USE OF THESES

This copy is supplied for purposes of private study and research only. Passages from the thesis may not be copied or closely paraphrased without the written consent of the author.
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
STATEMENT

This is to certify that, unless otherwise indicated, this thesis is entirely my own work. It is the result of research carried out by me while a candidate for the degree of Doctor of Philosophy, in the Division of Philosophy and Law, Research School of Social Sciences, at The Australian National University.

Joseph Mintoff
30 April 1993
ACKNOWLEDGMENTS

In the course of writing this dissertation I have incurred many debts of gratitude I would now like to record. My greatest debt is, of course, to my parents, Mary and Frank Mintoff, who supported and encouraged me during the long period of my education, and to my brother, John, and his fiancée, Tracy. My greatest philosophic debt is to Peter Menzies, Frank Jackson and Philip Pettit, whose objections and comments forced me to state more clearly, and to argue more forcefully, the views presented in this dissertation. Thanks also to David Braddon-Mitchell, Andre Gallois, Andrew Gleeson, Dominic Hyde, Robert Sparrow, Michael Tooley, and the other staff and students of the Department of Philosophy, for their friendship, their comments, and their provision of a stimulating environment in which to work. The production of a dissertation requires, however, more than just emotional and academic support - I wish in addition to thank the Commonwealth government for its financial assistance. Finally, but certainly not least, I wish to thank my partner Nicole Gerrand not only for her philosophical contribution to my dissertation, but for her companionship (and that of our cat Mica) over the four year period it took to produce.

Joseph Mintoff
30 April 1993
ABSTRACT

Sometimes, the intentions most promoting one's interests might result in the performance of actions not most promoting one's interests. For example, being the sort of person who intends to keep promises might most promote one's interests – since then others would be more inclined to enter into beneficial agreements with one – even though some of the resulting actions of promise-keeping do not most promote one's interests. A plausible view about the nature of rationality is that the rational intentions for an agent to have are those which most promote their interests, but what are we to say when such intentions result in the performance of actions not most promoting their interests? On the one hand, some say that since such actions are the result of intentions it is plausibly rational to have, then they too must be rational: on this view of the matter, if it is rational to have or to adopt a certain intention, then it is always rational to act upon it. On the other hand, others say that since such actions do not most promote the agent's interests, then they must be irrational: on this view of the matter, since the rational intentions are those which most promote the agent's interests, then surely the rational actions must also always be those most promoting the agent's interests. What, then, is the relation between the rationality of intention and that of action? Addressing this issue is the task for this thesis.
Contents

Statement ii
Acknowledgements iii
Abstract iv
Contents v
Introduction vi

Part I: Rational Irrationality 1

1 The Self-Interest Theory 2
2 Some Toxin Puzzle is a Rational Dilemma 28
3 The Toxin Puzzle – Reply to Objections 49

Part II: Rational Cooperation 77

4 David Gauthier on Rational Cooperation 78
5 Some Prisoner’s Dilemma is a Rational Dilemma 104
6 The Prisoner’s Dilemma – Reply to Objections 128

Part III: Rational Dilemmas 141

Preliminaries 142
7 Michael Slote on Rational Dilemmas 146
8 Derek Parfit on Rational Irrationality 169
9 Gregory Kavka and the Paradox of Deterrence 188

Conclusion 216
Appendix: Parfit’s ‘Endnote 14’ 220
References 227
Rational Intention – Rational Action?

A distinction needs to be made between actions and intentions to act. A distinction therefore also needs to be made between the issue of whether an action maximises expected-value – that is, roughly, whether it most promotes the agent’s interests, or what they value – and the issue of whether an intention to act maximises expected-value.\(^1\) The further distinction needs to be made because, sometimes, the intentions most promoting the agent’s interests might result in the performance of actions not most promoting the agent’s interests. For example, being the sort of person who intends to keep promises might most promote one’s interests – since then others would be more inclined to enter into beneficial agreements with you – even though some of the resulting actions of promise-keeping do not most promote one’s interests. Similarly, being the sort of person who intends to retaliate if encroached upon might most promote one’s interests – since then others would likely refrain from encroaching – even though any particular act of retaliation (if, indeed, any occurs) does not most promote one’s interests. In other words, and not so roughly, the intentions which maximise expected-value may result in the performance of actions which do not maximise expected-value.

