the likely effect of doing so is that I will get caught and sent to jail, I rationally ought not to rob the bank. Nevertheless, as a whim and quite irrationally, I do. I have all the reason I need to return the money, since, in the likely event I am caught, this will lessen my sentence; therefore, I rationally ought to return the money. Yet, clearly, it is not logically possible for me, thus, I cannot be such as, to return the money I stole from the bank and not have robbed the bank in the first place. (b) The Paradox of the Procrastinator. I receive an invitation to review a book, and must decide first whether to accept, and second whether to write it. I can do both, and it is (let us suppose) best if I did both: I rationally ought to accept and write the review. However, I am a procrastinator: if I were to accept I would not write the review. This is the worst possible outcome: I rationally ought not accept. Yet, clearly, it is not logically possible for me, and so I cannot be such as, to accept and then write, and not accept.3

Principles (OPn) and (OCn) obviously need to be altered. But how? The answer lies in the fact that, in both of these examples, there is something I rationally ought to do which I rationally ought to do only because I did something, will do something, or am doing something, I rationally ought not. In the Good Samaritan Paradox, I rationally ought to return the money only because I stole the money in the first place, which I rationally ought not to have done. In the Procrastinator Paradox, I rationally ought to say no to the review only because I will fail to both say yes and afterwards write it, which (both saying yes and writing) is what I rationally ought to do. If I rationally ought to A, and this is so only because I have done, will do, or am presently doing, something I rationally ought not, then we will say that my rational


3 There are, of course, various solutions offered for these paradoxes. For attempted solutions to both, see W. Sinnott-Armstrong, Moral Dilemmas, (Oxford: Basil Blackwall, 1988), pp. 143 ff. Instead of offering my own solution, I will, in this thesis, sidestep these issues by suitably modifying principles (OCn) and (OPn).
obligation to A is a second-best (rational) obligation. My returning the money, and saying no to the review, are both second-best rational obligations. In light of all this, the qualification we need to introduce to principles (OP_n) and (OC_n) should be clear:

\((OC'_n)\) If an agent rationally ought to \(S_1\), rationally ought to \(S_2\), ..., and rationally ought to \(S_n\), and none of these obligations are second-best, then he can be such that he \(S_1\)'s, he \(S_2\)'s, ..., he \(S_n\)'s. (for \(n = 1, 2, \ldots\))

\((OP'_n)\) If an agent rationally ought to \(S_1\), rationally ought to \(S_2\), ..., and rationally ought to \(S_n\), and none of obligations are second-best, then it is logically possible that he \(S_1\)'s, he \(S_2\)'s, ..., he \(S_n\)'s. (for \(n = 1, 2, \ldots\))

In short then, and roughly speaking, the principles I wish to endorse claim that if none of one's rational obligations arises from a (past, present, or future) mistake on one's part, then they will all be jointly satisfiable.

[2] This complication does not effect the arguments of previous chapters against the Self-Interest Theory, and nor will it affect the conclusions to be reached in the following chapters.

On the one hand, the conclusions concerning the Toxin Puzzle will not be affected. If the First Counterexample is consistent, and the Self-Interest Theory is true, then you rationally ought to adopt the intention, but you rationally ought not to drink the toxin. Neither of these obligations is a second-best obligation: you rationally ought to adopt the intention, not because you have done (or will do, are doing) something irrational, but simply because this is what will get you the million; you rationally ought not to drink the toxin, not because you have done (or will do, are doing) something irrational, but simply because this is what will prevent you from suffering a day's severe illness. But if the First Counterexample is consistent, you are not able to both adopt the intention and yet not drink the toxin. It again follows

---

4 These principles need further elaboration, particularly with respect to the times of the actions and rational obligations. I shall not assume that the time of the rational obligations is the same for all \(S_j\), but I will assume (what I believe independently to be correct, but shall not argue) that the time of the rational obligation to \(S_j\) is the same as the time of \(S_j\). See H. Goldman, 'Dated Rightness and Moral Imperfection', Phil Rev 85 (1976): 449-487, for an explanation of the sort of temporal indexing of actions, abilities and obligations I have in mind, and for the reasons one might think such indexing is necessary.
- this time by \((OC'_2)\) rather than \((OC_2)\) – that the Self-Interest Theory is false, and that you are rationally permitted to drink the toxin, even though you are free not to, and it has the best outcome for you not to. Similar points apply, \(mutatis mutandis\), to the discussion of the Second Counterexample.

On the other hand, the conclusions concerning the Prisoner’s Dilemma also will not be affected. If the Third Counterexample is consistent, and the Self-Interest Theory true, you rationally ought to become an Agreement-Keeper, rationally ought to believe we have agreed, but rationally ought not to cooperate. None of these obligations is a second-best obligation: you rationally ought to become an Agreement-Keeper, not because you have done (or will do, are doing) something irrational, but simply because this is what will get the other to cooperate with you; you rationally ought to believe we have agreed, not because you have done (or will do, or are doing) something irrational, but simply because this is what the evidence points to; and you rationally ought not to cooperate, not because you have done (or will do, are doing) something irrational, but simply because, whether or not the other cooperate, you always do better not to. But, whether or not the Third Counterexample is consistent, it is not logically possible for you to become an Agreement-Keeper, believe we have agreed, and yet not cooperate. It again follows – this time by \((OP'_3)\) rather than \((OP_3)\) – that the Self-Interest Theory is false, and that you are rationally permitted to cooperate, even though you are free not to, and it has the best outcome for you not to.

This change also does not much affect the arguments in the following chapters. Most discussions of principles like \((OP_n)\) and \((OC_n)\) have occurred – appropriately enough – in discussions of deontic logic, and have resulted in counterexamples such as those above. Instead of going over old ground, though, I want in this final part of the thesis to examine the implications of other discussions – those of Micheal Slote on rational dilemmas, Derek Parfit on rationally causing oneself to act irrationally, and Gregory Kavka on the paradox of deterrence – for principles \((OP_n)\) and \((OC_n)\), to see if they provide any independent reason for rejecting these principles, or their cousins \((OP'_n)\) and \((OC'_n)\). In the rest of the thesis, then, I will put this annoying complication firmly to one side.
Chapter Seven
Michael Slote on Rational Dilemmas

In Chapter Five of his *Beyond Optimising*, Michael Slote argues that what he calls *rational dilemmas* are possible. However, if rational dilemmas were possible, then (as we shall see) principles *(OP_n)* and *(OC_n)* would be false. As a result, I will in this Chapter examine his arguments for the possibility of rational dilemmas. We shall see that they do not establish this possibility, and so do not threaten principles *(OP_n)* and *(OC_n)*.

§1 Rational Dilemmas

The model for so-called rational dilemmas are *moral* dilemmas. One of the paradigms of someone facing a moral dilemma is Agamemnon, who has been commanded by Zeus to lead the Greek expedition to Troy. Zeus is chief amongst the gods: Agamemnon therefore has an obligation to Zeus and to his troops to lead the expedition to Troy. Another god, Artemis, has however storm-bound the expedition in Greek waters, and demands that he sacrifice his daughter, Iphigenia, before she will let the expedition get under way. But to kill his daughter by his own sword would be to perform a horrible and guilty act: he has an obligation not to slaughter his own child. He faces a bitter choice, for he cannot satisfy both of these solemn obligations, in a situation not of his own making. Agamemnon faces a moral dilemma, it is said, because there are things each of which he morally ought to do, though both of which he cannot do.

---

1 All page references in this chapter are to M. Slote, *Beyond Optimising: A Study of Rational Choice*, (Cambridge, Mass.: Harvard Univ. Pr., 1989), Ch. 5. Slote is the only one who, to my mind, explicitly endorses the possibility of rational, rather than moral, dilemmas.


In Chapter Five of his *Beyond Optimising*, Michael Slote argues that *rational* dilemmas are possible. Slote offers a definition of such dilemmas, which I shall need to generalise. I say that an agent faces a *rational dilemma* (with respect to states $S_1, S_2, ..., S_n$) if and only if, through no rational fault of their own, and all at the same time, they rationally ought to $S_1$, rationally ought to $S_2$, ..., rationally ought to $S_n$, but either they cannot be such that, or it is logically impossible that, they $S_1$, they $S_2$, ..., and they $S_n$. This definition differs slightly from that of Slote's. He refers only to conflicting *actions*, but I intend to refer, in addition, to beliefs. I need to do this because, in my discussion of the Prisoner's Dilemma, it appeared you rationally ought to *become an Agreement-Keeper* (an action), rationally ought to *believe you have agreed* (a belief), but rationally ought *not to cooperate* (an action). I am justified in doing this because, if – as Slote argues – rational dilemmas involving only conflicting *actions* are possible, then, presumably, so are those involving conflicting actions and beliefs.

If Slote's arguments for the possibility of rational dilemmas are correct, then principles $(OP_n)$ and $(OC_n)$ are false. Slote's arguments, if they are correct, show that, through no rational fault of one's own, and all at the same time, one could be rationally required to $A$, rationally required to $B$, even though one cannot both $A$ and $B$. If it is possible one rationally ought to $A$ and rationally ought to $B$ but cannot do both, and this is so all at the same time, then, presumably, it is also possible if the obligation to $A$ and the obligation to $B$ do not occur at the same

---

4 In fact, Slote suggests a number of definitions. (a) 'For there to be a rational dilemma, there must be a situation in which, through no practical fault of her own, an agent finds it impossible to act rationally, to make a rational choice among two or more alternatives' (p. 100-1). Rational dilemmas, for Slote, are conflicts between two or more alternatives not resulting from a mistake on the part of the agent. (b) He suggests, later on, that a rational dilemma occurs when '[t]he cumulative effect of the rational principles in force in a given situation may be to prescribe $a$ without qualification and prescribe $b$ without qualification, even though they cannot both be performed.' (p. 102), and that '[i]n order to have an example of a rational dilemma, we need to describe a situations where, all things considered, every practical solution is fatally flawed and rationally unacceptable' (p. 103). I take it, then, that the conflicts of obligation with which Slote is concerned are conflicts between what I am calling rational oughts.

5 Note that if one rationally ought to $A$ and rationally ought to $B$, through *no rational fault of one's own*, then these rational obligations are not *second-best* obligations, and so principles $(OC_n')$ and $(OP_n')$ will be just as applicable to Slote's discussion as are $(OC_n)$ and $(OP_n)$.
time. In a word, if *synchronic* dilemmas are possible, then surely so are *diachronic* dilemmas. Therefore, if Slote’s argument’s are correct, they also show it is possible that one rationally ought to A, rationally ought to B, even though one cannot both A and B. If his arguments are correct, then, (OP2) and (OC2) are false, and so, by extension, are (OPn) and (OCn). This fact explains the attention I will give in this chapter to Slote’s arguments for the possibility of rational dilemmas. From this point onward, then, I will have no occasion in this chapter to refer to principles (OPn) and (OCn), but will focus simply on the possibility of rational dilemmas.

The onus of proof is on someone like Slote who would defend the possibility of rational dilemmas. In the first place, if rational dilemmas were possible, we would have to admit there could be a true deliberative, (strongly) action-guiding, absolute and unconditional judgement that you ought to perform some action A, a true deliberative, (strongly) action-guiding, absolute and unconditional judgement that you ought to perform some action B, even though you cannot do both. This is, absolutely and unconditionally, incredible.6 In the second place, even amongst those who claim that *moral* dilemmas are possible, some, such as Bernard Williams, would claim it is *not* possible for there to be what I am calling rational dilemmas. Williams admits that ‘I have to act in the conflict; ... In thinking about this, or asking another’s advice on it, the question I may characteristically ask is “what ought I to do?” The answer to this question, from myself or another, cannot be “both”’. What he denies, though, is that this ought is the moral ought. Even though someone may face a moral dilemma, he will say, they need to make a final decision, and this cannot involve a dilemma.7 In the third place, there is a long history of thought that

---

6 Apologies to Earl Conee, in ‘Against Moral Dilemmas,’ *Phil Rev* 91 (1982): 87-97, for this (rhetorical) way of putting the point.

even moral dilemmas are not possible, and Slote needs to address the concerns of those who believe this.8

The burden of proof is therefore on someone like Slote who would defend the possibility of rational dilemmas. What arguments does he have to discharge this onus?

§2 Rational Dilemmas, and Infinite Choices

What Slote needs to defend the possibility of rational dilemmas is to show that there could be a situation in which one rationally ought separately to perform each of a group of actions, even though one cannot, or it is not logically possible to, jointly perform all the actions in that set. Slote provides an original and ingenious example to show this. Unfortunately, as we shall see, it fails to establish what Slote wants.

[1] Slote introduces a number of such examples, each similarly structured.9 Here is one such example:

Imagine that God has condemned some wine connoisseur to an infinite life with only as finitely much of his favourite wine, Chateau Effete, as he asks for on a certain occasion. How many bottles of Chateau Effete should he ask for as a consolation for the unpleasant tedium of his largely wineless immortality? (p. 114)

Suppose you are the unfortunate (or is it fortunate?) wine connoisseur. Ought you, or ought you not, pick zero bottles? Since you would enjoy it more, there is obviously more reason to pick any number of bottles than to pick none at all. Hence you ought not to pick zero bottles — you ought to pick at least one bottle. Ought you to pick (exactly) one? No,


9 Slote considers other versions of this example (e.g., pp. 111 ff.), but I have chosen to concentrate on the one which, at first sight, I believe most plausible. Furthermore, the interpretation of the example is my own, and not Slote’s, who says that in the situation described, whatever you do is irrational. It fits more neatly with my definition of rational dilemmas to take the interpretation I have, but, in any case, I believe they amount to the same thing.
since you would to better to pick (exactly) two, there is more reason for you to pick two, and so you ought not pick one. Ought you to pick (exactly) two? Again no, and for exactly the same type of reason. Continuing ad nauseam leads to the conclusion that (i) you rationally ought to pick at least one bottle of Chateau Effete, but (ii) for each amount, $n$, of bottles, you rationally ought not to pick $n$ bottles. This is an (infinite) rational dilemma, since it is not logically possible to fulfil all these 'oughts'.

This example represents a formidable challenge to my claim that rational dilemmas are not possible. Even though it concerns infinite rational dilemmas, it is still worth examining the argument in more detail. What the situation described establishes is the first premise to the effect that:

(1) It is logically possible that there be an infinite sequence of actions $A_0, A_1, ...$ such that (i) you are free to perform each action or not, (ii) you logically cannot fail to perform one of these actions, and (iii) performing any action in the sequence has better consequences for you than its predecessor.

The action $A_n$ is the speech act of requesting $n$ bottles of Chateau Effete for the rest of eternity, each action of which you are free to perform or not, though you logically cannot but perform one such action. And it is part of the example that picking more rather than less bottles has better consequences. The next premise of the argument is that

(2) If action A has better consequences for you than action B, then there is more reason for you to perform A than there is to perform B.

This premise asserts that the strength of reasons for action goes by the value of the outcome to the agent of those actions. It is a premise Slote needs for his argument (though he does not explicitly introduce such a claim), and also happens to be a consequence of the Self-Interest Theory’s view – $(S')$ in Chapter One – of reasons. The third premise is that

(3) If there is more reason for you to do A than there is to do B, then you rationally ought not to do B.
This premise is a consequence of the analysis, (R), of the rational 'ought' I introduced in Chapter One, and is a claim Slote does introduce (pp. 112, 115). The inference to be drawn from these premises, as we have seen, is that (infinite) rational dilemmas are logically possible. In the end, I shall deny either the truth or the applicability, in this case, of premise (2). Before doing so, however, I briefly consider the other two premises.

[2] In the first instance, one may deny the first premise, and claim that the situation Slote describes is not, contrary to appearances, logically possible. I believe this objection is not successful in the final analysis, but is nevertheless worth examining.

An important part of Slote's example is that you are in fact able to ask for any amount of bottles of Chateau Effete 'on a certain occasion' (p. 114). This means Slote faces a problem. Either this 'certain occasion' extends only a finite stretch of time, or it does not. On the one hand, if it does (and this is what Slote probably intends), then there will be some number of bottles of Chateau Effete you cannot ask for on that occasion. To ask for a certain number of bottles, you will have to name a particular number. But naming numbers takes time. For any naming system for the natural numbers – 0, 1, 2, 3, ... – there will presumably be some number whose name takes longer to express than the finite stretch of time allowed for choosing. This is clearly the case with the decimal naming system – '0', '1', '2', '3', ... – and it will be true for any naming system. On the other hand, if this 'certain occasion' does not extend only for a finite stretch of time, then it is no longer clear that picking more bottles is always better than picking less. In such a case, the costs and benefits of picking n bottles will be twofold: the costs attaching to naming n, and the benefits deriving from having n bottles of wine. Yet, the costs attaching to naming n (eg, spending ten millennia doing so) may outweigh the benefits of that much wine, and so it may have better consequences to pick some lesser amount of

10 Frank Jackson, in conversation, made this point. This formulation of the objection, though, is my own.

11 This argument assumes the number needs to be named, and, it may be objected, this is not so. God could present you with the semi-open interval (0,1] and ask you to pick a point on this line, assuring you that, if you pick point x, then he will give you 1/x bottles of wine. In response, however, it will still follow on this scheme that there are some amounts of wine you cannot request, since, it is clear, beyond some (sufficiently small) value of x, it would be false to say you could pick point x.
wine. It follows, then, that either there is some amount of wine you cannot request (on the 'certain occasion'), or it will no longer be clear that picking more bottles is always better than picking less.

While these points are effective against Slote's formulation of the example, it seems to me that the difficulties with naming the relevant number are not what lies at the heart of the example. There should be some way of re-formulating the example to avoid these problems. There is.

Suppose we have some type of act which can only be done once, but the longer it is postponed the greater the expectation value for doing it. Thus (5) and (8) [Pollock's formulation of Consequentialism] would counsel never doing it, because for each strategy prescribing doing it at a particular time, there is a strategy with higher expectation value which prescribes doing it at a later time. For example, a bottle of fine wine normally improves with age for a while, but then goes bad. Consider, however, a bottle of EverBetter Wine which continues to get better forever. When should we drink it?

Pollock does not assume - but, like Slote, should - that God has offered you an infinite life. (After all, if you are going to die, then you should presumably drink the wine just before your death.) The same problem arises. Should you drink it today? No - it will be better tomorrow. Should you drink it tomorrow. No - it will be better the day after. And so on. Denying the possibility of the type of example Slote has in mind, then, will not defuse the problem.

[3] A second possibility for responding to Slote's argument would be to deny the third premise - that if there is more reason to do A than there is to do B, then the agent rationally ought not to do B. I myself am not

---

12 This argument assumes there will in fact be a point at which the costs of naming n outweigh the benefits of having n bottles of wine, and, it may be objected, this need not be so. In response, I am not so sure.

13 J. L. Pollock, 'How do you Maximize Expectation Value?', *Nous* 17 (1983), pp. 417 ff. Strictly speaking, this is not a rational dilemma, as I am using the term, since the conflicting (non-) actions fail to occur at the same time. Still, it is sufficiently troublesome to warrant the consideration I give it.

14 Bernard Williams suggests, in 'The Makropolis Case: Reflections on the Tedium of Immortality,' in his *Problems of the Self*, (Cambridge: Cambridge Univ. Pr., 1973): 82-100, that beyond a certain point (about 200 years), it is not better for one to live longer. If this is so, then the example Slote describes will not support the first premise, and one will have reason to doubt whether any such example is possible. However, I believe Williams is wrong, and will not examine this response.
attracted to doing so, but I will briefly examine a suggestion of Michael Slote's (pp. 115 ff.) which might motivate such a denial.

What we could call morally 'acceptable' actions are ones you are morally permitted to perform even though ones are available you have more moral reason to perform. What the current example shows, or so it might be claimed, is that there is corresponding to this a notion of rationally 'acceptable' actions - those you are rationally permitted to perform even though ones are available you have more reason to perform.