A plausible view about the nature of rationality is that the rational intentions for an agent to have are those which maximise expected-value, but what are we to say when such intentions result in the performance of non-expected-value maximising actions? There are two schools of thought on this question. On the one hand, some say

---

\(^1\) This further distinction between the intentions maximising expected-value and the actions maximising expected-value is loosely based on David Gauthier’s distinction between, respectively, agents who choose in such a way as to maximise the satisfaction of their desires and those who choose to maximise the satisfaction of their desires. See ‘The Unity of Reason: A Subversive Reinterpretation of Kant,’ Ethics 96 (1985), p. 85, ‘Reason and Maximisation,’ Can J Phil 4 (1975), pp. 411-2, and ‘Morality and Advantage,’ Phil Rev 76 (1967), p. 461.
that since such actions are the result of intentions it is presumably rational to have, then they too must be rational. On this view of the matter, if it is rational to have or to adopt a certain intention, then it is always rational to act upon it. On the other hand, others say that since such actions are not expected-value maximising, then they must be irrational: on this view of the matter, since the rational intentions are those which maximise the expected-value, then surely the rational actions must also always be those maximising expected-value. The truth, as always, lies in the middle. Given certain conditions (to be specified below), the actions resulting from rational 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 may very well be irrational absent those conditions). Or so I shall argue in this thesis.

[1] I shall be dealing almost exclusively with the notion of rationality, and in particular with only one (quite well-known) theory of rational action. I want to note, however, that this type of issue is of general relevance to all broadly consequentialist, or goal-based, theories of what one ought to do.2

This type of issue is closely related to the debate, in moral theory, between act- and rule-utilitarianism.3 The utilitarian supposes that it is the greatest happiness for the greatest number which is the particularly moral goal. But is such a goal to be pursued directly or indirectly? Take, for example, truth-telling. It may be that being disposed to telling the truth in certain types of situations maximises happiness, even though particular truth-tellings resulting from such a disposition do not. For it may produce most happiness if you are the sort of person who can be believed, and who (within certain limits) tells the truth regardless of the unhappiness this produces. You morally ought to be this sort of person. Yet on some occasions (and

\[2\] This point is made by L. Alexander, 'Pursuing the Good - Indirectly', *Ethics* 95 (1985): 315-32.

within these limits) just such trustworthiness will lead you to tell the
truth even though this does not produce most happiness. Are such
actions moral or not? On the one hand, the rule-utilitarian says that
since such actions are the result of a disposition it is moral for you to
have, then they too are moral. On the other hand, the act-utilitarian
says that since such actions do not produce the most happiness, then
they are immoral. The concerns of this thesis are thus closely related to
the concerns of this well-known debate.

Indeed, we should expect this type of issue to exercise not only
utilitarian theories, but, more generally, all broadly consequentialist, or
goal-based, theories of what one ought to do. This is because the
chance is not necessarily a small one that the dispositions or
intentions which best promote a particular goal (be it a rational, moral,
legal or political goal) are ones requiring the performance of actions
not in themselves promoting that goal. A disposition or intention to
perform some action is, after all, distinct from the action itself, and this
means that such dispositions or intentions have (at least) two types of
effects. First, there are the standard effects, which include the resulting
action and all of its effects. For instance, the standard effects of a
disposition to retaliate if encroached upon will include particular
actions of retaliation (if any), and their effects. Second, there are the
autonomous effects, which do not include the relevant action or its
effects. The autonomous effects of a disposition to retaliate might (and
hopefully will) include successful deterrence; which effect may not be a
function of any actions of retaliation (particularly if no-one ever
encroaches, and so one never has occasion to retaliate). Where one's
dispositions or intentions are somewhat translucent to other agents, or
where due to one's fallibility value is likely to be best promoted
through the guidance of a limited number of relatively simple rules,
the chance is not necessarily a small one that the standard and
autonomous effects of some disposition, intention or rule will differ.4

4 I first came across this way of putting the point in G. Kavka, 'Deterrent Intentions
and Retaliatory Actions,' in D. MacLean (ed.), The Security Gamble: Deterrence
Dilemmas in a Nuclear Age, (Totowa: Rowman & Allanheld, 1984), p. 155, but for
a more detailed taxonomy of the effects of a disposition or intention, see J.
Kilcullen, 'Utilitarianism and Virtue,' Ethics 93 (1983), pp. 452 ff. In claiming the
standard and autonomous effects of an intention or disposition can differ, I claim
(a) Williams's 'Act-Adequacy' thesis is false ('A Critique of Utilitarianism,' in J.
J. C. Smart & B. A. O. Williams (ed.s), Utilitarianism: For and Against,
(Cambridge: Cambridge Univ. Pr., 1973), pp. 119 ff.), and (b) the difference is not
We can expect, then, that the issue of the relation between the rationality of intention and that of action is of broader relevance than just the debate between act- and rule-utilitarians.

Such an expectation turns out to be well justified. It is the type of issue arising in discussions of whether any form of consequentialism can incorporate rights — such as the right to private property, the right to demand fulfilment of one's promises, and so on — having real moral force. Or, indeed, whether any form of consequentialism can positively evaluate traits — such as spontaneity, love, friendship, integrity, and so on — which it seems require a disregard for goal-maximisation. It arises also in discussions concerning the possibility and desirability of an esoteric morality — a morality which requires that it itself not be promulgated. The concerns of this thesis are also closely related to the concerns of these discussions.

[2] The concerns of this thesis, however, are relevant to more than just the in-fighting amongst consequentialists. The relation between the rationality of intention and the rationality of action is also at the heart of David Gauthier's recent attempt, in Morals by Agreement, to take moral scepticism — or, as I shall call it, moral deflationism — seriously.

Moral deflationism takes many forms. On the one hand, the moral deflationist may urge that moral utterances do not (despite appearances) have truth values — they are neither true nor false. The


INTRODUCTION

simplest such view, that of the emotivists, is that to utter some such sentence is merely to express one's attitudes about the matter in hand, either pro or con.9 On the other hand, the deflationist might doubt that there are any true moral utterances at all, even if they admit that such utterances do have truth values. Such is the view of John Mackie, who believes that while moral utterances are either true or false, it turns out that they all are false, since they presume the existence of objective moral objects, and no such "queer" objects do in fact exist.10 On these views, moral talk either lacks truth-value or, if it has a truth-value at all, it is always false.

There is, however, another, and to my mind more troubling, form of moral deflationism. Even if one were to grant that moral utterances have truth values, and even that some of them are true, one might doubt nevertheless that they have the practical force so often attributed to them. Moral considerations, famously, are understood as not being circumscribed by the particular interests or values of the agents to whom they are addressed. It is in the nature of morality that, sometimes, it requires agents to do things they are free not to do, and it has the best outcome for them not to do. But if rational agents are those who act in a way they can expect to maximise what they value, then it seems the only type of consideration providing agents with reasons of any practical force are those which, in some way, are related to the values or interests of the agent. The sad conclusion is that it is not rational to be moral when acting morally does not promote one's interests. One might in this way be lured by the subjectivist threat to morality, and tempted to issue the challenge: why should I be moral?11

How might one respond to the threat posed by this third form of moral deflationism? On the one hand, some will want to repudiate the conception of rationality in general as the maximisation of

expected-value as 'liberal-capitalist ideology'. On the other hand, one might accept for the sake of argument this broad conception of rationality, and then try to argue for the possible rationality of morality within its bounds. I am interested in examining the second approach (being the liberal-capitalist ideologue that I am!), and, in particular, interested in examining David Gauthier's attempt to take this last form of moral deflationism seriously. The exciting task is to show how (or even if) it could be rational to be moral, while accepting the moral deflationist's broad conception of rationality.

I will have more to say about the details of Gauthier's attempt in later chapters, but let me now provide an overview. He argues that, under certain conditions, it is rational to do the moral thing, even if doing the moral thing involves not doing what maximises expected value. His argument has two basic parts.