Does this notion of rationally 'acceptable' action make sense? Here is an example Slote provides to show that it is a notion more common than might be thought:

A man is angry at his boss and believes ... that it would be in his own interest, and on the whole a good thing, to tell the boss off when they next meet. He believes that it would be best to do so in a loud enough voice so that everyone in the office will know what is happening. But both he and his fellow employees have long been intimidated by the boss, and the employee knows it will be difficult to stand up to the boss and tell him what he thinks, and even more difficult (though not impossible) to do so in the loud and angry tones he thinks most appropriate. When the time comes to confront his boss, he manages with considerable effort to speak his mind, but allows himself to be intimidated to the extent of not daring doing so loudly or angrily; ... Even though his performance may be less than the best that lay within his power, it may be good enough for him and for us not to regard it as rationally unacceptable. (p. 119)

The employee has managed to do something that, while not the best he could have done, we feel is rationally acceptable, given the difficulty he faced, and anyone else would have faced, in doing what he actually managed to do. Slote says 'the performance of the man who does everything but speak loudly may count as rationally acceptable, because of the great difficulty of doing (rationally) better than he has done. (It may be unreasonable, for example, to expect others to do as well as he has done.)' (p. 120) The point, then, which separates those less-than-the-most-rational actions which are, and those which are not, rationally 'acceptable' is the point at which the 'average person' would mostly likely give up.

There are two comments to be made in response. First, it is not clear that Slote has made the correct sense of our intuitions in this case,
for another interpretation can be given. We might say that in not
doing what he took to be supported by the strongest reasons, he did in
fact do something that was not rational. Nevertheless, we would want
to add, the degree of irrationality he exhibited was only minor, and
certainly much less than that of the average person, who would have
protested not at all to the boss. So it is correct to say that there might be
reason to admire this person, though not because he did something
totally rational (he didn't), but rather because he did something more
rational than anyone else was prepared to do. Thus, we can make sense
of the example other than the way Slote does.

The second comment to be made is that even if Slote's were the
correct interpretation, it would be irrelevant to the example with
which we started – the case of the accumulating Chateau Effete. What
amount of bottles separates those less-than-the-most-rational requests
which are, and those which are not, rationally acceptable? Perhaps the
quantity the 'average person' would request. But this will not do. First,
there might not be a specific amount that the average person would
request. And second, even if there were, it would be of no significance.
In the Intimidating Boss example, the significance of the employee's
doing more than the average person would have done is that this
displayed an admirable personal quality, such as courage, which his
workmates lacked. But picking a number larger than than the average
person would choose does not display any such personal quality – the
only thing it displays is that you happened to pick a number larger
than average.

Slote's suggestion, even if it could be sustained (which I believe it
cannot), would be of no help in the Chateau Effete case. It is also
implausible, then, to deny in the case in question the claim that if there
is more reason to do A than there is to do B, then the agent rationally
ought not to do B. Rejecting either the first or last premise of the
argument is implausible.

[4] Only the remaining, second, premise needs to be examined – that if
some action A has better consequences than another B then there is
more reason to A than to B. There are in fact two strategies one may
adopt with respect to this premise, though I am not sure that they are,
in the end, essentially different. One may deny the truth of the
premise, or may deny its applicability.
I consider first the strategy of denying the truth of the premise, and consider, in particular, the choice between having two bottles and having three. Suppose — for the sake of argument — it has better consequences to pick three than to pick two. I claim this does not mean there is more reason to pick three than to pick two. But how could this be?

To answer this question, consider, firstly, a finite choice between two and three bottles: these are the only options, and it is necessary to choose one or the other. The fact that you would enjoy having exactly two bottles is a reason for picking two bottles; the fact that you would pass up having exactly three bottles is a reason against it. The facts relevant to picking two bottles, and those with respect to picking three, are summarised thus:

(F2) You would enjoy having exactly two bottles; you would be passing up having exactly three bottles.
(F3) You would enjoy having exactly three bottles; you would be passing up having exactly two bottles.

Since the fact, (F3), you would enjoy three bottles and pass up only two, as a reason for picking three, is stronger than the fact, (F2), you would enjoy two bottles and pass up three, as a reason for picking two, we conclude there is more reason for you to pick three than there is to pick two, and so that you rationally ought to pick three bottles.

What, then, when you face an infinite choice, as in Slote's example? In this case the relevant facts about picking two, and picking three, can be summarised thus:

(F2') You would enjoy exactly two bottles; you would be passing up exactly three bottles, and you would be passing up exactly four, five, ..., googolplex, .... bottles.
(F3') You would enjoy exactly three bottles; you would be passing up exactly two bottles, and you would be passing up exactly four, five, ..., googolplex, .... bottles.

Obviously, in either case, you will be forgoing an awful lot of Chateau Effete. Indeed, there are no significant differences between the facts stated in (F2') and (F3'). True, there are differences (the bits before the 'and's), but they are not significant in the context of the infinite
amount of other considerations (the bits after the 'and's), which are, equally, reasons against picking two and picking three. Whether you pick two or three bottles, you will thereby be forgoing a great deal—an inconceivable amount—of Chateau Effete, so there is no justification for supposing the reasons for or against picking three are stronger than those for picking two.

It is clear where Slote has made his mistake. He says, correctly, that the reasons for taking three are stronger than those for taking two—it would be more enjoyable to take three. He concludes, incorrectly, that this means one rationally ought not to take two. He is wrong to infer this because whether one rationally ought to do something depends on both the reasons for, and the reasons against, that action; it depends on the benefits and the costs of performing that action. But, in the cases we are concerned, the reasons against taking three are equal in strength to those against taking two, given the infinite number of such considerations; the cost of taking three—the outcome the agent could have obtained in the best alternative—is precisely the same as the cost of taking two. The reasons for or against taking three are thus equal in strength to those for taking two; more generally, the reasons for or against taking n bottles are equal in strength to those for taking two, and so it is false one rationally ought not to take two bottles.¹⁵

Perhaps mindful of these sorts of considerations, the defender of the Self-Interest Theory will suggest that just as there is, in the infinite case, no more reason for picking three rather than two, so the consequences of picking three are no better than those of picking two, in light of the amount of Chateau Effete one would be forgoing with either decision. This, I believe, is the right thing for the Self-Interest Theorist to say about this example, but it means that the argument will still have failed to establish that rational dilemmas are possible. For if there is no difference in the consequences of picking three rather than two in the context of an infinite choice, then it is not possible to apply to the example the second premise—that if action A has better

¹⁵ This is the type of response Pollock also endorses, in 'How do you maximize expectation value?'. It is a view, though, with some surprising consequences. Suppose that instead of making your life infinite, and then offering to give you the number of bottles of Chateau Effete for which you ask, God does not make your life infinite, but offers to give you the number of happy days of life you ask for. How many should you ask for? If my argument is right, you have no more reason to ask for a millenia than you have to ask for day of happiness; and this seems surprising. I have Robert Sparrow to thank for drawing my attention to this point.
consequences than action B, then there is more reason to perform A than there is to perform B – and thus one will not have an argument to the desired conclusion.

I conclude that the second premise of the argument is either false or inapplicable, and thus that Slote has thus not been able to show rational dilemmas are possible, even in the infinitisic cases he draws on. Infinitisic cases such as this one cause problem’s for anyone’s intuitions, and though it seems strange, there is indeed no more reason in the case Slote describes to pick two bottles of Chateau Effete than there is to pick three, or pick a million.

§3 Rational Dilemmas, and Tragic Choices

The first example of Slote’s, then, fails to establish rational dilemmas are possible, but it is not the only type of example he provides. Slote also introduces a rationalistic formulation of a more popular (but less ingenious) argument for dilemmas. Unfortunately again for Slote, this argument fares no better than does its moralistic cousin.

[1] The example Slote is most convincingly introduced in parts. Here is the first part:

Consider a young man, living in the United States before the American entry into World War II, who has a strong interest in being a lawyer but an equally strong interest, persisting from student days, in archaeology. ... The man had been drafted, and his army unit is due to be transferred to Cambodia in a few months’ time. Having graduated from law school and passed the bar in his home state, he needs only to be sworn in as a member of the bar in order to be eligible for legal work when his unit is transferred abroad. But he chose this particular unit because he also wants to be near certain temples on which he would like to be able to do extended archaeological research. He needs to obtain permission from the Cambodian embassy in Washington in order to work at the temples ..., but he expects to be able to go to Washington to get the permission before his unit leaves for Cambodia. (p. 107)

What ought the young man to do? There is no problem answering this question. He rationally ought to see to it that he attends a swearing-in ceremony for the bar, since practising law is something enormously, centrally, important to him and to his conception of his own good. If something is that important, and he can see that it is achieved, then
the correct deliberative, (strongly) action-guiding and absolute judgement is that he ought to see to it. Similarly for the young man's getting his credentials checked. Archaeology is also something equally enormously, centrally, important to him and to his conception of his own good. The young man also rationally ought see to it that he gets his credentials checked. The two are in the relevant respects the same, and so the same judgements are required.

Suddenly the Japanese attack Pearl Harbour, and his unit's departure date is rescheduled for two days later. On the intervening day he has time to attend a swearing-in ceremony for admission to the bar or to visit Washington to obtain clearance for his archaeological work, but not to do both. So whatever he does, the results for him will be highly unsatisfactory. If he gets his credentials checked in Washington, ... [he] will for a number of years be restricted to peeling potatoes in the army, rather than developing his skills as a lawyer. What a waste! But if becomes a lawyer and misses out on the archaeological research, then he will have missed a unique opportunity ... [I]f he doesn't get a chance to work at the temples, he will rue and regret it the rest of his life. Again, what a waste! (p. 107)

Now, the argument goes, even though the young man cannot do both, it nevertheless remains true of each option that he rationally ought to pursue it. The second (and last) argument I will examine draws, like this example, on the phenomena of rational remainders (an explanation of which I shall introduce presently). This young man has separate obligations to pursue his own development. Nevertheless, it is argued, it would have been appropriate for the young man to feel distraught at the opportunity he will have to forgo.

[1.1] This type of argument has been much discussed in connection with the possibility of moral dilemmas.16 The tragedy of Agamemnon, it is claimed, shows that moral dilemmas are possible, for it would have been appropriate for him to feel guilty whatever he did, and such

---

guilt is appropriate only if he did something wrong. Agamemnon's guilt would be a moral remainder.

The notion of moral guilt, which is said to be applicable in Agamemnon's case, needs to be distinguished from that of moral regret. Both are similar, in being accompanied by similar moral phenomenology – the sense of anguish that typically (always?) attends guilt or regret. They differ, however, in their cognitive components. Feeling regretful about something you have done, on the one hand, necessarily includes believing that what you did was something you had moral reason not to do, whether or not you feel you had an excuse for inflicting such harm. Feeling guilty, on the other hand, includes the belief that what you did was morally wrong, and that you have no excuse for having done what you did. Guilt includes, where regret does not, the judgment that one has acted wrongly, against what is morally required.

If the argument from rational remainders is to be at all possible, then there must be rational remainders corresponding to the moral remainders of regret and guilt. I will assume, for the sake of argument, that such rational remainders exist. The rational analogue of moral regret I think can properly be termed rational regret. Feeling rationally regretful about something you have done necessarily includes believing you had reason to do otherwise, whether or not you felt you had an excuse for doing what you did. It would, however, not be appropriate to describe the rational analogue of moral guilt as rational guilt, for 'guilt' is a particularly moral term. I will instead, following Slote, call the relevant emotion (rational) rue. Feeling rueful includes the belief that what you did was irrational, and that you have no rational excuse for having done it. Rue includes, where rational regret does not, the judgment one has acted irrationally, against what is rationally required.

[1.2] Exactly what form, then, does this argument from moral and rational remainders take? Here is what Slote has to say about the fact that this young man cannot realise both of his dreams:

---


18 Note that 'rational regret' means 'regret pertaining to judgements of rationality', and not 'regret which is rational'.
The man himself may recognize all these difficulties and so realize that whatever he decides to do on the intervening day will result in an irreparable sacrifice of something enormously, centrally, important to him and to his conception of his own good. Whatever he chooses, it will be appropriate for him to have a profound and painful sense of dashed hopes ... His choice, whichever one he makes, will put a blight on a substantial portion of the rest of his life, and so afterwards he too can, and realistically must, feel he has done a terrible thing to himself. In him a sense of rue will be as appropriate and as realistic, as Sophie’s [famous of Sophie’s Choice] sense of guilt. (p. 108)

The interpretation I make of this particular argument, and of the argument from moral remainders in general, is as follows. The situation the young man faces, like Agamemnon faces, shows that:

(1) There are situations in which the agent in question has knowledge of all the relevant facts, and faces actions A and B such that (i) it would be rational for the agent to feel guilty (rueful) were he to fail to do A, (ii) it would be rational for the agent to feel guilty (rueful) were he to fail to do B, but (iii) he cannot do both.

It is, however, rational to feel guilty (rueful) only under certain sorts of conditions. In particular:

(2) If an agent has knowledge of all the relevant facts of the situation, then it would be rational for him to feel guilty (rueful) were he to fail to perform some action only if he morally (rationally) ought to perform that action.

---

19 See W. Sinnott-Armstrong, Moral Dilemmas, (Oxford: Basil Blackwall, 1988), pp. 44 ff. There are two further ways of interpreting this example. First, we might suppose the value of becoming a lawyer is not comparable with that of gaining accreditation, and that this explains why he ought to do each. This is the so-called ‘incommensurability’ argument, and is discussed, for example, by J. Raz, ‘Value Incommensurability: Some Preliminaries,’ Proc Aris Soc 86 (1985-6): 115-134 and W. Sinnott-Armstrong, ‘Moral Dilemmas are Incomparibility,’ Amer Phil Quart 22 (1985): 321-329. Second, we might suppose the value of becoming a lawyer is comparable, and furthermore equal to, that of gaining accreditation, and that this explains why he ought to do each. This is the so-called ‘symmetry’ argument, and is discussed by R. Marcus, ‘Moral Dilemmas and Consistency,’ J Phil 77 (1980): 121-136, D. Odegard, ‘Deep Moral Dilemmas,’ Theoria 53 (1987): 73-86, and F. Feldman, Doing the Best We Can, (Boston: Reidel, 1986), pp. 200 ff.
The conclusion to be drawn from these two premises is that there are situations in which a person morally (rationally) ought to do A, morally (rationally) ought to do B, but cannot do both, and thus that moral (rational) dilemmas are possible.

My major concern in this chapter is with rational, rather than with moral, dilemmas; the major concern of the literature on dilemmas has been moral, rather than rational, dilemmas. I shall assume - solely for the sake of argument - that the discussion of moral remainders are relevant to discussions of rational remainders. I shall want to deny the soundness of the above argument, in both its moral and rational forms, and will consider objections which have been put almost exclusively in moral terms. I shall assume - solely for the sake of argument - that the same type of objections apply in rational terms. Most of the discussion, then, will actually concern moral remainders. Needless to say, it could be rephrased in terms of rue and (rational) regret rather than guilt and (moral) regret.

The general strategy in replying to the argument from remainders is to claim that there are no remainders about which both premises are true. For some remainders, such as regret, the first premise is true, but the second false; for others, such as guilt and rue, the second premise is true but the first is false. The argument is thus unsound. Or so I shall argue.

[2] Consider the first premise of the argument: that there are situations in which the agent in question has knowledge of all the relevant facts, and faces actions A and B such that (i) it would be rational for the agent to feel guilty (rueful) were he to fail to do A, (ii) it would be rational for the agent to feel guilty (rueful) were he to fail to do B, yet (iii) he cannot do both.

I want to say that the premise is false as it stands, but becomes true when the emotion is regret rather than guilt. It is rational for Agamemnon to feel deep, even traumatic, (moral) regret at his having to kill his daughter, but not rational for him to feel guilty about it, in

---

20 As does Slote: 'I believe that the case for practical rational dilemmas is as strong as what can be said in favour of the possibility of moral dilemmas' (p. 100).
the sense that this would imply that it was rational for him to believe that he morally ought not to have done so. It is rational for the young man to feel deep, even traumatic, (rational) regret at his having to miss out on one of his life goals, but not rational for him to feel rueful about it, in the sense that this would imply that it was rational for him to believe that he rationally ought not to have done so. The account I favour thus finds a place for moral and rational emotion, but I dispute exactly which emotion is involved.

Some, however, are not convinced by this response, so I will spend the rest of this section defending it. The first objection I consider is that of Bas van Fraassen. He accuses those who make this standard response of treating cases like that of Agamemnon as if they counted as fundamentally the same as that of the philanthropist who regrets that he has but one fortune to give for mankind and has agonized over the choice between endowing the arts and furthering birth control. ... But such cases are the same only if regret is the same as guilt, or if it is necessarily appropriate to feel guilt if and only if it is appropriate to feel regret. And that can tenably be denied.22

In what way is the case of the philanthropist supposed to be different from that of Agamemnon? Perhaps that the philanthropist rationally regrets he has but one fortune to give to humanity, but would certainly be irrational to feel guilty about it, since the notion of regret is appropriate only when one faces a choice between great goods both of which one cannot achieve. By contrast, when the choice is between bads both of which one cannot avoid, the appropriate emotion is guilt rather than regret. Agamemnon does something bad, and so ought to feel guilty. The idea may be that failing to do something good calls for regret, while doing something bad calls for guilt (or rue).

Neither of these claims is in general true. Consider the first. A motor accident occurs just as I am driving by and it is obvious help is required, but I continue on my way. In ignoring the injured people, I have failed to do something good (after all, helping injured people is good, isn't it?), but it is guilt rather than just regret that is called for. Failing to do something good sometimes calls for guilt, rather than

mere regret. Consider the second claim. My child has something stuck in her throat and will choke to death unless I perform an immediate tracheotomy with the only knife, a blunt one, I have on hand. Amidst the screams of the child, this is what I bring myself to do. In cutting the child’s throat with a blunt knife I have done something bad (after all, why does my child look in horror as I approach her with knife in hand?), but it is regret rather than guilt that is called for. Doing something bad sometimes calls only for regret, rather than guilt as well.

The second objection I consider derives from Ruth Marcus, and is to the effect that to say of Agamemnon that he should only feel regret is to underestimate the strength of the appropriate emotion, and is to grossly misrepresent the facts. She has a different case in mind, though:

To insist that “regret” is appropriate rather than “guilt” or “remorse” is false to the facts. It seems inappropriate, for example, to describe as “regret” the common feelings of guilt that women have in cases of abortion even where they believe (perhaps mistakenly) that there was moral justification in such an undertaking.23

The magnitude of the situation, Marcus insists, is not adequately captured by merely attaching the label of ‘regret’ to it. Similarly, we may suppose, she would think Agamemnon ought to feel distraught at killing his daughter, and such an emotion is inappropriate if he only ought to feel regret at doing so. The idea is that in Agamemnon’s case a violent emotion is appropriate, and that regret is just not such an emotion, while guilt is.

A preliminary point in response. It seems consistent with the above passage that Marcus thinks that the woman may very well have been morally justified in having her abortion. (She says the woman might believe this, though ‘perhaps mistakenly’. But if perhaps mistakenly, then perhaps correctly?) If so, the relevant emotion is not what I have been describing as guilt. For, as I made clear above, the notion of guilt should be understood in such a way that it conceptually includes, in some fashion or other, a judgement to the effect the action performed was not morally justified. We can agree with Marcus that

23 R. Marcus, ‘Moral Dilemmas and Consistency,’ J Phil (1980), p. 133, fn. 11. This footnote is all she has to say to address the sort of response I have made.
the emotions appropriate to the situation she describes are powerful ones indeed, and if we feel the word 'regret' does not capture such intensity, then perhaps we will need to use another. One word we cannot use, though, is the word 'guilt', at least not if it retains the meaning I indicated above.

Suppose, though, Marcus would insist that the strength of the appropriate emotion indicates it is guilt (as I have understood the term) which is appropriate, rather than mere regret. This still seems incorrect. Early in life you made a career choice having many considerations in favour of it, but, due to a great deal of bad luck and through no fault of your own, it has led to a miserable and wasted life. Looking back on your life, you come to regret bitterly the choice you actually made. Such regret is surely much more intense than guilt one feels at, say, taking more than one's share of dessert. These points about the phenomenology of various moral emotions do not touch my notions of guilt or regret as I have defined them. For the way in which such emotions differ is not in their felt quality, but rather in the beliefs appropriate for each. And this Marcus (and van Fraassen) do not address.

Further, there are a vast range of emotions it is perfectly consistent with my response to claim it is rational to have. It would be more than appropriate for Agamemnon to feel distraught, anguish, horror, grief, and so on, at his killing his daughter. Just so long as it is not the case that it is rational to have the emotion \textit{vis a vis} some action only if you rationally ought not to have performed that action. This presumably leaves quite a large range of emotion available to Agamemnon.

[3] The first premise is true only of regret, and not of guilt – there are indeed situations in which it is rational to be regretful whatever you do. If the argument is to be valid, then, the second premise must be modified, as follows, to reflect this fact: if an agent has knowledge of all the relevant facts of the situation, then it would be rational for him to feel regret only if he morally (rationally) ought to perform that action.

There has been some recent discussion of the issue of the reasoning proper for a moral, or for that matter, rational agent.\textsuperscript{24} Contrast two types of Samaritan, both of whom give money to the beggar they pass. The reason the first gives money is that the beggar needs help and the

money would provide some. The reason the second gives money is that this is what he morally ought to do. What directly moves the first person are the facts about the beggar (namely, that the beggar needs the money), but what directly moves the second are facts about his own duties (namely, that he ought to help the beggar). It is commented, in these discussions about the proper reasoning for a moral agent, that the first Samaritan seems preferable to the morally self-conscious second Samaritan, because the second seems to have a disproportionate regard for his own moral goodness, and this is what makes him seem a priggish and an unattractive fellow. The reasons of a moral agent need not, and it seems preferably do not, include facts about his own moral obligations.

A similar point underlies my denial that it is rational to feel regret about having done some action only if you morally ought not to have done it. Consider two reasons Agamemnon might have had for feeling bitter regret for having killed Iphigenia.25 His reason might, on the one hand, be the fact that he has killed his innocent daughter, whom he loved dearly. Or his reason might have been, on the other hand, that he was breaking a solemn duty of care towards his daughter. In the first case, Agamemnon is motivated by facts about Iphigenia (namely, that she is his daughter, that he loves her, that she is innocent) whereas in the second case he is motivated by facts about his own obligations. It needs commenting upon that the first type of motivation seems preferable to the second, since the second again seems to involve too high a regard for the agent's own moral goodness. Since the fact Iphigenia is his daughter is a reason for Agamemnon to feel regret about killing her, a fact consistent with his nevertheless being morally permitted to do so, then it is indeed rational for Agamemnon to feel such regret, even if he was morally permitted to have done so.26

25 M. Nussbaum, in 'Aeschylus and Practical Conflict', Ethics 95 (1985): 233-267, points out that Agamemnon, in Aeschylus's play of the same name, not only kills his daughter, but does so with completely the wrong attitude. He emotionally treats her killing as if it were the slaughter of a goat. Nussbaum claims this emotional reaction is immoral: Agamemnon morally ought not to react in this way. I agree. Nussbaum claims further, though, that the immorality of this reaction shows the action itself was wrong: Agamemnon morally ought not to have killed his daughter. I disagree, and cannot find in this paper any definite argument for this further claim. It seems to me that Agamemnon's reaction is more than enough to explain why we find him morally reprehensible.

26 This is a common response. See H. Steiner, 'Moral Conflicts and Prescriptivism,' Mind 91 (1973), pp. 586-7, T. C. McConnell, 'Moral Dilemmas and Consistency in Ethics,' Can J Phil 8 (1978): 269-87, E. Conece, 'Against Moral Dilemmas,' Phil Rev
There are, nevertheless, a number of objections to this idea. The first one I consider comes from Bernard Williams, who claims that it is artificial to separate these two different types of motivation, and to claim one can be present but the other not. Williams puts this response in terms of a distinction between natural motivations — having regrets because one has had to do something distressing or appalling or which in some way goes against the grain (what I have called simply 'regret') — and moral motivations — having regrets because one thinks that one has done something one ought not to have done (what I have called simply 'guilt'). Williams has two points to make in response to this. First:

The sharp distinction that this argument demands between these natural and moral motivations is unrealistic. Are we really to think that if a man (a) thinks that he ought not to cause needless suffering and (b) is distressed by the fact or the prospect of his causing needless suffering, then (a) and (b) are just two separate facts about him? Surely (b) can be one expression of (a), and (a) one root of (b)?

Williams asks how Agamemnon could think he had killed Iphigenia without also thinking he had done something he ought not.

It is an easy question to answer. Just because there are two very closely related ways of describing the same act — as one of killing his beloved daughter, or as one of acting against an obligation to safeguard his daughter — it does not follow that to be motivated by one is to be motivated by the other. It does not follow that if Agamemnon's emotions are motivated by his belief in the innocence of his daughter, then they must also be motivated by a thought he ought not to kill his daughter. And if he can be motivated by the first consideration alone, then, presumably, he could be motivated by it while thinking it false that he morally ought not to kill his daughter. It makes perfect sense, then, to say Agamemnon was motivated in his regret by the fact he killed his beloved daughter, while realising at the same time, but was


not thereby motivated by, the fact he was acting against his paternal obligation.

Williams's second criticism of the suggestion it is regret rather than guilt which is appropriate in the cases we are considering is to claim that such a suggestion will not, in any case, work in all situations:

A man may, for instance, feel regret because he has broken a promise in the course of acting (as he sincerely supposes) for the best; and his regret at having broken the promise must surely arise via a moral thought. Here we seem just to get back to the claim that such regret in such circumstances would be irrational...28

If one breaks a promise one ought to feel regret at having done so, but the only reason for feeling regret, so the criticism goes, is that you did something which you morally ought not to have done.

This, however, is not the only possible reason one might have for regret at breaking a promise. This case is no more convincing than the previous one, concerning the reason Agamemnon has for feeling regret. The fact that he broke his promise is also a reason for our man to feel regret, just as the fact that Agamemnon killed his daughter is a reason for him to feel regret. Indeed, we can make the following speculation concerning the connection between reasons for action and reasons for regret: if the fact (if it is a fact) that \( p \) is a (strong) reason for an agent not to perform action A, then the fact (if it is a fact) that \( p \) and they did A is a reason for them to regret having done A.29 The fact Iphigenia is his daughter is a strong reason not to kill her. Because of this, the fact that he killed his daughter Iphigenia is a reason for Agamemnon to feel regret at having done so. The fact that he made a promise is a strong reason (let us suppose) for our man to keep it. Because of this, the fact he made a promise he didn’t keep is a reason for him to feel regret about not keeping it.

This second argument for moral (rational) dilemmas is a failure. On the one hand, there are never any situations in which, through no fault of one’s own, one rationally ought to feel guilty whatever one does, though there may be many in which one rationally ought to feel regret whatever one does. On the other hand, the fact that it might be

28 B. Williams, 'Ethical Consistency', p. 113.
rational to regret having done something does not entail that you morally (rationally) ought not to have done it, and so even if one rationally ought to feel regret whatever one does, this does not mean one has acted irrationally. For some remainders, such as regret, the first premise is true, but the second false; for others, such as guilt and rue, the second premise is true but the first is false. The argument is thus unsound.

Conclusion

Slote has not shown that rational dilemmas are possible, fortunately for my arguments in the first two parts of the thesis, for were they possible (OPn) and (OCn) would be false. But what, then, of Agamemnon, who faced a choice, in a situation not of his own making, between fulfilling his obligation to Zeus and leading his men to Troy, and fulfilling his parental obligation not to kill his daughter? Such situations are disturbing, both for the agents involved, and for us spectators who know what they are going through. Slote thinks that such examples, and others, show moral and rational dilemmas are possible: but he is wrong. Does this mean, then, that we are wrong to find these situations disturbing? Of course not. There is enough reason to be disturbed when a father kills his innocent daughter without supposing he acted immorally or irrationally.
No treatment of the Self-Interest Theory and rational irrationality would be complete without a discussion of Derek Parfit's work in the first part of his *Reasons and Persons*. Parfit argues that an adequate theory of rationality can imply it is rational to cause oneself to act irrationally, but were this so, then (as we shall see) principles (OCn) and (OPn) would be false. As a result, I will examine his arguments for the possibility of rational irrationality. We will see that they do not establish this possibility, and therefore do not threaten principles (OCn) and (OPn).

§1 Rational Irrationality

Parfit thinks the Self-Interest Theory, S, implies we cannot avoid acting irrationally (p. 13). He tries to show as much in Section 6 of *Reasons and Persons* ('How S implies that we cannot avoid acting irrationally'), and this is where I will begin my discussion. He thinks this because of someone he calls Kate.

---

1 All references to Parfit's work will be to *Reasons and Persons*, (Oxford: Oxford Univ. Pr., 1984).


3 Parfit's formulation of the Self-Interest Theory employs two important notions. First, rational action is identified in terms of what the agent has *most reason* to do. I shall assume, what is in any case implicit in Parfit's discussion (pp. 8, 16), that this is also to be identified with what the agent *rationally ought* to do. Second, the rational action is characterised in terms what is *best* for the agent in question. Parfit canvasses three theories of what is best for an agent, but concentrates only on one: the Hedonistic Theory, which says that the action which is best for an agent is the one which *gives the agent most happiness*. I shall assume, as does Parfit (p. 4), that the discussion can be altered to suit any other theory of what is best for an agent.
[1] Kate is a writer (pp. 6-7, 14). Because of her will to succeed, she has made it that she desires most that her books be as good as possible. Because of this desire, she writes quite good books and finds her work very rewarding, though she also often ends up working very hard for very long periods of time. As a result of this, she collapses with exhaustion and becomes, for a while, very depressed. Still, if this desire were not so strong, then while she would not end up in such bouts of exhaustion and depression, she would also find her work rather boring. Since rewarding work together with occasional depression is better for her than boring work without, Kate rationally ought to have made it that she desires most that her books be as good as possible. Even so, Kate rationally ought not to work so hard. As we have seen, because she works so hard, she writes good books and finds it rewarding, though she occasionally ends up in deep depression. But if she didn’t work quite so hard, then while her books would not be quite as good, she would still find her work as rewarding, and also not suffer from those bouts of depression. Since only slightly worse books without depression is better for her than the best possible books with bouts of depression, Kate ought not to work so hard. Thus Kate rationally ought to make it that she has a desire which causes her later to do things she rationally ought not to do.

After referring back to this original description (pp. 6-7) of Kate, and rehearsing it (p. 14), Parfit launches, in Section 6, into a discussion of whether or not Kate can in fact have the desires that are best for her—including the desire her books be as good as possible—without sometimes freely choosing to act in ways—such as overworking—which will be worse for her. There are two cases, Parfit believes. On the one hand, it might be that Psychological Determinism is true (pp. 14-5). On this view, our acts are always caused by our desires and dispositions. Given our actual desires and dispositions, it is not causally possible we act differently. If Psychological Determinism is true, then Kate cannot have the desires that are best for her without sometimes freely choosing to act in ways which will be worse for her. On the other hand, Psychological Determinism might be false (pp. 15-6). If this is the case, then Kate can have the relevant desire without overworking, but, Parfit says, an objector to the Self-Interest Theory must also admit it would be very hard for Kate to have the desire without overworking.

Parfit then says (p. 16) that Kate might add, in her own defence, that ‘[i]t 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 were 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’ (p. 16). We will meet this argument again later in the chapter, where some of its terms will receive more extensive definition — for the moment I am concerned merely with exegesis. Parfit calls a person never self-denying if they never do what they believe will be worse for them (p. 6), and what Parfit calls the ‘relevant’ sense of impossible to do A is ‘doing A would have been impossible, even if my desires and dispositions had been different’ (p. 15).

Parfit thinks, then, that if Psychological Determinism is true, Kate cannot have the desire that her books be as good as possible without overworking, and, whether or not Psychological Determinism is true, she cannot have one of the best possible sets of motives and never do what she believes to be irrational. He recognises that some might see in this an objection to the Self-Interest Theory:

It may now be said that, as described by Kate, S lacks on [sic] one of the central 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. (p. 16)

This is the type of objection I have been concerned to put to the Self-Interest Theory in this thesis. Yet to put this objection it is not necessary to assume Psychological Determinism or to endorse Kate’s argument in the previous paragraph.4 If Psychological Determinism is

4 It is important for my purposes not to have to rely on Parfit’s claims concerning Kate’s inability. This is because (a) I myself take no stand concerning the truth of
true, or if Kate's argument is sound, then it follows, as a general law, that Kate cannot have the desires and dispositions she has, including the desire that her books be as good as possible, and not work so hard. We need, for our discussion, though, to make no such general claim, but only a specific speculation about Kate. Unlike other persons, we may suppose, she lacks a certain capability: desiring most that her books be as good as possible without thereby overworking. Not necessarily that this is a psychological law of nature, or necessarily because Kate's argument in the previous paragraph is valid. Rather, this is a just humble fact about Kate. So, whether or not Parfit's specific diagnosis of Kate's inability is correct, we may make the weaker assumption, for the rest of the chapter, that Kate cannot both have the best desire and yet not overwork. Parfit and I agree at the very least on this much.

[2] Parfit claims, in response to those who would put this objection, that it can be rational to cause oneself to act irrationally. He provides a number of arguments for this claim (which we will come to presently), and so he thinks it is not a good objection to S that Kate cannot always avoid doing what S claims to be irrational (p. 16).

If Parfit is right, and it actually can be rational to cause oneself to act irrationally, then principles (OC_n) and (OP_n) are false. To see why, concentrate on Kate. Parfit's arguments, if they are correct, show that Kate rationally ought to adopt one of the best set of motives (including the desire that her books be as good as possible) which she cannot have without freely and intentionally doing something irrational (such as overworking). Hence, if what Parfit says is right, it is possible that one

---

5 Psychological Determinism, and (b) the argument quoted above which Parfit puts into Kate's mouth is invalid, as I argue in the Appendix.

As much as Parfit ever says about Kate is that S implies it is rational for Kate to adopt a desire which will cause her to act irrationally. He does not specifically say that in fact it is rational for Kate to adopt a desire which will cause her to act irrationally. Still, Parfit thinks there are such cases (pp. 12-3), and it will simplify matters if I assume – solely for the sake of argument – that Kate is one such case.

6 It is clear from this example, furthermore, that neither of these rational obligations is a second-best rational obligation: she rationally ought to adopt one of the best set of motives, not because she has done (or will do, or is doing) something irrational, but simply because this is the only way she can ensure that she does not find her work boring; she rationally ought not to overwork, not because she has done (or will do, or is doing) something irrational, but simply because if she worked just a little less, her books would still be quite good, and she would no
rationally ought to A (for example, adopt the desire that one's books be as good as possible), rationally ought to B (for example, not overwork), even though one cannot both A and B. If his arguments are correct, then, ($OP_2$) and ($OC_2$) are false, and so by extension are ($OP_n$) and ($OC_n$). This fact explains the attention I will give in this chapter to Parfit's arguments for the possibility of rational irrationality. From this point onward, then, I will have no occasion to refer to principles ($OP_n$) and ($OC_n$), but will focus simply on the possibility of rational irrationality.

The onus of proof is, I believe, on someone like Parfit who would defend this possibility. Parfit, as we have seen, thinks it is not a good objection to S that Kate cannot always avoid doing what S claims to be irrational. To show this, he provides three arguments. I will spend the remainder of this chapter, and the Appendix, examining each in turn.

§2 Could it be Rational to Cause Oneself to Act Irrationally?

Immediately after introducing the objection above, employing Kant's dictum, Parfit directs the reader's attention to the previous section in his book, Section 5 ('Could it be Rational to Cause Oneself to Act Irrationally?') in which he claims to show that an acceptable theory of rationality can imply we cannot avoid acting irrationally. I will argue that his discussion in this section does not show this.

[1] Since the example provided in Section 5 is Parfit's major positive argument for this claim, I will quote it in length:

*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 than 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 for me 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.

longer have bouts of depression. Therefore, principles ($OC'_n$) and ($OP'_n$) will be just as applicable to Parfit's discussion as are ($OC_n$) and ($OP_n$).
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 this 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*. I also have a special drug, conveniently at hand. This drug causes one to be, for a brief period, very irrational. I reach for the bottle and drink a mouthful before the man can stop me. 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, the man is less likely to believe that I would record the number on his car. He therefore has less reason to kill me.

While I am in this state, I shall act in ways that are very irrational. 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 risk that this man will kill us all. (pp. 12-13)

Any plausible theory, Parfit claims, would say it would be rational for me, in this case, to cause myself to become for a brief period very irrational. He says *Schelling's Answer* shows an acceptable theory about rationality that can tell us to cause ourselves to do what, in its own terms, is irrational. If it is to show this, however, there must be (a) some action described in the example which any acceptable theory would tell us to do, but which would cause later irrationality, and (b) some action described in this example, caused by that earlier action, which is irrational in terms of that theory.

(a) In *Schelling's Answer*, what action would any acceptable theory tell me to do? The answer is plain: take the drug. For if I do not, then it is very likely that the robber will kill all of us; and if I do, then it is very likely he will not kill all of us, though there is a slight risk I will hurt either myself or my children. If I take the drug, I will behave in such a

---

7 My discussion of this particular example was greatly aided by conversations with Frank Jackson and Peter Menzies.
way as to give the robber reason to believe that threats and torture will do him no good, and that I will not tell the police the number of his car. I will behave in such a way that I might hurt either myself or my children. But since this is much better than having all of us dead, I plainly ought to risk it, and take the drug. Clearly, Parfit intends his example to be interpreted such that taking the drug is what any acceptable theory would tell me to do.

(b) In Schelling's Answer, what action, caused by my taking the drug, is irrational in terms of any acceptable theory? Within seconds, Parfit says, it becomes apparent that I am crazy, and this in a number of different ways (p. 13). Reeling about the room, I say to the robber 'Go ahead. I love my children. So please kill them'; he tries to get the gold by torturing me, however I cry out: 'This is agony. So please go on'. Furthermore, while I am in this state, and after the robber has fled (his threats and torture being ineffective), there is a risk that, before the police arrive, I may harm myself or my children. Plainly, Parfit intends his example to be interpreted such that some, if not all, of these actions, caused by my taking of the drug, are ones any acceptable theory would say is irrational.

I shall now argue that, if the type of behaviour described in the previous paragraph - ignoring the man's threats to my children and the torture he inflicts on me, and harming myself or my children - can indeed be interpreted as free and intentional action, then Parfit has not shown that they are actions which any acceptable theory would say is irrational. Of course, if it is not free and intentional action, then Schelling's Answer fails to address the case of Kate who, it will be recalled, is a person who adopted desires which cause her, quite voluntarily (p. 14), and intentionally, to overwork. I shall provide two arguments for my conclusion.

[2] The first argument for my conclusion is concerned with precisely what the effects of the drug are supposed to be. There are three broad possibilities: the drug might make me good at faking irrationality; the drug may alter my values or my beliefs (I examine this in [2.1]); and, finally, the drug may alter the relation between my values and beliefs, on the one hand, and my actions, on the other (I examine this in [2.2]). I shall examine each of these in turn.
(i) First, perhaps the drug makes me good at faking irrationality. As the man is holding the gun to my child's head, I am quaking inside, terrified he might shoot her, but (due to the effects of the drug) none of this shows, and I say, looking for all the world like a madman: 'Go ahead. I love my children. So please kill them'. As the man is turning the thumbscrews just that little bit more, I am suffering terrible pain, and wish he would stop, but (due to the effects of the drug) none of this shows, and I say, again looking for all the world like a madman: 'This is agony. So please go on'. Seeing my behaviour (but not knowing my real values or beliefs), the man becomes convinced I am crazy, and, deciding to cut his losses, he leaves.