First is a contractarian analysis of morality: one morally ought to perform some action when it is what one would agree to do were one to employ a rational bargaining procedure, from a rational initial bargaining position, in a situation of perfect information. Gauthier attempts to characterise the rational bargaining procedure (Ch. V), defend the rational initial position (Ch. VII), and argue that the contractarian conception of morality is indeed the correct one (Ch. VIII).

Second is a rationalistic justification for cooperation: when others are sufficiently cooperatively disposed and sufficiently knowledgeable about how each is disposed to behave, it is rational to do what one would in this manner rationally agree to do. To argue this second part, Gauthier begins with a distinction between straightforward maximisers - who are disposed always to do what maximises expected value - and constrained maximisers - who are disposed to comply with rational agreements with those whom they expect would also comply, but are disposed to straightforwardly maximise otherwise (Sect. VI.2.1). Gauthier then argues that if others can know with a better-than-guesswork chance how one is disposed to act, then it may very well be rational to be disposed to constrained rather than a straightforward maximisation (Sect. VI.2.3). He then claims that if it is rational to be disposed to keeping one's agreements with those who one expects also to keep them, it is rational actually to keep one's agreements, when one expects others to do so (Sect. VI.3). He concludes that, on condition that others are sufficiently cooperatively
inclined and sufficiently knowledgeable about how each is disposed to behave, it is rational to be moral.

Doubts have arisen about this analysis of morality. Some have doubted the rationality of his preferred bargaining procedure, others the rationality of his preferred initial bargaining position, yet others the practical import of his assumption of perfect information. Others are unhappy with the contractarian conception of morality as what agents would agree to under certain hypothetical conditions, believing that the moral judgements such an approach delivers fit poorly with our basic intuitions regarding morality. Having neither the space nor the current inclination to consider these objections, I will say no more concerning this conception of morality, and will concentrate only on the rationality of cooperation. For even if morals are not to be defined by agreement, Gauthier will still have said something interesting about the nature of the rationality of making, and keeping, agreements.

Doubts have also arisen about his justification for cooperation. Some doubt the acceptability of Gauthier's definition of constrained maximisation, and others are concerned about the abstraction and idealisation underlying the argument. The doubts I am primarily concerned with, however, are those arising from the fact that, for


Gauthier, cooperation is a non-expected-value maximising action resulting from an expected-value maximising disposition. Are such actions really rational? On the one hand, and as we have just seen, Gauthier thinks so, since he thinks one may move from the rationality of being disposed to keep agreements to the rationality of actually doing so: in his view of the matter, if it is rational to have or to adopt a certain disposition or intention, then it is rational to act upon it. On the other hand, some think not. Gilbert Harman, Gregory Kavka, Derek Parfit, Stephen Darwell and David Lewis all think such actions are yet another instance of so-called rational irrationality. Based on that view of the matter, since the rational dispositions are those which maximise the expected-value, then surely the rational actions must also be those maximising expected-value. They believe constrained maximisation is a rational disposition to perform irrational actions, and that Gauthier has not shown it is rational to be moral.

The truth, as I indicated above, lies in the middle. GIVEN that one rationally ought to adopt, or cannot but have, the enduring disposition to do what one intends (or believes one has agreed) to do, THEN 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 they may very well be irrational absent those conditions. To take the examples with which I started: if it most promotes your interests, and so is rational, for you to become disposed to keep the promises you make, then you are rationally permitted to keep the promises you rationally ought to make; even if you are free not to keep them, and it has the best outcome not to keep them. Similarly, if it most promotes your interests, and so is rational, for you to become disposed to retaliate if you believe you have been encroached upon, then you are rationally permitted to retaliate whenever you rationally ought to believe you

---

INTRODUCTION

have been encroached upon; even if you are free not to retaliate, and
the best outcome results from not retaliating. Or so I shall argue in this
thesis.