If the drug makes me very good at faking irrationality, then Parfit has not shown that an acceptable theory would say I was irrational to act as I did. Certainly, before I had taken the drug, to act in this way might have been irrational, as Parfit correctly points out (p. 12), for the robber would most likely have killed one of my children to make me believe that, unless he gets the gold, he will kill my other children. Before I had taken the drug, my attempts at faking irrationality would have led to disaster. But after I have taken the drug, I am not doing anything irrational when I do this. The drug makes me a very good actor: if I were to act in this way, the robber would become convinced I was crazy, and would leave; if I were not to act in this way, then I would have to hand over my gold, since he would remain convinced that threats and torture would get the gold. Clearly, if this is the effect of the drug, then it is much better to act in the way I did, and any adequate theory would say this is precisely the way I should act.

[2.1] Second, perhaps the drug does not make me very good at faking irrationality, but, instead, radically alters my values or my beliefs.

(ii) On the one hand, perhaps the drug radically alters my beliefs. It gives me, for a brief time, strange beliefs about how I can best ensure the safety of my children, and that the torture stops. These strange beliefs are to the effect that the best way of ensuring my children's safety is to say 'Go ahead. I love my children. So please kill them', and that the best way of ensuring that the torture stops is to say 'This is

---

8 It is clear from the way he describes the example, that this interpretation is not what Parfit intends, and I consider it only for the sake of completeness. It is not clear, however, precisely what interpretation Parfit has in mind, so I will consider a number of possibilities.
agony. So please go on. It is not that I believe I will be able to say these things in a way to convince the man I am crazy, because (we are now supposing) it is not part of the effect of the drug that I receive this ability. Perhaps I think he will take pity on me if I display my desperation by making such ridiculous statements. No matter why I have this belief, though, the drug’s effects are only brief, and so it will likely not lead to a bad outcome.

If the drug radically alters my beliefs in this way, then Parfit has not shown that any acceptable theory would say I was irrational to act as I did. For if I really do think that speaking as I do will promote my children’s safety, and will promote the torture’s stopping, then in saying ‘Go ahead. I love my children. So please kill them’ and ‘This is agony. So please go on’ I am acting in a way my (rather strange) beliefs lead me to expect will promote my (reasonable) values for the safety of my children and the cessation of my pain. Given that the drug alters my beliefs in the way I have suggested, it seems I am acting rationally in pursuing these values in the way I am. It will be straight away objected, of course, that these beliefs clearly might be irrational, and so any action taken on the basis of them (however effectively it may take them into account), is therefore also irrational. The drug may alter my beliefs in this radical way, but these are irrational beliefs, and so I do act irrationally – and any adequate theory would say so – when I speak as I do. Saying ‘Go ahead. I love my children. So please kill them’ and ‘This is agony. So please go on’ will not ensure that my family remains unharmed, even if I really believe it will.

If the drug radically alters my beliefs in the way suggested, then there are two things we might say. On the one hand, we might agree that such beliefs are indeed irrational, but insist that the rationality of an action is to be determined by its relation to the beliefs one actually has, rather than the beliefs it would be reasonable for one to have. This is the view of those I called subjectivists (with respect to expectations), who we met briefly in Chapter One. On the other hand, we might insist that not only are these beliefs irrational, but also that their irrationality infects the actions based upon them. This is the view of those I called rationalised subjectivists (with respect to expectations).

---

and also who we met briefly in Chapter One. If this is the correct interpretation of the effect of the drug, then the example in and of itself provides no reason for picking the second of these alternatives rather than the other. Even if we agree that the beliefs the drug induces are irrational, the example Parfit gives does not, by itself, determine that the actions based on such a belief are also irrational. If the drug radically alters my beliefs in the way suggested, then Parfit’s example does not show that any acceptable theory would say I was irrational to act as I did.

(iii) Alternatively, maybe the drug radically alters my values. It makes me, for a brief time, a bitter person who greatly disvalues loving relationships, so I speak sincerely when I say: ‘Go ahead. I love my children. So please kill them’. It also makes me, for a brief time, a masochist who greatly values painful experiences, so I again speak sincerely when I say: ‘This is agony. So please go on’. The drug’s effects are only quite brief, though, and will likely not lead to bad outcomes.

If the drug radically alters my values in this way, then Parfit has not shown that any acceptable theory would say I was irrational to act as I did. For if I really do think that the fact I love my children is a reason for the man to kill them, and if I really do think that the fact the man’s torture is painful is a reason him to continue, and if I believe I can realise these ends by speaking as I do, then in saying ‘Go ahead. I love my children. So please kill them’ and ‘This is agony. So please go on’ I am acting in a way I (reasonably) expect will promote my (rather strange) values. Given that the drug alters my values in the way I have suggested, it seems I am acting rationally in pursuing these values in the way I do. It will be straight away objected, of course, that these values are clearly irrational, and so any action taken on their basis (however effective it may be in promoting them), is therefore also irrational. The drug may alter my values in this radical way, but these are irrational values, and so I do act irrationally – and any adequate theory would say so – when I speak as I do. Saying ‘Go ahead. I love my children. So please kill them’ and ‘This is agony. So please go on’ is to promote deeply irrational values, even if they are values I now hold.

However, for the same type of reason as the previous case, I conclude that if the drug radically alters my values in the way suggested, then Parfit has not shown that any acceptable theory would say I was irrational to act as I did.

[2.2] Third, perhaps the drug does not make me very good at faking irrationality, and changes neither my values nor my beliefs – so that I still value my children and disvalue my own pain, do not now believe that the best way of ensuring my children's safety is to say 'Go ahead. I love my children. So please kill them', and that the best way of ensuring my torture stops is to say 'This is agony. So please go on' – but, instead, alters the relation between my values and beliefs, on the one hand, and my behaviour, on the other.

(iv) In particular, perhaps the drug makes me act randomly, so that there is no systematic relation between my beliefs and values, on the one hand, and my actions, on the other. I love my children, and do not want them harmed, and believe by acting as I do I only endanger their lives, yet I ask the man to kill them. I detest pain, wish it would stop, and realise that by acting as I do I only endanger myself, yet I ask the man to continue torturing me. My actions are simply random in relation to my prior values and beliefs. In my view, this is most likely the interpretation of Parfit's intention. Within a few seconds of taking the drug, he says, it becomes apparent I am 'crazy' (p. 13). I reel about the room, and say things which seem in no way related to my beliefs or values. I am in no state, furthermore, to record the number of his car. The image Parfit seems to be providing us is of a person completely out of control.

If the effect of the drug is to make me act randomly, then none of the behaviour Parfit describes can plausibly be interpreted as free and intentional action. No doubt others would describe me as very irrational when I behave in the way I do, but the issue is whether what I do can be interpreted as free and intentional action. So very irrational and out of control am I, it seems what I do cannot be interpreted in this way. On the one hand, an action is free only if I could have done otherwise; but it seems I could not have done otherwise. I am totally under the control of this drug, and so, arguably, my behaviour does not constitute free action. On the other hand, my behaviour A can be interpreted as an intentional action only if two conditions are satisfied: (a) I take myself to have some reason R for me to A; and (b) I do A, and
do it because I take R to be a reason for me to A. Clearly, though, if the
drug causes my actions to be random, then my behaviour could not be
interpreted as intentional action. First, condition (a) is likely not to be
satisfied, given that my beliefs and values remain as they were before I
took the drug. Second, and more importantly, condition (b) will not be
satisfied, given that my actions are random in relation to my beliefs
and values, and so to my reasons. As a consequence, even if I did take
myself to have reason R for doing what I ended up doing, A, I could
hardly have done A because I took myself to have reason R to do it. My
actions are random in relation to my reasons.

Now, to summarise this first argument for the conclusion that, if
the type of behaviour described above – ignoring the man's threats to
my children and the torture he inflicts on me, and harming myself or
my children – can be interpreted as free and intentional action, then
Parfit has not shown that they are actions which any acceptable theory
would say is irrational. There are four effects the drug might have: it
might (i) make me good at faking irrationality, (ii) radically change my
beliefs, (iii) radically change my values, or (iv) make my actions
random in relation to my beliefs and values. If, on the one hand, the
drug has effect (iv), Parfit has not displayed actions which can be
interpreted as free and intentional. If, on the other hand, the drug has
effects (i), (ii), or (iii), Parfit has not shown that they are actions which
any acceptable theory would say is irrational. Hence, either Parfit has
not displayed actions which can be interpreted as free and intentional,
or he has not shown that they are actions which any acceptable theory
would say is irrational. Therefore, if the type of behaviour described
above can be interpreted as free and intentional action then he has not
shown they are actions any acceptable theory would say is irrational.

[3] Now for the second argument the conclusion that, if the type of
behaviour described above – ignoring the man's threats to my children
and the torture he inflicts on me, and harming myself or my children
– can be interpreted as free and intentional action, then Parfit has not
shown they are actions which any acceptable theory would say is
irrational. I shall first [3.1] describe another form Parfit's argument
might take, and then [3.2] critically examine it.

[3.1] To get the conclusion he wants, recall, Parfit has to show there is
(a) some action described in the example which any acceptable theory
would tell us to do, but which would cause later irrationality, and (b) some action described in this example, caused by that earlier action, and which is irrational in terms of that theory. (a) In Schelling's Answer, what action would any acceptable theory tell me to do? Clearly, Parfit intends his example to be interpreted in such a way that taking the answer is: taking the drug. Any acceptable theory would say this because taking the drug is what I expect will best promote the values I have. (b) In Schelling's Answer, what action, caused by my taking the drug, is irrational in terms of any acceptable theory? Clearly, Parfit intends his example to be interpreted such that some, if not all, of the actions above, caused by my taking of the drug, are ones any acceptable theory would say is irrational. I argued in the previous subsection, [2.1], that if these are free and intentional actions, then (cases (i), (ii) and (iii)) they are also actions I expect will best promote the values I have, and so Parfit has not shown they are irrational in terms of any acceptable theory.

However, let us suppose – solely for the sake of argument – that Schelling’s Answer is indeed a situation in which what I expect would have the best outcome for me is to cause myself (by taking the drug) to freely and intentionally perform an action I expect would not have the best outcome for me. Suppose, in particular, this free and intentional action caused by my taking the drug is my harming myself or my children. Let us, suppose, that is, the situation is as follows:

(SA(D)) Schelling’s Answer (Diachronic case). A man breaks into my house. He hears me calling the police, etc. etc. I have a special drug, conveniently at hand. This drug has some very strange effects. I cannot take the drug without behaving in a very strange way, so strange that it gives the man reason to think that threats and torture will do him no good, and that I will not tell the police the number-plate of his car. If I do behave in this way, the man would flee. But I cannot take the drug without at some time later (after he has fled) quite freely and intentionally harming myself or my children. Still, it is much better that I harm myself or my children than that we all be dead.

Any acceptable theory, so the argument goes, would say I ought to take the drug, but would also say I ought not harm myself or my children, even though I cannot take the drug without, later, harming myself or
my children. Any acceptable theory would say I ought not to harm myself or my children since, clearly, this will achieve nothing.

[3.2] Even this reconstructed argument will not show that if harming myself or my children can indeed be interpreted as free and intentional action, they are actions any acceptable theory would say is irrational. To see this, consider the following example:

(SC(S)) Sophie's Children (Synchronic case). Sophie has two children, a boy and a girl, who have been swept away by the rising river. She knows if she doesn't now save the girl, she will drown, and if she doesn't now save the boy, he will drown. She also knows that, while she can now save the girl, and can now save the boy, she cannot now save both.

It might seem to some that Sophie faces a dilemma: that she rationally ought to save the boy, and rationally ought to save the girl. Others, like myself and Parfit (p. 506), believe otherwise. Why do we believe otherwise? The answer is plain: it is false Sophie rationally ought to save each child because, quite simply, she cannot save both.

Consider now a second case involving Sophie, where the relevant actions occur not at the same time (as previously, in (SC(S))), but one after the other:

(SC(D)) Sophie's Children (Diachronic case), What Sophie knows is if she doesn't now save the girl, she will drown, and if she doesn't later save the boy, he will drown. She also knows that, while she can now save the girl, and can later save the boy, she cannot both save the girl now and save the boy later.

Again we should say it is false Sophie rationally ought to save each child. For even though the action of saving her son would occur later than would that of saving her daughter, it nonetheless remains a fact she cannot save both. In the original example we saw it was because she could not save both that she was excused from saving both. Again we should say it is false Sophie rationally ought to save each child because, quite simply, again Sophie cannot save both. The point of these two examples is that if we change only the time of one of the
actions involved in a conflict, then what we have to say about whether or not the agent faces a dilemma should not be different.

So it is with Schelling's Answer. In the situation described above, my taking the drug now would cause me, freely and intentionally, to harm myself or my children later. This is case (SA(D)), above. But suppose instead we have the following case:

(SA(S)) Schelling's Answer (Synchronic case). This is the same as the previous case, except that (for some strange reason we need not go into) I cannot take the drug now without at the same time harming myself or my children. The drug will have its desired effects only if I take it while I am doing this. It still remains true, of course, that it is much better that I harm myself or my children than that we all be dead.

The only difference between this synchronic version of Schelling's Answer, (SA(S)), and the previous diachronic version, (SA(D)), is the time at which I would harm myself or my children. In this synchronic case, it would be at the same time; in the previous diachronic case, this would be at a time later than that of my taking the drug. Hence, what we have to say about whether or not I face a dilemma should not be different in either of the two cases. But it is obvious that, in this second synchronic case, (SA(S)), I do not face a dilemma. It is obvious I rationally ought to take the drug, and am rationally permitted to harm myself or my children while I do so. Therefore, in the previous diachronic case, (SA(D)), I also do not face a rational dilemma. It should be just as obvious that I rationally ought to take the drug, and am rationally permitted to harm myself or my children after I have done so. In taking the drug, I did what I rationally ought to have done, and in harming myself or my children I did what I was rationally permitted to do.

One might suggest there is a disanalogy between the two pairs of cases and that this vitiates my argument. In Sophie's Children, the considerations for the conflicting actions balance – there is just as much reason to give up the boy as there is the girl; in Schelling's Answer, however, they do not balance – there is more reason to take the pill than there is not to harm my children. My argument, however, depended on the analogy between the Sophie's Children cases and the
Schelling's Answer cases, and since the cases are in this way disanalogous, the argument is defective.

In response, I admit the cases are in the indicated way disanalogous, but deny that this means my argument is defective. The pairs of cases are different in that in one pair the conflicting considerations balance, while in the other they do not. But this is not a relevant difference. What is relevant, and what my argument is based upon, is the fact that in both pairs of cases, one cannot perform both of the conflicting actions. In both of the Sophie's Children cases, Sophie cannot save both of her children, and this is why we say it is false she ought to save each. (The balanced considerations in this case are a reason to believe that Sophie ought to save at least one of her children, but is rationally permitted to give up her girl, and is rationally permitted to give up her boy.) In both of the Schelling's Answer cases, I cannot take the pill without harming myself or my children, and this is why we should also say it is false that I both ought to take the pill but ought not harm myself or my children. (The unbalanced considerations in this case are a reason to suppose that I ought to take the pill, and am rationally permitted to harm myself or my children.)

Parfit is wrong, then, to claim that Schelling's Answer shows an acceptable theory about rationality that can tell us to cause ourselves to do what, in its own terms, is irrational. For all that has been said to this point, it is indeed a good objection to the Self-Interest Theory that Kate cannot always avoid doing what it claims to be irrational.

§3 Could it be Impossible to Avoid Acting Wrongly?

Parfit thinks (and I agree) that theory S implies Kate cannot avoid acting irrationally. He recognises that some may see in this implication an objection to theory S, and he asked us to consider Schelling's Answer. After considering this example and its relevance to the case of Kate, Parfit says: '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' (p. 17). Since I believe Parfit's claims concerning Schelling's Answer do not fully answer the objection, I will examine what Parfit has to say in Section 15 ('Could it be Impossible to Avoid Acting Wrongly?'). I will argue his discussion in this section still does not show this.
The focus in this section is not the Self-Interest theory of rationality, S, but on what Parfit calls the Consequentialist theory of morality, C. This theory claims that: (C2) What each of us ought to do is whatever would make the outcome best (p. 24). Parfit thinks both the Self-Interest theory and Consequentialism are subject to the same type of objection. His objection to the truth, in Kate's case, of the dictum that 'ought' implies 'can' goes as follows:

In most cases, when someone acts wrongly, he deserves to be blamed, and should feel remorse. This is what is most plausible in the doctrine that ought implies can. It is hard to believe that there could be cases where, whatever someone does, or might have earlier done, he deserves to be blamed, and should feel remorse. It is hard to believe that it could be impossible for someone to avoid acting in a way that deserves to be blamed. C does not imply this belief. If I saved my child rather than several strangers, I will believe that I am doing what will make the outcome much worse. I would therefore believe that I am acting wrongly. But this would be a case of blameless wrongdoing. According to C, we can always avoid doing what deserves to be blamed. This is enough to satisfy the doctrine that ought implies can. (p. 36)

In this passage Parfit has the following situation in mind. It is morally best for Clare, as he calls her, to cause herself to desire most the well-being of her children (just as it is best for Kate that she desires her books be as good as possible), morally best for her to save several strangers in difficulty rather than saving her own child in difficulty (just as it is best for Kate not to overwork), even though, were she most to desire the well-being of her children, she would save her child rather than the strangers (just as were Kate to desire most her books be as good as possible, then she would overwork).

Parfit's argument, when applied to the case of Kate, must be like this. There are two interpretations of Kant's dictum. These are:

(O) An agent rationally ought to A only if they can A.
(P) An agent would be rationally worthy not to A only if they can A.

Parfit seems to be saying (O) is false, and only appears true because we confuse it with (P), which is true, but consistent with S and C. It is exactly Kate's case, and Claire's, which show this. For Kate rationally ought both to desire most that her books be as good as possible and not
work so hard, even though (in discord with (O)) she cannot. But it is the very fact she cannot which explains (in accord with (P)) why she would not be rationally blameworthy not to have this desire without overworking.

[2] This explanation will not do, however, since someone who believes the Self-Interest theory should not, in consistency, also be committed to a principle such as (P). Parfit is wrong to say that 'According to C [and so, presumably, according to S], we can always avoid doing what deserves to be blamed' (p. 36, my emphasis). The reason is simple. Such a person thinks they ought to perform some action if and only if it has the best outcome to perform that action. Since blaming someone is itself an action, such a person should think that they ought to blame an action when it has the best outcome to blame it. That is, the defender of the Self-Interest Theory, like all consequentialists, should think that an action is blameworthy when it has the best outcome to blame it. And this, of course, need not necessarily be if the action in question was itself wrong (for it might have the best outcome to blame an action which has the best outcome – in cases, for example, of moral luck), or even if the person in question could not have done otherwise (for it might have the best outcome to blame a person’s action even when they could not have done otherwise – as in cases, for example, of strict liability).

The supporter of S cannot suppose, consistently with their own position, that interpretation (P) is true, and so cannot use it in an attempt to explain why the standard interpretation of Kant’s dictum – (O) – only appears true in Kate’s case.

Parfit’s attempt to supplant the standard interpretation of Kant’s dictum with one of his own – and one more congenial to the position he advocates – is a failure. After considering Clare’s plight, and again referring to his discussion of Schelling’s Answer, Parfit suggests further that ‘[t]o meet the objection to C, Clare might appeal to other cases where we cannot avoid acting wrongly. That there are such cases has been claimed by some of the writers who are most opposed to C. I discuss this answer in endnote 14.’ (p. 37). Since I believe more needs to be said to meet this objection, I will examine what Parfit has to say in Endnote 14. But just as Parfit relegates his discussion to a four-page-long endnote, I relegate my further discussion of Parfit to an Appendix. The conclusion of the Appendix, though, is much the same as that of
the previous two sections. For all that Parfit says, it remains a good objection to S that Kate cannot always avoid doing what S claims to be irrational.

Conclusion

Parfit thinks that an adequate theory of rationality can tell us to cause ourselves to do something which is, in its own terms, irrational, but he is wrong. This is fortunate for my argument in the first two parts of the thesis, for were they possible (OPn) and (OCn) would be false. He thinks this sort of rational irrationality is possible because of the case of Kate. What we know about Kate is that were she to desire most that her books be as good as possible, then she would quite freely and intentionally work so hard that she would occasionally suffer from exhaustion and deep depression – she cannot both have this desire and not work so hard. What we know about the Self-Interest theory is that it says she rationally ought to make it she has this desire, and that she rationally ought not to work so hard. Since by hypothesis she cannot do both, the Self-Interest Theory implies Kate cannot avoid acting irrationally. Contrary to what Parfit says, this is indeed an objection to the theory, and no reason to suppose principles (OPn) and (OCn) are false.
Chapter Nine

Gregory Kavka and the Paradox of Deterrence

The aim of part II of the thesis was to argue that it could be rational to cooperate in the Prisoner's Dilemma. I addressed three objections to Gauthier's attempt, introduced in Chapter Four, also to argue for this conclusion. The last objection claimed not only that Gauthier failed to justify the move from the rationality of dispositions and intentions to that of actions, but introduced an example which seemed to show that such an inference could not be justified at all. It seems that any attempt such as Gauthier's - and, presumably, such as mine - to forge a link between the rationality of dispositions and the rationality of actions is doomed to failure. The example in question was the Paradox of Deterrence. As we saw, Kavka and others think it shows it could be rational to be disposed to performing irrational retaliatory actions, and thus that it could also be rational to be disposed to perform irrational cooperative actions. Gauthier's response to this paradox I examined, and dismissed as incomplete; my own response to the paradox I did not introduce at all. The purpose of this chapter is to rectify this omission, and to argue that, given certain conditions (to be specified below), the actions resulting from rational deterrent intentions are also rational, even if one is free to do otherwise and it has the best outcome to do otherwise (though these actions may very well be irrational, absent those conditions).

§1 The Self-Interest Theory and Rational Deterrence

The objectors to Gauthier's position - and, implicitly, to mine - would deny such a claim, and insist that if the actions resulting from expected-value maximising, and so presumably rational, deterrent intentions are not themselves expected-value maximising, then those actions are irrational. They are committed to the claim, then, that it could be rational to intend to perform an irrational retaliatory action.
[1] You and I are still in the state of nature. I would like to have some way of getting you to cooperate unilaterally, by casting off your own right to use force while I retain mine. Unfortunately for me, there is rough equality between us, and I cannot force you to do so. One day, however, I discover a weapon of amazing destructive potential—the Bomb. So powerful is this weapon that, even were I to trigger it from a distance, doing so would result not only in your death, but in mine as well. Even though you are aware that I now have this weapon, you realise it would do me no good to detonate the Bomb—I would only kill myself. You do not cooperate, I have conclusive evidence to believe you have not, and I believe you have not. Should I trigger the Bomb as a result of your non-cooperation, or not?

This is not the world’s most difficult decision problem. I rationally ought not to trigger the Bomb: if I trigger the Bomb, the outcome will be that both of us will die; if I do not, then nothing untoward will happen, and we will remain in the state of nature. The state of nature is a bad state-of-affairs indeed, but not as bad as my death. It should be painfully obvious that I rationally ought not to trigger the Bomb. We can justify this evaluation on the basis of the following two claims:

(S) If an agent is free to perform an action A, then they rationally ought to A if and only if the agent-relative expected-value of doing A exceeds that of doing any alternative to A.

(D1) You are free to make an independent choice between the actions of cooperating or not, you do not cooperate, and I have conclusive evidence to believe you have not cooperated. I am now free to trigger the Bomb in response to your non-cooperation or not, and the expected value of not doing so greatly exceeds that of doing so. This is common knowledge.

The first assumption, (S), is of course the standard formulation of the Self-Interest Theory. The second assumption, (D1), is a summary of the central features of situation I have just described. Since it will be important, I will describe it in a little more detail.

The assumption states that you face an initial independent choice between the actions of cooperating or not, and, after you have made your decision, I am free to trigger my Bomb in response to your non-cooperation (if you do not cooperate). It will save words if, henceforth, we say that I retaliate (and denote this by ‘R’) if I trigger the Bomb in
response to your non-cooperation. In short, the assumption says that you face an initial choice between cooperating or not, that you do not cooperate, and that as a result I then face a choice between retaliating or not. The possible outcomes of these choices, and the values each of us attaches to them, may be depicted as follows:

\[
\begin{align*}
\text{You} & \quad \text{Me} \\
\sim C & \quad R \\
C & \quad \sim R
\end{align*}
\]

Doom for Both (h,h)  
War of all against all (d,d)  
I exploit You (t,s)

The values I assign to the possible outcomes are listed first, and are: my exploiting you (=t, Temptation payoff); your non-cooperation and my non-retaliation, leaving us were we started in the state of nature and the war of all against all (=d, Defection payoff); and, my retaliating and doom for both (=h, Holocaust payoff). Clearly, \( t > d > h \), since I value most the outcome of my exploiting you, second the status quo, and a distant third my death. The values you assign to the possible outcomes are listed second, and, in addition, include: my exploiting you (=s, Sucker payoff). Clearly, \( d > s > h \), since you value most the war of all against all, next your servitude to me, and a distant third your death.

This little example is a paradigm of the application and plausibility of theory (S). It seems perfectly clear I rationally ought not to retaliate, since this will bring only my death. It seems that even with such a weapon in my hands, I cannot get you to cooperate unilaterally.

[2] However, suppose I could convincingly threaten to use the Bomb if you were not more obliging. I could say to you 'I intend that if you do not cooperate, then I will trigger the Bomb as a result (and that if you do, I won't'). And suppose that such a threat is necessary and very

---

1 This intention needs to be distinguished from two other internal states: (a) my being such that if you do not cooperate, then I intend to retaliate. These are different since, in the text, I have an intention whether or not you retaliate; but in this first case, I may not have an intention if you do in fact cooperate; (b) my being disposed such that if you do not cooperate then I will retaliate. These are different since it is consistent with the text that I believe you have not cooperated but do not retaliate (I change my mind), but not with the second. See G. Kavka,
likely to be sufficient to ensure your cooperation: I must intend to
trigger the Bomb if you are to be induced to cooperating; and the
intention to retaliate if you have not cooperated must very likely get
you to cooperate. Ought I to adopt the intention to retaliate or not?

This decision problem is less straightforward than the previous
one. I realise, as do you, that the situation has changed, and may now
be represented as follows:

\[
\begin{align*}
&\text{State of Nature (d,d)} \\
&\quad \left\{ \\
&\quad \quad \text{Me} \\
&\quad \quad \quad \begin{cases}
&\text{Me} \\
&\text{You}
\end{cases}
\quad \left\{ \\
&\quad \quad \quad \text{I} \\
&\quad \quad \quad \quad \begin{cases}
&\text{I} \\
&\text{You}
\end{cases}
\quad \left\{ \\
&\quad \quad \quad \quad \quad \begin{cases}
&\text{\sim I} \\
&\text{\sim C (\mu)}
\end{cases}
\quad \left\{ \\
&\quad \quad \quad \text{\sim R} \\
&\quad \quad \text{I dominate You (t,s)}
\end{cases}
\quad \left\{ \\
&\quad \quad \text{\sim C (1-\mu)}
\end{cases}
\quad \left\{ \\
&\quad \quad \text{Doom for Both (h,h)}
\end{cases}
\end{cases}
\end{align*}
\]

I face a choice between adopting the deterrent intention (I) or not. If a
certain condition – to be introduced below – obtains, it turns out that I
rationally ought to adopt the intention to retaliate.

To see this, consider the outcomes of adopting this intention or
not. On the one hand, adopting the deterrent intention is necessary if
you are to cooperate: if I were not to adopt the intention, then we
would remain in the state of nature. Since the value to me of the state
of nature by d, then the expected-value for me of not adopting the
deterrent intention is d. On the other hand, adopting the intention is
very likely to be sufficient for you to cooperate: if I were to adopt the
intention then there would be a very small chance – which I shall
denote by \( \mu \) – that you will not cooperate, I will retaliate, and death for
both would result, but a very high chance – (1-\( \mu \)) – that you will

‘Deterrent Intentions and Retaliatory Actions’, in D. MacLean (ed.), The Security
Gamble: Deterrence Dilemmas in a Nuclear Age, (Totowa: Rowman & Allanheld,
1984), p. 155, and ‘A Paradox of Deterrence Revisited’, in his Moral Paradoxes of
Nuclear deterrence, (New York: Cambridge Univ. Pr., 1987), p. 36, for an
interpretation of the conditional intention the same as that in the text. Note,
furthermore, that the conditional intention needs also to include the claim that if
you do cooperate, then I will not retaliate.
cooperate, and that I will be able to exploit you. Since the value to me of my doom by \( h \), and that of my exploiting you by \( t \), then the expected-value to me of adopting the deterrent intention is \( \mu h + (1-\mu) t \).

Given a certain condition, it now follows that I rationally ought to adopt the intention. This is because, on the Self-Interest Theory’s view of the matter, I rationally ought to adopt it if and only if the expected value of doing so is greater than of remaining an SM — that is, if and only if \( \mu h + (1-\mu) t > d \). Rearranging terms, we see that I ought to adopt the intention if and only if

\[
\frac{t-d}{t-h} < \mu
\]

I assumed above that if I were to adopt the deterrent intention, then it would be very unlikely – \( \mu \) – that you will not cooperate. Assume, in particular, that \( \mu \) is less than \( (t-d)/(t-h) \), as condition (1) requires. It follows I rationally ought to adopt the deterrent intention.

In assumption (D1), above, we supposed that you did not cooperate, and that it had the best outcome for me not to retaliate as a result. Let us suppose further, that this non-cooperation on your part was in fact preceded by a threat that I would retaliate if you did not cooperate, and that such a threat was necessary and very likely sufficient to get you to cooperate. Let us suppose further, then, that

(D2) (a) I am free initially to adopt the intention to retaliate; the expected value of my adopting the intention is substantially greater than that of not adopting it, since it is necessary and very likely sufficient to adopt the intention if you are to cooperate. (b) If I were to adopt the intention, you would be free later to choose between cooperating or not. As a matter of fact, however, you do not cooperate, and I have conclusive reason to believe so. I am later still free to retaliate, and the expected value of not doing so greatly exceeds that of doing so. This is all common knowledge.

I simply add clause (a) to assumption (D1) to get this new assumption, (D2). It may seem inconsistent to suppose, in clause (a), that my adopting the retaliation intention would very likely result in your cooperation, and, in clause (b), that you do not as a matter of fact
cooperate. But it is not. The fact it is very likely you will be deterred does not entail that you will be deterred; hence, it is possible that it is very likely you will be deterred – (a) – but that, unfortunately – (b) – you are not. Though logically possible, it is, of course, very unlikely. I will call the sorts of situations described by assumption (D2) Unhappy Deterrence Situations, or UDSs for short.

[3] Unhappy Deterrence Situations are closely related to what Gregory Kavka calls Special Deterrence Situations, or an SDSs for short. This is how he defines them:

an agent is in an SDS when he reasonably and correctly believes that the following conditions hold. First, it is likely he must intend (conditionally) to apply a harmful sanction to innocent people, if an extremely harmful and unjust offence is to be prevented. Second, such an intention would very likely deter the offence. Third the amounts of harm involved in the offence and the threatened sanctions are very large, and the relevant probabilities and amounts of harm are such that a rational utilitarian evaluation would substantially favour having the intention. Finally, he would have conclusive moral reasons not to apply the sanctions if the offence were to occur.2

This definition generalises a situation in which it is likely that a nation must conditionally intend to launch a retaliatory nuclear strike if a first strike against it is to be deterred, that such a policy would very likely deter the first strike and would, on utilitarian moral grounds, be the required policy, even though, were the nation the object of a first strike, there would be conclusive moral reasons for them not to launch a retaliatory salvo of missiles. Kavka’s definition and mine are closely related, though not exactly the same, for where Kavka’s concerns are moral, mine are rational.3 Consider each of the four conditions in turn.

---

2 G. Kavka, ‘Some Paradoxes of Deterrence’, in his Moral Paradoxes of Nuclear Deterrence, (New York: Cambridge Univ. Pr., 1987), p. 17. This is a reprinted version of an article in J Phil. Since it includes some changes from the earlier article (some in this passage), I shall concentrate on the later version.

First, in order for an agent to be in an SDS, Kavka says it is likely he must intend (conditionally) to apply a harmful sanction to innocent people, if an extremely harmful and unjust offence is to be prevented. In the context of the little story I introduced above, this corresponds to the claim that in order for me to be in a UDS, I must intend (conditionally) to apply a sanction harmful to myself (namely, to trigger the Bomb) if an offence which is harmful to me (namely, your not cooperating) is to be prevented. Note, however, that Kavka requires it to be ‘likely’ to be necessary, while I say it ‘must’ be necessary. This simplifies the discussion, and (if anything) strengthens Kavka’s case for rational irrationality.

Second, in order for an agent to be in an SDS, Kavka says the intention to apply the sanction would very likely deter the offence. In the context of the little story I introduced above, this corresponds to the claim that in order for me to be in a UDS, the intention to retaliate if I believe that you have not cooperated must very likely get you to cooperate. It is worth noting in passing that this condition excludes the possibility you yourself face a decision about how likely you are to cooperate were I to intend to retaliate. If you do face such a decision, then it cannot be assumed, as the definition of an SDS requires, that if I adopt the retaliation intention then you will very likely cooperate. For if you do face such a decision, then whether or not such a conditional will obtain is exactly what you would be deciding upon. And if you are initially rational, you may decide to become a Threat-Ignorer, and not be such that if I adopt the deterrent intention then you will cooperate, as this would dissuade me from making any threats against you.

Third, in order for an agent to be in an SDS, Kavka says the amount of harm involved in the offence and the threatened sanctions are very large, and that the relevant probabilities and amounts of harm are such that a rational utilitarian evaluation would substantially favour having the intention. In the context of the little story I introduced above, this corresponds to the claim that in order for me to be in a UDS, the expected value of my adopting the intention is substantially greater than that of not doing so.

Fourth, in order for an agent to be in an SDS, Kavka says he would have conclusive moral reasons not to apply the sanctions if the offence were to occur. In the context of the little story I introduced above, this corresponds to the claim that in order for me to be in a UDS, the expected-value of not retaliating is greater than that of retaliating, even
if you were not to cooperate. The final condition as I understand it importantly differs, however, from the way Kavka does. He would suppose, in the rationality case, that the agent in question would have conclusive (rational) reasons against retaliating, or, in short, that they rationally ought not to retaliate. But to provide such a condition as part of a characterisation of SDSs is to beg the question against those such as Gauthier and myself who are concerned to argue that if it is rational to adopt the conditional intention (or disposition) then it might be rational to act upon it. A non-question-begging final condition would be that the agent in question values the outcome of not retaliating to that of retaliating. According to the Self-Interest Theory, of course, this just means that the agent rationally ought not to retaliate, but since the issue is whether retaliation is sometimes rational, a condition such as the one Kavka introduces should not be part of the definition of an SDS.

Note finally an important difference between Kavka’s SDSs and my UDSs. It is no part of the definition of an SDS that the retaliation intention has failed; yet it is an important part of the definition of a UDS that it has. As we shall see, the sorts of Special Deterrence Situations which cause most problem for Gauthier’s position and my position are the ones where the deterrent intention has failed. The sorts of situations of most concern are Unhappy Deterrence Situations.

In Unhappy Deterrence Situations, it seems that I rationally ought to adopt the intention that if you do not cooperate then I will retaliate, rationally ought to believe you have not cooperated, even though I rationally ought not to retaliate. Toxin Puzzle Cases, we saw in Chapter Two, also show it could indeed be rational to adopt an intention to perform and irrational action – but in my view UDSs are the more convincing. Why? Acting on the toxin-drinking intention may not be expected-value maximising, but – at least – you end up in a situation better than if you had not adopted the intention in the first place: acting on a deterrent intention is also not expected-value maximising,

4 Kavka’s form of the fourth assumption unfortunately seems to play a prominent role in the arguments of ‘Some Paradoxes of Deterrence’ (esp. pp. 290, 292). His discussion in ‘A Paradox of Deterrence Revisited’, however, drops this assumption. See p. 36, condition (E).

5 This follows from the fact that, after you have not cooperated, I have conclusive evidence to believe that you have not done so, and the theory (B) of rational belief, which I introduced in Chapter Four. I will make scant reference to this theory in the remainder of this chapter.
and - worse still - you end up in a situation much worse than if you had not adopted the intention in the first place. Acting on a deterrent intention seems clearly much more irrational than acting on a toxin-drinking intention.

§2 The Paradox of Deterrence

It seems paradoxical that it could be rational to adopt the intention to perform an irrational action. Yet, as we have seen, this is precisely what the Self-Interest Theory again entails. And it seems the appropriate conclusion is that rational intentions to cooperate are also intentions to perform irrational actions. Yet this is not so – in neither Gauthier’s case nor my own is this the appropriate conclusion to draw from Unhappy Deterrence Situations.

[1] But first some revision. Gauthier’s suggestion, as we saw at the end of Chapter Four, is that we should evaluate the rationality of an action, not in terms of its consequences, as we have already done, but rather in terms of the rationality of the intention which has it as an object. In the present case, this means that

(B2) If it is rational for me to adopt an intention to do $x$ in circumstances $c$, and if $c$ come about, and if nothing relevant to the adoption of the intention is changed save what must be changed with the coming about of $c$. . . , then it is rational for me to carry out $x$.6

The rationality of an action and the intention to perform that action, Gauthier feels, stand or fall together. On this view, the rational agent is the one who takes the big picture in their aim to fulfil their values:

The fully rational actor is not the one who assesses her actions from now but, rather, the one who subjects the largest, rather than the smallest, segments of her activity to primary rational scrutiny, proceeding from policies to

---

performances, letting assessment of the latter be ruled by assessment of the former.7

It is the largest segments of a rational agent’s activities – her plans, intentions, and policies – which are for Gauthier the primary objects of rational evaluation, and the smaller segments – particular actions falling under these plans, intentions and policies – which derive their evaluation from them.

As we have seen, there are some who would simply deny the rationality of intentions and the rationality of actions are in this way connected. Even though Kavka, for example, admits that ‘there may be something to’ this wider segments view, he believes that

our normal view of rationality also implies being prepared to change previously formulated plans or intentions when there are significant stakes involved and relevant new information about outcomes is available. This is precisely the situation that arises when deterrence fails in a SDS. There is much harm to be done by retaliation, and the benefit that motivated the formation of the intention to retaliate – prevention of the offence – is now unobtainable.8

For Kavka the fact that one’s deterrence has failed is more than enough reason to reconsider one’s policy. If it is rational to intend nuclear retaliation, then, in the unlikely event that such deterrence fails, we need not admit it is therefore rational to retaliate.

[2] This much I argued in Chapter Four. What I did not point out there is that Gauthier can respond to this objection (though whether he would take up the suggestion I make is not an issue with which I shall deal). The response has two parts. First, Gauthier can agree with Kavka’s claim concerning ‘our normal view of rationality’ that it would imply that retaliation in Special Deterrence Situations would not be rational – for simplicity, I shall concentrate on SDSs rather than UDSs. Second, though, he may go on to claim that such a view of rationality would nevertheless still imply that cooperation in the Prisoner’s Dilemma could be rational.

7 Gauthier, ‘Deterrence, Maximization, and Rationality,’ p. 488.
The first part of this response has Gauthier agreeing with Kavka's claims concerning 'our normal view of rationality'. Kavka thinks that there is something to the view that the rationality of actions is in some way determined by the rationality of intentions and plans having those actions as objects. Kavka thinks there is nothing to the view that this relationship is captured by Gauthier's (B2), above. He says, instead, that 'our normal view of rationality also implies being prepared to change previously formulated plans or intentions when there are significant stakes involved and relevant new information about outcome is available.' Unfortunately, Kavka says nothing explicit about what our normal view of rationality implies when there are not both significant stakes involved and relevant new information about outcome is available. He seems to suggest, and Gauthier could make the speculation, that:

(B3) If it is rational for me to adopt an intention to do x in circumstances c, and if c comes about, and if either no significant stakes are involved or no relevant new information about the outcome is available, then it is rational for me to carry out x.