If the argument is indeed successful, then some real progress will
have been made in what Sidgwick calls "the profoundest problem in
ethics" – namely, responding to that most deflationary of moral
sceptics, who asks "Why should I be moral?"
RATIONAL IRRATIONALITY

PART I
Chapter One

The Self-Interest Theory

A plausible view about the nature of rationality is that the rational intentions are those maximising expected-value, but what are we to say when such intentions result in the performance of non-expected-value maximising actions? Some say that since such actions are the result of intentions it is plausibly rational to have, then they too must be rational; others say that since such actions are not expected-value maximising, then they must be irrational. Though I will be negotiating a middle way between these views, it is the second I will be particularly concerned to deny. In consequence, the aim of this first chapter is to introduce in more detail the view – which I shall dub the Self-Interest Theory – that all non-expected-value maximising actions are irrational, and to see why, at first blush at least, it is indeed a plausible theory of rational action.

§1 The Self-Interest Theory, Introduced

The Self-Interest Theory, as I shall construe it, is a theory about what one rationally ought, and ought not, to do. Stating the theory is quick enough:

(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.

1 Although referring to it as the 'Self-Interest' Theory, I wish to put to one side the connotations associated with this name. As will become clear momentarily, the theory is one of formal self-interest only. The term is employed in this sense by D. Gauthier, 'The Incompleat Egoist,' Tanner Lectures on Human Values 5 (Salt Lake City: Univ. Utah Pr., 1984): 67-119, and Morals by Agreement, (Oxford: Clarendon Pr, 1986), p. 73.
Not so quick is explaining the theory. In this and the next two sections I shall introduce some of its key terms.

There are a number of relatively minor terms, about which I shall say little. First, when I refer to actions, I mean action-tokens — a particular act of a particular agent at a particular time — rather than action-types. The actions to which I will be referring can be performed at most once, and are not the sort of things which can be instantiated. Second, I shall say an agent is free to perform some action A when they are able to A and are able not to A. I avoid the locution `they can refrain from doing A' since some non-performances of actions are not refrainings. Third, the actions with which I shall be concerned are intentional actions. Exactly what is to count as an intentional action, though, will be a matter for discussion in later chapters; I will leave a more detailed introduction until then. Finally, I shall suppose, in general, that two actions are alternatives to one another only if the agent in question cannot perform both together, and, in particular, that A'ing and not-A'ing are alternative actions. The major concern then, in this thesis, is with the rational evaluation of free intentional action-tokens, and, in particular, with the Self-Interest Theory's evaluation of them.

Two key terms remain: that of what one 'rationally ought' to do, and that of the 'agent-relative expected-value' of an action. I devote to each a section of explanation.

§2 The Rational Ought, Introduced

I will be concerned, in this thesis, with the normative evaluation of action, but there are many ways an action might be evaluated. We evaluate an action in terms of its value, by describing it as good or bad, useful or not; in terms of the attitudes one should have towards it, by saying it is praiseworthy or blameworthy, desirable or undesirable; or in terms of the virtue it displayed or failed to display, by saying it was

---

2 I shall not assume that an intentional action needs to be free, or vice versa. See H. C. Frankfurt, 'Alternative Possibilities and Moral Responsibility,' J Phil 66 (1969): 829-39.

courageous or cowardly, kind or unkind. I am concerned, in particular, with the evaluation we make of an action when we say it is *rational* or *irrational*, but there are also many interpretations of the notion of rationality. Being rational is variously equated with being calm and deliberative, versus being emotional or impulsive; with being cold and impartial, rather than not; or with having one's actions directed by monetary considerations, rather than by more 'human' values. I am concerned, in particular, with the notion of what one *rationally ought* to A, but there are also many understandings of 'ought'. What, then, distinguishes the rational 'ought' from all of these other evaluative notions?

[1] I shall not undertake a survey of all the conditions ever suggested, but will introduce those which seem to me to be central. The distinctiveness of the 'rational ought', in sum, lies in the functional roles judgements using such a notion should, and typically do, play in practical deliberation - that is, deliberation about what to do - and practical action. It is a judgement involving such an 'ought', I want to say, which properly forms the link from practical deliberation to practical action.