Thus, Gauthier could now agree that if it is initially rational to adopt a policy to A when p, and if p, then it is rational to A, unless there are significant stakes involved and it is clear that the policy cannot now do the job for which it was adopted.

The second part of the response has Gauthier claiming that, even with such an admission, this view of rationality would nevertheless still imply that cooperation can be rational. Consider what we might call Special Cooperation Situations (or SCSs). If such situations are to be characterised in a way corresponding to Kavka's characterisation of Special Deterrence Situations, they need satisfy four conditions. First, that I must intend (conditionally) to do the cooperative thing (namely, lay down my arms) if your cooperation is to be ensured. Second, that the intention to cooperate, if I believe that you would, must very likely get you to cooperate. Third, that the expected-value of my adopting this (conditional) cooperative intention, is greater than that of not doing so.

---

9 SCSs are almost identical to the Sequential Prisoner’s Dilemmas, which we met briefly in Chapter Four. As can be seen, the SCS corresponds more closely to the Toxin and Deterrence Cases; I concentrated, however, on the simultaneous PD in Part II because these are the most discussed.
And finally, that the expected value of not cooperating is greater than that of cooperating, even if you were to cooperate in the first place.

Special Deterrence Situations and Special Cooperative Situations are, in many respects, very similar. To get you to cooperate in an SDS, it is necessary and very likely sufficient for me to adopt the (conditional) retaliation intention; to get you to cooperate in an SCS, it is necessary and very likely sufficient for me to adopt the (conditional) cooperation intention. Thus: it is rational for me in an SDS to adopt the (conditional) retaliation intention; it is rational for me in an SCS to adopt the (conditional) cooperation intention. However, retaliating in an SDS would not be expected-value maximising; cooperating in a SCS would not be expected-value maximising.

Special Deterrence Situations and Special Cooperative Situations, though, are in one respect crucially different. To see this, concentrate on principle (B3).

On the one hand, in an SDS, I am considering whether to adopt an intention to retaliate (=‘x’) in the circumstances that you fail to cooperate first (=‘c’). This means that in an SDS, if the relevant circumstance (namely, that you fail to cooperate) comes about, then I will know that things have not turned out as I expected they would (since, recall, my adopting the retaliation intention is supposed to make it very likely that you will cooperate), and so the stakes are high and relevant new information about the outcome of adopting the intention is available. This means we may not infer from (B3) that it is rational to retaliate in an SDS – we can agree with Kavka that our normal view of rationality implies that when things are going exactly contrary to what we thought they would when we devised our plans, then it is indeed irrational to act on those plans.

On the other hand, in an SCS, I am considering whether to adopt an intention to cooperate (=‘x’) in the circumstances that you cooperate first (=‘c’). This means that in an SCS, if the relevant circumstance (namely, that you cooperate) comes about, then I will know that things have turned out as I expected (since, recall, my adopting the cooperation intention is supposed to make it very likely you will cooperate), and no relevant new information about the outcome of adopting the intention is available. This means we may infer from (B3) it is rational to cooperate in an SCS – Gauthier can insist that our normal view of rationality also implies that when things are going
exactly as we thought they would when we devised our plans, then it is indeed rational to act on those plans.

Kavka’s objection, then, is a blessing in disguise. By admitting the essential point of the objection, Gauthier can show why he may claim that it is rational to cooperate (a claim he dearly wants to be able to make), without thereby committing himself to the rationality of retaliation (a commitment with which he seems ill at ease\textsuperscript{10}). While SDSs may indeed show that deterrent intentions are intentions to perform irrational actions, Gauthier – if he is careful – can deny that this implies cooperative intentions are also intentions to perform irrational actions.

[3] It seems Gauthier has better luck than others defending some form of a bridging principle.\textsuperscript{11} However, in the course of this thesis, I have not been concerned myself to defend bridging principles, but rather to defend the following deontic principles:

\[(\text{OP}_n)\] If an agent rationally ought to \(S_1\), rationally ought to \(S_2\), ..., and rationally ought to \(S_n\), then it is logically possible that he \(S_1\)'s, he \(S_2\)'s, ..., and he \(S_n\)'s. \((n = 1, 2, 3, ...)\)

\[(\text{OC}_n)\] If an agent rationally ought to \(S_1\), rationally ought to \(S_2\), ..., and rationally ought to \(S_n\), then he can be such that he \(S_1\)'s, he \(S_2\)'s, ..., and he \(S_n\)'s. \((n = 1, 2, 3, ...)\)

The question arises: do Special Deterrence Situations where deterrence has failed – that is, Unhappy Deterrence Situations – pose a threat to these deontic principles, as they posed a threat to Gauthier’s original bridging principle \((B2)\)?

In short, the answer is no. Every UDS seems to involve only three rational obligations: that I rationally ought to adopt the retaliation

\textsuperscript{10} That keeping one's agreements and carrying out one's threats may be given parallel rational support may seem a very mixed blessing. It seems, therefore, only fair to acknowledge that this essay ['Deterrence, Maximization and Rationality'] in particular represents work still in progress, ... one of my current projects focuses on the differences, rather than the similarities, between the rationale for keeping agreements and the rationale for carrying out threats. Here then I should stress the tentativeness of my defense of the rationality of deterrence,' – from the introduction of D. Gauthier, \textit{Moral Dealing: Contract, Ethics and Reason}, (Ithaca: Cornell Univ. Pr., 1990).

\textsuperscript{11} But it only seems. Later we will see that the correct theory of reconsideration implies there are some cases where one ought \textit{not} to reconsider one's retaliation intention, even though things are not as one expected them to be.
intention; that I rationally ought to believe you have failed to cooperate; and, that I rationally ought not to retaliate. Hence, if a UDS is to be an objection to these two deontic principles then it must be an objection, in particular, to \((OC_3)\) or \((OP_3)\). On the one hand, it is an objection to \((OC_3)\) only if, in addition, I cannot adopt the intention to retaliate and believe you have not cooperated without retaliating. Clearly, though, it is no part of the definition of a UDS that this be so – I might for example be able to reconsider my intention to retaliate – and hence no UDS is an objection as such to \((OC_3)\). On the other hand, it is an objection to \((OP_3)\) only if it is not logically possible I adopt the intention to retaliate and believe you have failed to cooperate without retaliating. Even more clearly, it is no part of the definition of a UDS that this be so, and hence no UDS is an objection as such to \((OP_3)\). So, as defined, no UDS is an objection to principles \((OC_n)\) or \((OP_n)\).

Kavka's objection, then, entirely passes me by. While UDSs may indeed show that deterrent intentions are intentions to perform irrational actions, I can – consistently with my belief in \((OC_n)\) and \((OP_n)\) – deny that this implies cooperative intentions are also intentions to perform irrational actions.

The situation Kavka describes, and the one I base upon it, leads us, it seems, to the conclusion that I rationally ought to adopt the intention to perform an irrational action. Kavka is inclined to see these sorts of examples as refutations, in the first instance, of the claim that if it is rational to adopt a (conditional) deterrent intention then it is rational to carry it out. Since these cases do not touch principles \((OC_n)\) and \((OP_n)\), I can agree with Kavka, and, if Gauthier endorses principle \((B3)\) rather than his original \((B2)\), then he may as well. Kavka is also inclined to see this sort of example more generally as a refutation of the claim that if it is rational to adopt a (conditional) cooperative intention, then, in some cases, it is rational to carry it out. With this I do not necessarily agree, and Gauthier, if he were to adopt the strategy I have been discussing, would not agree.

§3 Rational Retaliation

Nevertheless, putative counterexamples to principles \((OC_n)\) and \((OP_n)\) do lurk within UDSs. In this section, I will introduce two, and respond to them in detail. I shall argue that given certain conditions (to be specified below), the actions resulting from rational deterrent
intentions are also rational, even if one is free to do otherwise and it has the best outcome for one to do otherwise (though these actions very well may be irrational absent these conditions).

[1] The Fourth Counterexample. Imagine, again, the situation described by assumption (D2) – the Unhappy Deterrence Situation – and add to its details the following:

While contemplating that, if my threat is successful, I will not need to retaliate in order to get you to cooperate, I realise that I cannot believe you have not cooperated without fulfilling my deterrent intention. I was brought up to be such that if I intend to do something when \( p \), and I believe that \( p \), then I do it. Furthermore, I realise, I cannot but be like this. Not only will I not, but I also cannot be such that I intend to do something when \( p \), believe that \( p \), and then fail to do that thing. I conclude, to my chagrin, that I cannot have the retaliation intention, believe you have failed to cooperate, and yet not retaliate.

Not that this is a grand logical truth about the inevitable relation between intention and unhindered action, but just a humble truth about my own makeup.

This is a special sort of UDS. The assumption we need to make if it is indeed to address principle (OC_n) is that situations of the following form are logically possible:

\[(D3a) \text{ (a) I am free initially to adopt the intention to retaliate; the expected value of my adopting the intention is greater than that of not adopting it, since it is necessary and very likely sufficient to adopt the intention if you are to cooperate. (b)} \text{ If I were to adopt the intention, you would be free later to choose between cooperating or not. As a matter of fact, however, you do not cooperate, and I believe so. I am later still free to retaliate, and the expected value of not doing so greatly exceeds that of doing so, but (x) I cannot adopt the intention to retaliate and believe you have failed to cooperate, without in fact retaliating. This is all common knowledge.}\]

I simply add clause (x) to assumption (D2) to get this assumption, (D3a). If the stipulation I have just described can indeed be consistently added to assumption (D2), then it seems we have a counterexample to (OC_3).
It seems plain, in this situation, that I rationally ought to have adopted the retaliation intention, that I rationally ought to believe you have not cooperated, and yet I rationally ought not to retaliate. Yet, if (as we have supposed) I cannot adopt this intention and believe you have not responded without retaliating, (OC3) says this is not possible, and so it seems that (OC3) is false.

The Fifth Counterexample. Imagine, again, the original situation described by assumption (D2), but this time add to its details the following:

You can tell whether or not I am the sort of person who would stick to their intentions. In particular, you can tell whether or not I would stick with my intention that if you do not cooperate then I retaliate. If you see that I would not, then, even though this is what I intend, you (sensibly) will not be deterred from defecting; if you see that I would stick with my intention, then likely you (sensibly) will be deterred. In a word, you take into consideration not just my intentions, but also my dispositions. I, therefore, face the following choice: to adopt, or not, the disposition that if I have the deterrent intention, and believe you have not cooperated, then I will in fact retaliate. It turns out that in order to get you to cooperate, it is a necessary and likely sufficient condition that I adopt this disposition.

It turns out, then, that I ought to adopt this disposition. Again, not that this is a grand logical truth about the normative relation between intention and unhindered action, but rather that this is a humble truth about my own situation.

This is another special sort of UDS. The assumption we need to make if it is indeed to address principle (OPn) is that situations of the following form are logically possible:

\[(D3b)\] (x) I am free to adopt the enduring disposition that if I intend to do something, and am aware that the relevant conditions obtain, then I do it, and the expected-value of adopting this disposition exceeds that of my not adopting it. (a') Whether or not I adopt this disposition, I am free then to adopt the intention to retaliate; the expected value of my adopting the intention is greater than that of not adopting it, since it is necessary and very likely sufficient to adopt the intention if you are to cooperate, (b') whether or not I
adopt the disposition, and whether or not I adopt the intention, you would be free later to choose between cooperating or not. As a matter of fact, however, you do not cooperate, and I believe so. I am later still free to retaliate, and the expected value of not doing so greatly exceeds that of doing so. This is all common knowledge.

I simply add a different clause (x) to assumption (D2), and slightly modify the other clauses, to get this assumption, (D3b). Note that there is a distinction to be made here between the deterrent intention – that is, the intention that if you do not cooperate then I will retaliate – and the disposition to carry such an intention out – that if I have the intention and believe you have not cooperated then I will retaliate. In the Fifth Counterexample, the intention and the disposition together are necessary and likely sufficient to get you to cooperate.

If the stipulation I have just described can indeed be consistently added to assumption (D2), then it seems we have a counterexample to (OP_n). It seems plain, in this situation, that I rationally ought to adopt the disposition to fulfil my intentions, I rationally ought to adopt the intention that if you cooperate then I retaliate, I rationally ought to believe you have not cooperated, and yet seems plain that I rationally ought not to retaliate. Yet, it is not logically possible for me (i) to adopt the disposition that if I intend that if p then I do A, and believe that p then I do A, (ii) adopt the intention that if you do not cooperate then I will retaliate, (iii) believe you have not cooperated, and yet (iv) not retaliate. All this contradicts (OP_4), and so it seems that (OP_4) is false.

The Fourth and Fifth Counterexamples have clear analogues in the First and Second Counterexamples, respectively, introduced in the discussion of the Toxin Puzzle in Chapter Two. To the description of the Toxin Puzzle, (T2), I added the stipulation that you could not have the intention to drink the toxin without drinking it, to get the First Counterexample, (T3a); to the description of Unhappy Deterrence Situations, (D2), I add the stipulation that you cannot intend to retaliate and believe you have not cooperated, without retaliating, to get the Fourth Counterexample, (D3a). I added the stipulation that it maximised expected-value for you to be disposed to keeping your intentions, to get the Second Counterexample, (T3b); I add a similar stipulation to get the Fifth Counterexample, (D3b).
[2] There are three relevant obligations in the Fourth Counterexample: I rationally ought to adopt the retaliation intention; I rationally ought to believe you have failed to cooperate; and, that I rationally ought not to retaliate. The Fifth Counterexample includes one more: I rationally ought to adopt the disposition that if I intend to retaliate and believe you have not cooperated, then I retaliate. In order to defend principles (OCn) and (OPn), I shall provide an independent argument for the claim that, given certain conditions, I am actually rationally permitted to retaliate in the Fourth and Fifth Counterexamples.

I will do so by concentrating on the notion of reconsideration, one which Michael Bratman, in his book *Intentions, Plans, and Practical Reason*, deals with at length. In the rest of this chapter I defend the following argument:12 [2.1] in a UDS, my opponent in right, and I rationally ought not to retaliate, only if I am rationally permitted to reconsider my (conditional) intention to retaliate; [2.2] in a UDS, if it maximises expected-value for me to be disposed not to reconsider the retaliation intention, then I rationally ought not to reconsider the intention; but, [2.3] in the UDSs described in the Fourth and Fifth Counterexamples, it does maximise expected-value for me to be disposed not to reconsider the intention. Hence, in the UDSs described in the Fourth and Fifth Counterexamples, I rationally ought not reconsider my retaliation intention, and so, in these circumstances, my opponent is wrong, and it is false that I rationally ought not to retaliate. I will consider each of the three major premises in turn.

[2.1] The first premise of this argument is that in a UDS, I rationally ought not to retaliate only if I am rationally permitted to reconsider my deterrent intention.13

In my view this first premise is one thoroughly embedded in 'the normal view of rationality' to which Kavka, above, refers. The

---

12 The details of this argument, below, are tentative. The general idea, though, is not. As will become apparent, the argument even as it stands now is long, and for this reason, and due to lack of space, I have chosen not to investigate all of its labyrinthine complexity. This I shall leave this for another occasion.

13 Though relying heavily on Bratman's work, I will need to point out the occasional differences. In this case, I start with the intention _cum_ policy that if you do not cooperate then I will retaliate. After I come to believe you have not cooperated, I actually face, according to Bratman, two choices: (a) whether or not to reconsider the policy, or (b) whether or not to block the application of the policy in this case. Bratman (p. 89) thinks that one can block the application of a policy without thereby giving it up. I am not sure, and so I will concentrate solely on case (a).
defender of the Self-Interest Theory, however, will not be happy with this claim, and will insist the premise is question-begging, since it is just one more bridging principle he will want to deny. I plead guilty, and can only comment that at this point the argument between myself and my opponent reaches stalemate. I shall have to content myself with trying to convince only those who are inclined to accept this first premise, and claim to the defender of the Self-Interest Theory only that if this first premise is true, then it is rational to retaliate.

At the risk of defending the obvious with the not-so-obvious, it is in fact possible to provide an argument for this first premise. The way to do so is on the basis of the following claim about the relation between intention, belief, reconsideration, and attempted performance:

(I2) it is not logically possible that (i) x adopt the intention at t1 that if p at t then he A's at t, (ii) x not reconsider this intention from t1 to t2 inclusive, (iii) x believes at t2 that p obtains, and (iv) x not (try to)\(^\text{14}\) A at t2.

We briefly considered, in Chapter Two, an inevitability principle such as this, but it faltered because it failed to take into account the possibility of fickleness. This new inevitability principle – (I2) – does not falter on this point.\(^\text{15}\) If this principle is correct, deontic principle (OP\(_4\)) entails with (I2) that

(2) it is not all the case that (i') x rationally ought to adopt the intention at t1 that if p at t then he A's at t, (ii') x rationally ought not to reconsider this intention from t1 to t2 inclusive, (iii') x rationally ought to believe at t2 that p obtains, and (iv') x rationally ought not (try to) A at t2.

In a UDS, we can all agree that (i'') I rationally ought initially to adopt the intention that if you do not cooperate then I retaliate, and can all agree that, after you have failed to cooperate, (iii'') I rationally ought to

\(^\text{14}\) Without this type of qualification, the principle would be false, since x might not be able to A at t2. I shall largely ignore this complexity in what follows, and shall assume that were I to try to retaliate, then I would succeed.

\(^\text{15}\) Something like it seems correct, but I will not have the space to defend this principle in detail. I am aware, of course, of (failed) attempts to argue for some sort of necessary connection between intention and action.
believe that you have not cooperated. It follows from these facts, and (2), that

(3) it is not both the case that (ii") I rationally ought not to reconsider the intention to retaliate from the time I initially adopt it to the time I realise you have failed to cooperate, inclusive, and (iv") I rationally ought not to (try to) retaliate at the time I realise you have failed to cooperate.

It follows from this that I rationally ought not to retaliate after you have failed to cooperate only if I am rationally permitted to reconsider my intention to retaliate sometime between when I initially adopted it and the time I realise you have failed to cooperate, inclusive. And this is simply the first premise.

The defender of the Self-Interest theory will object that this argument is also question-begging, since it employs, in the move from (I2) to (2), the principle (OP4), which is the very principle in doubt. I again plead guilty, but in mitigation would point out that the use of the relevant deontic principle in justifying the move from (I2) to (2) seems less problematic than its use in the Fifth Counterexample.

[2.2] The second premise of the argument is that, in a UDS, if it maximises expected-value for me to be disposed not to reconsider the deterrent intention, I rationally ought not to reconsider the intention. The argument for this second premise goes as follows. [2.2.1] If it maximises expected-value for me to be disposed not to reconsider the intention to retaliate, then I rationally ought to unreflectively not reconsider the intention; and [2.2.2] in an SDS, if I rationally ought to unreflectively not reconsider the intention to retaliate, then I ought simpliciter not reconsider the intention. I will examine these in turn.

[2.2.1] If it maximises expected-value for me to be disposed not to reconsider the intention to retaliate, I rationally ought to unreflectively not reconsider the intention. Understanding what this means, and why it is true, requires a number of steps.
First, we may identify two types of reconsideration or non-reconsideration. (Non) reconsideration of whether to A is, on the one hand, *unreflective* when it comes about as a result of ingrained habits or dispositions of (non) reconsideration. Thus, I might be such that if I come to believe an event of type E has occurred, I automatically reconsider any prior intention to A. (Non) reconsideration of whether to A is, on the other hand, *deliberative* when it comes about as a result of (implicit or explicit) deliberation about whether to reconsider. Thus, I (implicitly or explicitly) deliberate about whether to reconsider my decision to A, come to a decision to reconsider this decision, and then (presumably) go ahead with the deliberations involved in reconsidering whether to A. Bratman provides a good example of this distinction. Early this year I deliberated about whether to get earthquake insurance but decided not to. Most of the time I do not reconsider this decision. Such nonreconsideration is unreflective—it comes about as a result of ingrained habits or dispositions of (non) reconsideration. Occasionally, though, I receive unsolicited mail from insurance companies, which, on some occasions, prompts me to think about whether to reconsider my earlier decision. I think about whether to reconsider, decide that such reconsideration would take too much time, and thus do not reconsider my decision not to have earthquake insurance. Such nonreconsideration is deliberative—it comes about as a result of deliberation about whether to reconsider.