The rational 'ought', first, is what I call *deliberative*. The judgement that I rationally ought to do something is one with which my practical deliberations about whether to do that thing properly terminate. In practical deliberation, one is often confronted by considerations each recommending the performance of actions all of which cannot be done. You might have a reason to save the valet from drowning, because he is your father; and yet also have reason to save Fenlon from drowning, since he is a famous author and archbishop. In such a situation, you want to say 'yes, but what ought I really to do,'

---

4 See S. F. Sapontzis, 'The Obligation to be Rational,' *J Val Inq* 13 (1979), pp. 295 ff. for some discussion of the following, and other, interpretations of the word 'rational'.


6 What I am calling the *deliberative* aspect of the rational 'ought' is also identified, for example, by B. Williams, 'Ethical Consistency,' *Aris Soc Supp Vol 39* (1965), p. 108, and R. Trigg, 'Moral Conflict' *Mind* 80 (1972), p. 52. Note I do not claim that all actions require deliberation, only that deliberation, when it occurs, is properly terminated by a conclusion employing a 'rational ought'.

save the valet or the archbishop, granted I cannot do both? You want a word to express a final decision, and 'ought' seems the appropriate one. When such is the meaning of the word 'ought', it is being used in what I have called a deliberative sense. For example, it is irrational both to judge you rationally ought to save the archbishop, and yet continue with practical deliberation about whether to save the archbishop.

This sense of the word 'ought' should not be confused with others closely related. We want to say, and often do, that you ought to save the valet, because of the special relationship he bears to you, and yet ought to save the archbishop, because of his greater importance to the community. To ask the question 'Yes, but what (in the deliberative sense) ought I to do?' is not to ask either 'What do the considerations concerning my special relations to others dictate I ought to do?' or 'What the considerations concerning the general good dictate I ought to do?' You already know what the first dictates — saving the valet; you already know what the second dictates — saving Fenlon. You want to know what you, in the end, ought to do.8

The rational 'ought', second, is what I call (strongly) action-guiding. The judgement that I rationally ought to do something is one which my doing it properly follows.9 Consider again Fenlon and the valet. After some deliberation you come to the judgement that you rationally ought to do what the common good demands — namely, saving Fenlon. Now that deliberation is over, all that remains is to act. Understandably, though, you save the valet — your father — instead. Not only do you feel guilty for (let us suppose) doing something immoral, but are by your own lights irrational, since you failed to do that thing — saving the bishop — which you thought you rationally ought to do. That you rationally ought to A is the sort of statement it is irrational both to judge to be so and yet fail to do what it demands.

---


This feature of the rational 'ought' needs to be distinguished from another one closely related. For

[even if it is true that 'ought' has often to be used at two different stages in situations of moral conflict, it is not true that it only preserves its action-guiding function at the level of actually deciding what we are to do. ... if either of the 'oughts' in 'I ought to [save my father] and 'I ought to [save the bishop]' ... had nothing to do with the recommendation of action, it is very strange that we find ourselves at all worried about what we should.]

We can explicate the distinction as follows. My heart, we have seen, recommends saving my father, but the common good demands saving the bishop. On the one hand, the fact that the valet is my father and Fenlon the bishop, are action-guiding considerations in a weak sense, in that both should have some causal influence over the action I perform; I should be moved at least to some degree by each. On the other hand, the consideration concerning the common good is, and that of the special relation to my father is not, an action-guiding consideration in my strong sense, in that the first should, whereas the second should not, have decisive causal influence over the action I perform: though I should be moved at least to some degree by each, it is the first which should be decisive. It is the aim, so to speak, of both considerations to guide action, but the one I shall call strongly action-guiding is the one that should prevail.

[2] To these two necessary conditions making for the distinctiveness of the rational 'ought', there is, in addition, a third. The practical 'ought' is what I shall call absolute, or unqualified. There are a number of different qualifications I mean to exclude by this claim.

First, the rational 'ought' is not relative to evidential considerations. A judgement involving the notion is, I shall say following Davidson, not an all-things-considered, but rather an all-out,
judgement. It is one thing to judge something simply as what you ought to do, and quite another to judge it as what you ought to do relative to some fact or facts. A distinction needs to be made between those practical judgements which are relative to evidence, and those which are not. So, for example, to judge that, all-things-considered, you ought to A, is effectively to judge that relative to all the available relevant evidence you ought to A. The contrasting all-out judgement, that you (simply) ought to A, is, of course, typically made on the basis of a previous all-things-considered judgment about what you ought to do, but still needs to be distinguished from it.

Second, the rational 'ought' is not conditional on the desires or aims of the agent. It is, I shall say, following Kant, not a hypothetical but rather a categorical notion. It is one thing to judge something simply as what you ought to do, and quite another to judge it as what you ought to do given, for example, that you want to keep your honour. A distinction needs to be made between those practical judgements which are relative to one's desires and aims, and those which are not. So, for example, to judge that, given your desire to keep your honour you ought to kill so-and-so is effectively to judge that the best possible way to keep your honour is for you to kill so-and-so. The contrasting categorical judgement, that you (simply) ought to kill so-and-so, will (hopefully) differ in truth value from the hypothetical judgement about what you ought to do.

Finally, the rational 'ought' is not qualified by various epithets, such as 'morally' ('prudentially', 'legally') or suchlike. As we have seen, it is one thing to judge something simply as what you ought to do, and quite another to judge it as what, for example, you morally


13 Others who insist on such a condition are, for example, B. A. O. Williams, *Ethics and the Limits of Philosophy*, (London: Fontana, 1985), pp. 5-6, and H-N. Castenada, *The Structure of Morality*, (Springfield, Ill.: Thomas, 1974), pp. 177, 192 ff.
ought to do. A distinction needs to be made between those practical judgements which refer to some distinctive normative category, and those which do not. For example, to judge that you morally ought to donate your spare cash to Oxfam is effectively to judge (on certain common views about the nature of morality) that from the impartial perspective you ought to do this. The contrasting unqualified judgement, that you (simply) ought to donate the money, may differ in truth value from the moral judgement about what you ought to do.

In sum, then, to make a judgement that you rationally ought to do something is to make a judgment satisfying three requirements: (a) it is deliberative – it would be irrational for you to judge you rationally ought to do something and yet still deliberate about whether or not to do it; (b) it is (strongly) action-guiding – it would be irrational for you to judge you rationally ought to do something and yet not do it. And, finally (c) it is absolute, rather than relative – to make a judgement that you rationally ought to do something is to make a judgement not relativised to evidence (though typically made on the basis of evidence), not relativised to desires or aims (though such desires will be related to what you ought to do), and not relativised to moral or prudential norms (though what you ought to do may well be moral or prudent). While there are indeed many ways we might evaluate some action, many meanings the notion of 'rationality' may have, and many different ways of understanding what one 'ought' to do, these three features of what I have called the rational ought ensure it is a notion deserving of a central place in our ethical theorising.

§3 Expected Value, Introduced

Like other theories of rationality, what I am calling the Self-Interest Theory is concerned with identifying some property distinguishing those actions an agent rationally ought to do from those they are rationally permitted not to do. In particular, it says an agent rationally ought to perform some action, which they can perform, if and only if the agent-relative expected-value of that action exceeds that of any alternative possible action.

What is the 'expected-value' of an action? To answer this, suppose Q is a partition of states-of-affairs – that is, a set of states-of-affairs mutually exclusive and jointly exhaustive – interpreted as possible outcomes of the agent's doing A. Denote by $\mathcal{O}_X(A,q)$ the claim that q
THE SELF-INTEREST THEORY

would be (part of) the outcome of x's doing A (were x to do A). Suppose that associated with the agent, x, is a set of expectations regarding what the outcome would be of x's doing A (were x to do A), and denote by \( P_x(O_x(A,q)) \) the agent-relative probability, \( P_x \), that q would be (part of) the outcome of x's doing A (were x to do A). Suppose further still that associated with the agent, x, is a set of values regarding the possible outcomes of action, and denote by \( V_x(q) \) the agent-relative value of outcome q. (I will have something to say about these 'associations' presently.) The expected value x's doing A, denoted by \( EV_x(A) \), is then given by the expression:

\[
EV_x(A) = \sum_{q \in Q} P_x(O_x(A,q)) \cdot V_x(q)
\]

The expected-value of an action is therefore an agent-relative-probability weighted-sum of the agent-relative-values of the possible outcomes of the action.