Second, we must now ask the question: under what conditions ought one to engage in *unreflective* (non) reconsideration? I will concentrate only on consequentialist answers to this question, of which there are two. On the one hand, an act-consequentialist theory would claim, roughly, that I ought to conduct unreflective (non) reconsideration on the basis of your having failed to cooperate if and only if it maximises expected-value (not) to reconsider that intention in such a situation. The act-consequentialist intuition is applied here to a particular class of actions—unreflective (non) reconsiderations. In particular, and supposing that if I were to reconsider then I would not retaliate, the act-consequentialist theory implies I ought to reconsider the deterrent intention. On the other hand, a rule-consequentialist

---

16 Following Bratman, I shall abbreviate the locution ‘reconsideration or non-reconsideration’ by ‘(non) reconsideration’.

theory would claim, roughly, that I ought unreflectively to conduct reconsideration on the basis of your having failed to cooperate if and only if it maximises expected-value for me to be disposed such that if you were not to cooperate then I would reconsider any intention to retaliate. The rule-consequentialist intuition is applied here to a particular class of actions – unreflective (non) reconsiderations. What, in particular, this theory says we will come to below.

It is clear, third, that a rule-, rather than an act-, consequentialist theory of rational unreflective (non) reconsideration is the correct one. I use a reductio argument to show this. Suppose, for reductio, that the act-consequentialist theory of rational unreflective reconsideration is the right one. This theory implies that

(1) I rationally ought to unreflectively reconsider my intention to retaliate.

Statement (1) is true because it maximises expected-value for me to reconsider (since we have supposed – quite reasonably – that if I were to reconsider then I would not retaliate). Unreflective reconsideration, though, is reconsideration coming about as a result of ingrained habits or dispositions of reconsideration, and not coming about as a result of (implicit or explicit) deliberation about whether or not to reconsider. In light of this, it follows from (1) that

(2) I rationally ought to be such that an ingrained (non-deliberative) habit of reconsideration causes me (in the right way) to reconsider my intention to retaliate.

In the case at hand, it maximises expected-value for me to reconsider. But if I were motivated to reconsider by this fact, then my reconsideration would have come about as a result of (implicit or explicit) deliberation. Anyone moved to A by the fact that the expected-value of A exceeds that of any alternative is moved by implicit or explicit deliberation. Anyone moved to reconsider by the fact that the expected-value of reconsidering is greater than not reconsidering is moved by implicit or explicit deliberation. The (implicit or explicit) deliberation involved, of course, is just that which takes place in the

'Retaliation Rationalised: Gauthier's Solution to the Deterrence Dilemma.' Pac Phil Quart 72 (1991), pp. 16, 23.
calculation of expected-values. To be moved by expected-values, then, is not to be moved by an ingrained (non-deliberative) habit of reconsideration. It follows from (2), then, that

(3) It is false that I rationally ought to be such that the fact it maximises expected-value for me to reconsider causes me (in the right way) to reconsider.

The act-consequentialist theory of rational unreflective reconsideration, hence, implies that (1) I rationally ought to reconsider (because it maximises expected-value), but (3) it is false that I ought to be moved to reconsider by the fact that it maximises expected value. And this is our absurdity. The theory says \( R \) (the fact it maximises expected-value to reconsider) is the reason why I ought to \( A \) (reconsider), but false that \( R \) should move me to \( A \). However, what is a reason to act apart from something which should move one to act? If the fact it maximises expected value is the reason I ought to reconsider, then this should move me to act. The act-consequentialist theory under consideration, though, denies this, and so is false.\(^{18}\)

In short, the act-consequentialist theory fails to provide an account of the rationality of unreflective (non) reconsideration – of reconsideration which is not deliberative. And it is just such a theory we are after. The conclusion, then, is that the only adequate consequentialist theory of the rationality of unreflective (non) reconsideration will be a rule-consequentialist one. That is, I rationally ought to unreflectively reconsider whether to \( A \) if and only if it maximises expected-value for me to have habits of which would lead to reconsidering whether or \( A \). Therefore, if it maximises expected-value for me to be disposed not to reconsider the intention to retaliate, I rationally ought to unreflectively not reconsider the intention.

---

\(^{18}\) The act-consequentialist may wish to appeal to the familiar distinction between theories offering truth-makers for 'ought' statements (which are reasons why they are true) and those offering decision-procedures (which are what should motivate people to act). He may insist that his theory of rational unreflective reconsideration aims to provide a truth-maker, and not a decision-procedure, for reconsideration. See R. E. Bales, 'Act-Utilitarianism: Account of Right-Making Characteristics or Decision Procedure?', *Amer Phil Quart* 8 (1971), 257-65. In my view, the distinction is irrelevant. The rational 'ought' is concerned with the guidance of action – particularly in the case of reconsideration – and so the appropriate theory of the rational 'ought' will provide a decision-procedure, and not just a truth-maker. I cannot elaborate on this point further, but see my brief discussion of 'objectivist' theories in Chapter One, §5[2].
[2.2.2] But is unreflective (non) reconsideration of the intention to retaliate the type which is of importance in a UDS? The answer is yes: in an UDS, if I rationally ought to unreflectively not reconsider the intention to retaliate, I ought simpliciter not reconsider the intention. The important point is that it seems reasonable to suppose, in the case of an UDS, that the sort of reconsideration required (or permitted) is unreflective reconsideration.

I shall argue that, in a UDS, we may assume I reconsider the intention to retaliate if and only if I unreflectively reconsider the intention. This breaks up into two claims: we may suppose that (a) I reconsider the intention if I unreflectively reconsider the intention, and (b) I reconsider the intention to retaliate only if I unreflectively reconsider the intention. This first claim is obvious, and I shall take it as proven.

The second not so obvious. Just after I have come to believe you have failed to cooperate, but before I trigger the Bomb, all of the following are true:

(1) I am not reconsidering the intention to retaliate;
(2) I am not reconsidering the non-reconsideration in (1);
(3) I am not reconsidering the non-reconsideration in (2);
(4) and so on ...

Suppose, now, that I reconsider the intention to retaliate (and so (1) has become false). Is such (first-order) reconsideration unreflective or deliberative? If it is deliberative, then (by definition) it must have come about as a result of my deliberating whether or not to reconsider the intention. But if I deliberated about whether or not to reconsider, (2) must have become false. Therefore, I must have reconsidered the non-reconsideration of the intention. Is such (second-order) reconsideration unreflective or deliberative? If it is deliberative, then (by definition) it must have come about as a result of my deliberating whether or not to reconsider the non-reconsideration of the intention to retaliate. But if I deliberated about whether or not to reconsider the non-reconsideration of the intention, then (3) must have become false. Therefore, I must have reconsidered the non-reconsideration of the non-reconsideration of the intention. And so on. ... Since this process cannot continue indefinitely, there must be a level – at line (n) – at which the (higher-order) reconsideration of nonreconsideration of
nonreconsideration of ... was unreflective – that is, just occurred on the basis of some change in circumstances and as a result of habits or dispositions of (non) reconsideration I happened to have. This means that (b') I reconsider the intention to retaliate only if there is some higher-order, \( n \), at which I have unreflectively reconsidered my non-reconsideration of non-reconsideration ... of my intention. It would be possible to reformulate all of the arguments below in terms of this \( n \)th level unreflective reconsideration, but that would be tedious. It is simpler to suppose that the unreflective reconsideration occurs at level-1. It is simpler to suppose that (b) I reconsider the intention to retaliate only if I unreflectively reconsider the intention.

Therefore, in a UDS, I reconsider my intention if and only if I unreflectively reconsider it. It follows that in an UDS, if I rationally ought to unreflectively not reconsider the intention to retaliate, then I ought not reconsider the intention.

The first two premises of the overall argument have now been established: [2.1] in a UDS, my opponent is right, and I rationally ought not to retaliate, only if I am rationally permitted to reconsider my intention; [2.2] in a UDS, if it maximises expected-value for me to be disposed not to reconsider the deterrent intention, I rationally ought not to reconsider the intention. It follows from both of these premises that, in a UDS, my opponents are right, and I rationally ought not to retaliate, only if it does not maximise expected-value for me to be disposed not to reconsider the deterrent intention.

[2.3] However, in the UDSs in the Fourth and Fifth Counterexamples – the ones causing me problems – it does maximise expected-value for me to be disposed not to reconsider the deterrent intention. This is the third, and final, premise for my argument that, in the Fourth and Fifth Counterexamples, it is indeed rational to retaliate. To show this final premise is true, consider each example in turn.

In the case of the Fourth Counterexample, I cannot but be disposed to not reconsidering this intention. Recall that, in this example, I cannot but be disposed such that if I intend that I retaliate upon your non-cooperation, I believe you have not cooperated, then I retaliate. Since, in this example, there are no alternatives to the disposition not to reconsider the retaliation intention, it follows trivially that the expected-value to me of this disposition is greater than that of all its
alternatives. It follows trivially, in other words, that it maximises expected-value for me to have this disposition.\footnote{Peter Menzies has raised the worry that if there are no alternatives to $A$, then while $A$ does indeed maximise expected-value as a trivial result, the issue of whether one ‘ought’ to $A$ also does not arise. In response, I say that (a) I am not sure that it doesn’t, and, in any case, (b) I am confident that the (already complex) argument can be modified to take account of this point. Lack of space prevents me from elaborating on these points.}

In the case of the Fifth Counterexample, it follows non-trivially that it maximises expected-value for me to have this disposition. Recall that, in this example, you take into consideration not my intentions, but my dispositions, and it turns out that in order to get you to cooperate, it is a necessary and likely sufficient condition that I adopt this disposition.

The argument is complete. Just above, we saw that in a UDS, my opponents are right, and I rationally ought not to retaliate, only if it does not maximise expected-value for me to be disposed not to reconsider the deterrent intention. However, in the UDSs in the Fourth and Fifth Counterexamples, it does in fact maximise expected-value for me to be disposed not to reconsider the deterrent intention. Therefore, in these two examples, it is indeed rational to retaliate, even though one is free not to do so, and it has the best outcome not to do so. My opponents are wrong.

This argument provides independent support for the claim – a corollary to the claim established in Chapter Two, §4[2] – that GIVEN that you rationally ought to adopt, or cannot but have, the enduring disposition to act on the deterrent intention if the other does not cooperate, THEN if you rationally ought to adopt the deterrent intention, and you rationally ought to believe the other has not cooperated, then you are rationally permitted to retaliate, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise, and even though you end up in a situation worse than if you had not adopted the intention in the first place (though retaliation may very well be irrational absent this condition).

\section{Reply to Objections}

I will end this chapter by dealing briefly with some, by now, familiar objections. The objections are threefold, and my treatment of them will be brief, and I will conclude that none are effective.
Is it really still possible for me not to retaliate? Consider the Fourth Counterexample. It assumes, on the one hand, that I can both adopt the intention to retaliate and believe you have failed to cooperate, and that even if I adopt the intention and have the belief, then I would still be free not to retaliate. It assumes, on the other hand, that I cannot: adopt the intention, believe you have failed to cooperate, and not retaliate. The objection is that this is just inconsistent.

The response, as we saw in Chapter Three, is just to claim that the objection depends on an invalid inference. Since this inference is invalid, I may act freely when I retaliate.

Does it really still have the best outcome for me not to retaliate? Consider the Fifth Counterexample. It assumes, amongst other things, that even if I adopt the disposition such that if I have the retaliation intention and believe you have not cooperated then I will retaliate, and even if I adopt the intention and have the belief, then the expected-value of not retaliating would still be greater than that of retaliating. The objection is that if I adopt the disposition, adopt the intention, and have the belief, then I will end up retaliating, and my retaliation will constitute an intentional action only if it now maximises expected value for me to retaliate.

This objection rests on a question-begging view concerning the nature of intentional action. Weakness of will to one side, what may be true is an agent performs some action A intentionally only if it maximises the expected-strength of reasons for or against A, but what is not true is that they perform this action intentionally only if it maximises the expected-value of A. In adopting the deterrent intention, I become a vengeful person. I think (rightly or wrongly) that the fact you have not cooperated is a sufficient reason for me to wreak a terrible vengeance upon you, even though I recognise that the outcome of doing so will be our mutual doom. If I do think this, then I may act intentionally in retaliating, even though it maximises expected-value for me to do otherwise.

What reason is there for me to retaliate? My retaliation in a UDS may be both free and intentional, and the previous two objections are

---


21 Some are inclined to think one has no reason to retaliate. See D. MacIntosh, ‘Retaliation Rationalised: Gauthier’s Solution to the Deterrence Dilemma,’ *Pac Phil Quart* 72 (1991), p. 16. S. I. Benn, ‘Deterrence or Appeasement? Or, on trying to be rational about nuclear war,’ *J Applied Phil* 1 (1984), sec. VI(b), and R. Dworkin,
mistaken. The conclusion of the previous section stands: in the Fourth and Fifth Counterexamples, I am rationally permitted to retaliate. I am rationally permitted to retaliate, however, if and only if I have a sufficient reason to retaliate. What is that reason? Simple: the fact you have not cooperated. The claim is that if I adopt the (conditional) intention to retaliate in the Fourth Counterexample, or I both adopt the intention-keeping disposition and the deterrent intention in the Fifth, then the fact that you have failed to cooperate is a sufficient reason for me to retaliate. More so than in previous cases, it is important to note two things about this claim. First, this is a claim about the reason only in the Fourth and Fifth Counterexamples, and not a claim that, in all cases, another’s failure to acquiesce in a threat is a reason to retaliate. Second, even in this restricted class of cases, the claim is only that I have reason to retaliate if the specified conditions obtain, and is not that I have a reason to retaliate, even if I do not have the relevant intention or the relevant disposition.

Conclusion

The Paradox of Deterrence presents no threat to my argument, in Chapter Five, for the rationality of cooperation. On the one hand, the supposedly problematic types of situation — the Unhappy Deterrence Situations — provide no counterexamples to the principles \((OC_n)\) and \((OP_n)\) which are central to my inference from the rationality of being disposed to cooperation to that of actually cooperating. Special cases of the situations — the Fourth and Fifth Counterexamples — do present some problems, but these are not insurmountable if one considers the conditions under which it would be rational to reconsider a prior intention to retaliate. Vengeance is not just the Lord’s, say I.

---

22 Gauthier suggests a different reason: ‘Her reason for sticking to her guns [and executing a failed threat] is not to teach the others by example, not to improve her prospects for successful deterrence in the future, or anything of the sort. Her reason is simply that the expected utility or payoff of her failed policy depended on her willingness to stick to her guns’ in ‘Deterrence, Maximization, and Rationality,’ (p. 489).
Conclusion

Non-Consequentalist Reasons for Action

Given certain conditions, the actions resulting from rational intentions (or agreements) are also rational, even if one is free to do otherwise and it has the best outcome for one to do otherwise (though these actions may very well be irrational absent those conditions). Or so I have argued in this thesis.

I started with the claim that a plausible view about the nature of rationality is that the rational intentions are those maximising expected-value. Some say that since such actions are the result of intentions it is presumably rational to have, then they too must be rational; others that since such actions are not expected-value maximising, then they must be irrational. The second group suggest, in other words, that the Self-Interest Theory - introduced in Chapter One - would be true:

(S) If an agent is free to perform an action A, then they rationally ought to A if and only if the agent-relative expected-value of doing A exceeds that of doing any alternative to A.

Though I have negotiated a middle way between these opposing views, it is the second I was particularly concerned to deny, and I devoted the first and third parts of the thesis to presenting, and defending, counterexamples to the Self-Interest Theory.

My argument against the Self-Interest Theory - introduced in Chapter Two - is simple. An eccentric billionaire placed before you a vial of toxin that, if you drink it, would make you painfully ill for a day, but would not threaten your life. The billionaire would pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasised that you need not drink the toxin to receive the money; in fact, the money would already be in your bank account hours before the time for
Conclusion

Non-Consequentialist Reasons for Action

Given certain conditions, the actions resulting from rational intentions (or agreements) are also rational, even if one is free to do otherwise and it has the best outcome for one to do otherwise (though these actions may very well be irrational absent those conditions). Or so I have argued in this thesis.

I started with the claim that a plausible view about the nature of rationality is that the rational intentions are those maximising expected-value. Some say that since such actions are the result of intentions it is presumably rational to have, then they too must be rational; others that since such actions are not expected-value maximising, then they must be irrational. The second group suggest, in other words, that the Self-Interest Theory – introduced in Chapter One – would be true:

(S) If an agent is free to perform an action A, then they rationally ought to A if and only if the agent-relative expected-value of doing A exceeds that of doing any alternative to A.

Though I have negotiated a middle way between these opposing views, it is the second I was particularly concerned to deny, and I devoted the first and third parts of the thesis to presenting, and defending, counterexamples to the Self-Interest Theory.

My argument against the Self-Interest Theory – introduced in Chapter Two – is simple. An eccentric billionaire placed before you a vial of toxin that, if you drink it, would make you painfully ill for a day, but would not threaten your life. The billionaire would pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasised that you need not drink the toxin to receive the money; in fact, the money would already be in your bank account hours before the time for
drinking it arrives, if you succeeded. (The presence or absence of the
intention was to be determined by the latest 'mind-reading' brain
scanner ...). It turns out, however, that you could not have had the
intention at midnight without, the next afternoon, drinking the toxin.
In such a situation:

\[\text{T3a} \quad \begin{aligned}
& (a) \text{You are free to adopt the intention to drink the toxin, and the} \\
& \text{expected value of your adopting this intention exceeds that of your} \\
& \text{not adopting this intention; and (b) whether or not you actually} \\
& \text{adopt the intention, you are free later not to drink the toxin, and the} \\
& \text{expected value of your not drinking it still exceeds that of your} \\
& \text{drinking it; but (x) you cannot both adopt the intention to drink the} \\
& \text{toxin, and then, later, not drink the toxin. You know all this is the} \\
& \text{case.}
\end{aligned} \]

One might be inclined to question the consistency of this
assumption, but – as we saw in Chapter Three – one would be
mistaken. On the one hand, it is consistent with your not being able to
have the intention without drinking that – if you have the intention –
you are still free not to drink. And on the other hand, it is consistent
with your having the intention, and acting upon it, that it still has the
best outcome not to drink. This assumption, (T3a), does however
imply that, if the Self-Interest Theory is true, then you rationally ought
to adopt the intention and rationally ought not drink the toxin, even
though you cannot adopt the intention without drinking the toxin.
This, however, is not possible:

\[\text{OC2} \quad \begin{aligned}
& \text{If an agent rationally ought to } S_1, \text{ and rationally ought to } S_2, \text{ then} \\
& \text{he can be such that he } S_1 \text{'s and he } S_2 \text{'s.}
\end{aligned} \]

One might be inclined to deny the truth of this so-called deontic
principle, but – as we saw in the third part of the thesis – one would be
mistaken. In Chapter Seven, we saw that Slote's novel and not-so-
 novel attempts failed to show that rational dilemmas are possible; in
Chapter Eight, that Parfit's attempts failed to show that it could be
rational to cause oneself to act irrationally; and, in Chapter Nine, that
Kavka's introduction to the paradox of deterrence failed to show that
this deontic principle is false. The conclusion is clear: the Self-Interest
Theory is false.
This claim has implications concerning what does, and what does not, give one reason to act. Statement (S) I have called the standard formulation of the Self-Interest Theory. Yet this statement merges two central ideas which – as we saw in Chapter One – it is best to keep separate. On the one hand, there is the general idea that there are many reasons for or against any action, and the influence a certain reason should have in determining the rationality of the action is directly proportional to the expectation that that reason obtains, and also to the (intrinsic) strength of that reason. There is the claim that

(R) If an agent is free to perform an action A, then they rationally ought to A if and only if the expected strength of reasons for doing A exceeds that of doing any alternative to A.

On the other hand, there is the specific idea that all and only considerations about outcomes provide reasons for action, and that the only relevant expectations and values are the agent's own. There is also the central tenet of the Self-Interest Theory, that:

(S') (1) A consideration p is a reason for or against performing some action A if and only if it takes the form 'q would be (part of) the outcome of x's doing A'; (2) the weight of such a consideration is to be given by the agent's expectation that q would be (part of) the outcome of x's doing A; (3) and the (intrinsic) strength of such a consideration is given by the value to the agent of outcome q.

Statement (R) and (S') together entail the standard formulation the Self-Interest Theory, (S), which, as we have seen, is false. Hence, one of these latter statements must also be false. Since (R) is true, I conclude that (S') should be abandoned. This statement itself consists of basically two claims: about the sorts of things which are reasons (namely, statements about outcomes of acting), and about the sorts values and expectations which are relevant (namely, yours). One of these claims must be false. It is clear, though, that the counterexample refutes the theory's claim that all reasons for drinking the toxin need to mention possible outcomes of drinking the toxin. I conclude, then, that there are non-consequentialist reasons for action.

It is possible to generalise the argument in the previous paragraphs. For it follows from (OC_n) (and (OP_n)) that GIVEN you
rationally ought to adopt, or cannot but have, the enduring disposition to do what you intend, THEN if you rationally ought to adopt the intention to perform some action, then you are rationally permitted to perform that action, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise (though this action may very well be irrational absent this condition). This statement has two important corollaries.

On the one hand, it allows one to argue that some agreements to act make it rational so to act. I gave the example that being the sort of person who keeps agreements might most promote your interests – since then others would be more inclined to enter into beneficial agreements with you – even though some of the resulting actions of agreement-keeping do not most promote your interests. GIVEN that this is so, THEN if you rationally ought to believe you have agreed to perform some action, you are rationally permitted to perform that action, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise. Under some conditions, promise-keeping is an act contrary to self-interest which is nevertheless rational.

On the other hand, this implication of principles \((OC_n)\) and \((OP_n)\) allow one to argue that some deterrent intentions make it rational to retaliate. I also gave the example that being the sort of person who intends to retaliate if encroached upon might most promote your interests – since then others would likely refrain from encroaching – even though any particular act of retaliation (if, indeed, any occur) does not most promote your interests. GIVEN that this is so, THEN if you rationally ought to believe you have been encroached upon, you are rationally permitted to retaliate, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise. Under some conditions, revenge is an act contrary to self-interest which is also rational.

I opened the thesis with the question: are the non-expected-value maximising actions resulting from expected-value maximising, and so plausibly rational, intentions themselves rational? I say yes – given certain conditions, it is rational to act on the basis of rational intentions, even if contrary to self-interest so to act.
Appendix

Parfit's 'Endnote 14'

Parfit thinks (and I agree) that theory S implies that Kate cannot avoid acting irrationally. He recognises some might see in this implication an objection to theory S, and he has asked us to consider Clare's plight which, he claimed, showed that the correct interpretation of Kant's dictum is simply that it is impossible for someone to avoid acting \textit{in a way that deserves to be blamed}, and not that it is impossible for someone to avoid acting \textit{irrationally}. After considering Clare's plight, and again referring to his discussion of \textit{Schelling's Answer}, Parfit suggests further that '[t]o meet the objection to C, Clare might appeal to other cases where we cannot avoid acting wrongly. That there are such cases has been claimed by some of the writers who are most opposed to C. I discuss this answer in endnote 14.' (p. 37). Since I believe more needs to be said to meet this objection, I will examine what Parfit has to say in the four-page-long Endnote 14.

Parfit's reference to these 'other cases where we cannot avoid acting wrongly' is a reference to Bernard Williams's and Thomas Nagel's discussions of the possibility of moral dilemmas (pp. 507-9). Since I have, in the previous chapter, discussed rational dilemmas (and, I believe by implication, moral dilemmas) I will not examine this part of the endnote. Rather, I want to discuss a further argument which may be teased from it. The argument is that, even if 'ought' did typically imply 'can', this dictum would in any case be inapplicable to Kate's situation. Parfit primarily discusses the moral case, but it will save words if I put his argument in terms directly relevant to the current discussion - in terms of rationality. His argument in this endnote is complex, and depends on material in the main body of the text, so I will [1] introduce and interpret it, before [2] conducting a critical examination.

[1] Parfit's objection to the applicability of doctrine that 'ought' implies 'can' hinges firstly on his understanding, provided early in \textit{Reasons and
Persons, of the words 'can' and 'cannot'. Here is what he has to say on this topic:

[(P1)] In the doctrine that ought implies can, the sense of can is compatible with Psychological Determinism [the view that our acts are always caused by our desires and dispositions]. 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. (p. 15)

For Parfit then, the notion of 'cannot' relevant to the Kant's dictum is the notion that one 'cannot, whatever one's desires and dispositions might have been.' (p. 16)

The second part of his objection is concerned with the type of inability with which Kate is afflicted. Parfit says that someone is never self-denying when they never do what they believe will be worse for themselves. For example, were Kate never self-denying, she would have made it that she most desire her books be as good as possible, and also she would not work so hard. The Self-Interest theory, recall, tells Kate that she ought to cause herself to have this desire, and that she ought not to work so hard. The second part of Parfit's argument (which I introduced in Chapter Eight, §1[1]), though, depicts Kate complaining that

[(P2)] It is not possible both that I have one of the best possible set of motives [which includes her desire that her books be as good as possible], ... and that I never do what I believe to be irrational [in particular, that she not work so hard]. This is not possible in the relevant sense: it is not possible whatever my desires and dispositions are. If I were 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. (p. 16)
Parfit thinks it quite generally true, even if we do not explicitly assume it, that she cannot do both of these things. The important thing to note, though, is that for Parfit the explanation of the impossibility of Kate's doing both of these things essentially involves claims about her desires and dispositions.

The final part of Parfit's objection to the applicability in Kate's case of the dictum that 'ought' implies 'can' - and the new element introduced in Endnote 14 - is that if the explanation of the relevant impossibility essentially involves this sort of reference to desires and dispositions, then this means that the doctrine is not applicable. Kate complains that it is impossible for her both to have the best desire and not overwork, and that this relieves her of any obligation to do each. Parfit, though, is not so sure. Here is what he says (with the relevant comments about the moral case replaced by suitable comments about the rational case):

\[ (P3) \] Is this impossible in the sense which justifies an appeal to the doctrine that ought implies can? Is it impossible that [Kate] never act in this way, whatever [her] desires and dispositions are, or might have been? This is true, but misleading. It suggests that this impossibility has nothing to do with what [her] desires and dispositions are. This is not so. This impossibility essentially involves claims about [Kate's] desires and dispositions. Why is it impossible that [she never do what she believes will be worse for her]? This is impossible because there is only one disposition given which it would be causally possible [for her to never do what she believes will be worse for her], and causing [herself] to have or to keep this disposition would itself be a case of doing what [she believes will be worse for her]. Because this impossibility essentially involves these claims about [her] desires and dispositions, it is not clear that this is the kind of impossibility that justifies an appeal to the doctrine that ought implies can. It can at least be said that this case is very different from the case where it is impossible for me to save both of a pair of lives. That impossibility had nothing to do with my desires or dispositions. (pp. 506-7, initial emphasis added)

In this passage Parfit has the following type of situation in mind. A person can either save one life, or save another life, but cannot save both. It might be, for example, that these people are drowning in a swollen river, and that the potential rescuer is not a strong enough
swimmer. Whichever life is saved, they are failing to save someone's life.

Parfit's argument must, then, be this. (P1) The relevant notion of 'cannot' in the dictum that 'ought' implies 'can' is the notion that one 'cannot, whatever one's desires and dispositions are'. It follows from this that (P3) if the impossibility occurring in a situation essentially involves claims concerning desires and dispositions, then it is not clear it is the kind of impossibility justifying an appeal to the doctrine. In particular, (P2) the impossibility in Kate's situation does essentially involve claims about her desires and dispositions. Therefore, so the objection goes, it is not clear that the dictum that 'ought' implies 'can' applies to Kate's situation.

[2] So much for introducing the argument; now to examine it. The proper analysis of 'can' is a minefield, and one in which I am not particularly keen to tread. In light of this, I will grant Parfit's first premise, that the relevant notion of 'cannot' is the notion that one 'cannot, whatever one's desires and dispositions are'. No doubt, as Parfit points out, one would want a notion that was compatible with Psychological Determinism, though it might be going too far to suggest that the relevant sense of 'can' is as completely independent of desires and dispositions as he suggests. I shall concentrate instead on other parts of the argument.

The first objection to the argument is that it seems at best question-begging, and at worst false, to suppose, as (P3) does, that just because the impossibility occurring in a situation essentially involves claims about desires and dispositions, then the doctrine that 'ought' implies 'can' is not applicable. Parfit himself admits that the impossibility Kate faces is the sort required by this doctrine, even though it does essentially refer to desires and dispositions. He says: 'Is it impossible that [Kate] never act in this way, whatever [her] desires and dispositions are, or might have been? This is true, but misleading' (p. 506, emphasis added). Though he claims this is misleading, it is hard to see this as anything but question-begging.

The main objection to Parfit's argument, though, is that it is not clear that the explanation of the impossibility in Kate's situation does necessarily involve a problematic reference to her desires and dispositions. Parfit thinks it does because he seems to think [2.1] the impossibility involved in Kate's case is to be explained completely by
the argument occurring in (P2), and [2.2] this argument essentially involves a problematic reference to Kate's desires and dispositions. But both of these claims can be denied.

[2.1] The first thing to note is that any impossibility involved in Kate's situation is not completely to be explained by the argument Kate gives in (P2), since this argument is invalid. What we are told about Kate's situation is that a certain conditional obtains: to wit, that if she were to have the relevant desire then she would overwork. Kate complains: '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' (p. 16). This may be true, but it does not follow, as the argument requires, that she cannot cause herself to have one of the best sets of motives without doing something irrational. It does not follow, in particular, that she cannot have the desire that her books be as good as possible without overworking. This afternoon I have some work to do in the library. But if I go to the library, I shall procrastinate, and read the day's newspapers. It does not follow that I cannot go to the library without procrastinating, just that (weak as I am) I won't. Similarly, even if we know that were Kate to have the relevant desire she would overwork, it does not follow she couldn't do both of these things together, it just might mean that (weak as she is) she won't. We can make no inference from the pertinent conditional to whether or not Kate can both have the strongest desire and yet not overwork herself. From what we know, it does not follow that she can not both cause herself to have one of the best sets of motives, and yet be never self-denying, it only follows that she will not do both of these things. The argument occurring in passage (P2) is invalid.

This means that, if there is to be any sort of impossibility occurring in the situation at all, we must suppose (as I did in Chapter Eight, §1) that it is a particular fact about Kate that she cannot both have this desire and not work so hard. Unlike other persons, she lacks a certain capability: desiring most that her books are as good as possible without thereby overworking herself. But if this is what is true of Kate, then her incapacity looks a lot more like that of the person above, who cannot save both of the drowning people. Unlike other persons (we may suppose), she lacks a certain capacity: being able to swim strong enough so as to be able to save both. If the inabilities in the two cases are in this
way analogous, then the doctrine that 'ought' implies 'can' will be as applicable to Kate as we all agree it is to the person above.

[2.2] The second thing to note is that, even if the argument in (P2) were valid and were the complete explanation for the impossibility which occurs in Kate's situation, it does not in any case essentially involve a problematic reference to Kate's desires and dispositions.

Suppose some theory tells Kate that (a') she ought to cause herself to have the desire that her books be as good as possible, and tells her, in addition, that (b') she ought not to have this desire. It might be theory S, since this theory in effect tells Kate to cause herself to have this desire (since it tells her to cause herself to have one of the best sets of motives, all of which include this desire), but also in effect tells her not to have this desire (since it tells her not to overwork, which is what is caused by this desire). As she did above, in (P2), Kate might complain as follows:

\[
\text{(P2')} \text{ It is not possible both that I cause myself to have the desire that my books be as good as possible, and that I not have this desire. This is not possible in the relevant sense: it is not possible whatever my desires and dispositions are. If I were not to have this desire, then I would have satisfied the requirement not to have this desire. But I would have acted irrationally in causing myself to lose, or allowing myself to remain without, this desire. If instead I cause myself to have this desire, then I will have this desire, and so will have failed to satisfy the requirement not to have this desire.}
\]

In this case, the explanation Kate would give for inability is valid, for it is necessarily the case that she cannot cause herself to have a desire without having it. But note that even though the explanation in (P2') essentially involves claims about her desires or dispositions, in a way structurally analogous to the case Parfit thinks is problematic, this does not mean that Kant's dictum is inapplicable to (P2').

So it is with the original explanation in (P2). If we assume that the explanation of Kate's inability is in this case valid, then even though it involves claims about her desires or dispositions, this again does not mean that the dictum that 'ought' implies 'can' is inapplicable. If we assume that the explanation of Kate's inability is in this case valid, as in the case I have just presented, then there will be no grounds for
supposing that the doctrine applies in the second, (P2'), and not in the first, (P2).

The analogy should be clear. The complaint Parfit puts into Kate's mouth, (P2), about the requirements S makes on her, involves a conflict between (a) her causing herself to have one of the best set of motives (including a desire that her books be as good as possible), and (b) her being never self-denying (and so not overworking). The complaint I put into Kate's mouth, (P2'), about the requirements a fictitious theory makes on her, involves a conflict between (a') her causing herself to desire most that her books be as good as possible, and (b') her not having this desire. As (a) is to (b) in (P2), so (a') is to (b') in (P2').

The supporter of the Self-Interest Theory cannot therefore show that Kate's inability to have the best desire without performing non-best actions is relevantly different from the sort of inability referred to in Kant's dictum that 'ought' implies 'can'. Therefore, for all that has been said in this Appendix, and in Chapter Eight, it remains a good objection to Self-Interest Theory that Kate cannot always avoid doing what it claims to be irrational.
References


Adams, R. M., 'Motive Utilitarianism,' *J Phil* 73 (1976): 467-481


Aqvist, L., 'Good Samaritans, Contrary-to-Duty Imperatives, and Epistemic Obligations,' *Mind* 69 (1960): 289-300


Barnes, G., 'Utilitarianisms,' *Ethics* 82 (1971): 56-64

Benn, S. I., 'Deterrence or Appeasement? Or, on trying to be Rational about Nuclear War,' *J Applied Phil* 1 (1984): 5-20


Bergstrom, L., 'Utilitarianism and Deontic Logic,' *Analysis* 29 (1969): 43-4
Brink, D. O., ‘Utilitarian Morality and the Personal Point of View,’ *J Phil* 83 (1986): 417-438
Castenada, H-N., *The Structure of Morality*, (Springfield, Ill.: Thomas, 1974)
Chisholm, R., ‘Contrary-to-Duty Imperatives and Deontic Logic,’ *Analysis* 23 (1963): 33-6
Conee, E., ‘Why Moral Dilemmas are Impossible,’ *Amer Phil Quart* 26 (1989): 133-141


Feldman, F., Doing the Best that We Can, (Dordrecht: Reidel, 1986)


Frank, R. H., Passions within Reason, (New York: Norton, 1988)


Gauthier, D., ‘Morality and Advantage,’ Phil Rev 76 (1967): 460-75


Gauthier, D., 'Deterrence, Maximization, and Rationality,' *Ethics* 94 (1984): 474-95
Gauthier, D., 'The Unity of Reason: A Subversive Reinterpretation of Kant,' *Ethics* 96 (1985), p. 85,
Gruzalski, B., 'Parfit's Impact on Utilitarianism,' *Ethics* 96 (1986): 760-783
REFERENCES

Hardin, R., & B. Barry (eds), Rational Man and Irrational Society?
(Beverly Hills: Sage, 1982)
Hardin, R., Pragmatic Intuitions and Rational Choice,' in A. Diekmann & P. Mitter
(Eds), Paradoxical Effects of Social Behaviour: Essays in Honour of
Harman, G., 'The Inference to the Best Explanation,' Phil Rev 74 (1965):
88-95.
Harman, G., ‘Reasons,’ Critica 7 (1975): 3-17
Harman, G., ‘Rationality in Agreement: a Commentary on Gauthier’s
Harsanyi, J., ‘Rule Utilitarianism and Decision Theory,’ Erkenntnis 11
(1977): 23-53
Hegselmann, R., ‘Rational Egoism, Mutual Advantage, and Morality –
A Review of D. Gauthier: Morals by Agreement,’ Erkenntnis 31
(1989): 143-59
Hobbes, T., Leviathan, (Ed.) C. B. MacPherson, (Harmondsworth:
Penguin, 1985)
Press, 1967)
Homer, The Odyssey, Trans. R. Lattimore (New York; Harper Perennial
1991)
Howard, N., The Paradoxes of Rationality: the Theory of Metagames
(1988): 489-502
Oxford Uni. Press, 1888)
Jackson, F. & R. Parfit, ‘Oughts, Options, and Actualism,’ Phil Rev
95 (1986): 233-55
Jackson, F., ‘Understanding the Logic of Obligation, I,’ Proc Aris Soc
Jackson, F., ‘Decision-Theoretic Consequentialism and the Nearest and
REFERENCES


Kant, I., Critique of Practical Judgement, Trans. L.W. Beck, (Indianapolis: Bobbs-Merrill, 1956)


Kuflick, A., 'A Defense of Commonsense Morality,' Ethics 96 (1986): 784-803


Lango, J. W., 'Is it Wrong to Intend to do that which it is Wrong to do?', Monist 70 (1987): 316-329


Luban, D., 'The Paradox of Deterrence Revived,' Phil Stud 50 (1986): 129-41


Mack, E., 'How to Derive Ethical Egoism,' *Personalist* 52 (1971): 735-743


Marcus, R., 'Moral Dilemmas and Consistency,' *J Phil* 77 (1980): 121-136


McConnell, T. C., 'Moral Dilemmas and Requiring the Impossible,' *Phil Stud* 29 (1976): 409-13

McConnell, T. C., 'Moral Dilemmas and Consistency in Ethics,' *Can J Phil* 8 (1978): 269-287


Mendola, J., 'Gauthier's 'Morals by Agreement' and Two Kinds of Rationality,' *Ethics* 97 (1987): 765-74


Nathenson, S., 'Nonevidential Reasons for Belief: a Jamsian View,' *Phil Phenom Res* 42 (1981-2): 44-54,
Nussbaum, M., 'Aeschylus and Practical Conflict,' *Ethics* 95 (1985): 233-267
Parfit, D., 'Comment,' *Ethics* 96 (1986): 832-872
Piper, A. M., 'Utility, Publicity, and Manipulation,' *Ethics* 88 (1978): 189-206
Quinn, W., 'The Right to Threaten and the Right to Punish,' *Phil Pub Affairs* 14 (1985): 327-73
REFERENCES

Raz, J., 'Reasons for Actions, Decisions and Norms,' *Mind* 84 (1975): 481-499
Rickman, H. P., 'Escapism: The Logical Basis for Ethics,' *Mind* 72 (1963): 273-4
Santas, G., 'Plato's "Protagoras" and Explanations of Weakness,' *Phil Rev* 75 (1966): 3-33
Sapontzis, S. F., 'The Obligation to be Rational,' *J Val Inq* 13 (1979): 294-298
Sartorius, R., *Individual Conduct and Social Norms*, (Encino: Dickenson, 1975)
Sinnott-Armstrong, W., Moral Dilemmas, (Oxford: Basil Blackwall, 1988)
Steiner, H., 'Moral Conflicts and Prescriptivism,' Mind 91 (1973): 586-591
Stocker, M., 'The Schizophrenia of Modern Ethical Theories,' J Phil 73 (1976): 453-466
Stocker, M., 'Moral Dilemmas: What they are and why they matter,' Phil Quart 68 (1987): 104-23
Tannsjo, T., 'Moral Conflict and Moral Realism,' *J Phil* 82 (1985): 113-7
von Fraassen, B., 'Values and the Heart's Command,' *J Phil* 70 (1973): 5-19
Vorobej, M., 'Gauthier on Deterrence,' *Dialogue* 25 (1986): 471-6