It is important to keep in mind that the notion of the expected-value of an action, in this way defined, is conceptually distinct from that of the value of an action. To illustrate this, consider the following example. The notion of water, as this is often understood, is conceptually distinct from that of \( H_2O \). Water may indeed be \( H_2O \) (though, I suppose, there must have been some dispute about this at one time), but not because 'water' means 'H2O'. To tell whether or not water is \( H_2O \), someone had to do more than look solely at the meanings of the terms – they had to do some chemistry. As for water and \( H_2O \), so for the value, and the expected-value, of an action. The value of an action may very well be its expected-value (as the Self-Interest Theory states, though I shall argue it is not), but not because 'value' means 'expected-value'. To tell whether or not value is the same as expected-value, we will have to do more than look solely at the meanings of the words – we will have to do some moral philosophy.

This definition of the expected-value of an action as a probability-weighted sum of values of outcomes is familiar in form. Its central features I take to be its consequentialism and its agent-relativism. I will briefly consider each of these features in turn.
(1) Expected-value is, first, a consequentialist notion: it is characterised in terms of the outcomes or consequences of the agent's performing the action, and, in particular, in terms of the values of outcomes.\(^{14}\) There are, though, many ways the notion of an outcome may be understood.\(^{15}\) However it is to be understood, it is necessary that it satisfies the following requirement: (1) an event \(q\) would be (part of) the outcome of \(x\)'s doing \(A\) only if were \(x\) to \(A\) then \(q\) would obtain.

Such a condition is too weak, and needs to be modified. To see this, suppose you are the sort of person who would take the garbage out only if you had previously promised to do so. What is true of you is that if you were to take the garbage out tonight, then you would have made a promise to do so in the morning. It is implausible to suggest, though, that your promise this morning is (part of) the outcome of taking the garbage out this evening, since the time of the putative outcome (namely, your promising) is earlier than that of the relevant action (namely, your taking the garbage out). This suggests that the relevant subjunctive not be purely back-tracking. It will do for our purposes if we say that the subjunctive \(\text{p} \rightarrow \text{q}\) is purely back-tracking when the time of the event referred to by the consequent (\(q\)) is earlier than that of the event referred to by the antecedent (\(p\)). However the notion of an outcome is to be understood, then, it seems necessary that it satisfies the following requirement: (1') an event \(q\) would be (part of) the outcome of \(x\)'s doing \(A\) only if were \(x\) to \(A\) then \(q\) would obtain, and this subjunctive is not a purely backtracking one.

Some may be inclined to think that the condition is still too weak, and that it needs to be modified further.\(^{16}\) To see why one might be

---

\(^{14}\) Unfortunately, the term 'consequentialism' has become associated with the view that (a) what one ought to do is to be determined by outcomes of actions, and (b) in particular, by the agent-neutral value of those outcomes. See S. Scheffler, *The Rejection of Consequentialism*, (Oxford: Clarendon Pr., 1982), pp. 1-2, and D. Parfit, *Reasons and Persons*, (Oxford: Clarendon Pr., 1984), p. 27. I shall, however, be using the term as I believe it should be used, to denote only condition (a).


\(^{16}\) J. J. C. Smart, in 'An Outline of a System of Utilitarian Ethics' in J. J. C. Smart & B. A. O. Williams (eds.), *Utilitarianism: For and Against*, (Cambridge: Cambridge Univ. Pr., 1973), p. 13, stipulates that keeping a promise is not the sort of thing which can be part of the state of affairs which is the outcome of an action, lest his own position be identical to that of W. D. Ross. Amongst those requiring the causal connection discussed in this paragraph is G. E. Moore, 'A Reply to my Critics,' in P. A. Schlipp (ed.), *The Philosophy of G. E. Moore* (Evanston and Chicago:
inclined to this view, note that condition (1) does not exclude the proposition that your keeping the promise is (part of) of the outcome of your taking the garbage out. Since if you were to take the garbage out then you would keep the promise, and since this is not a purely back-tracking subjunctive (the time at which you keep your promise is, presumably, exactly the time at which you take the garbage out), then, for all condition (1') says, your keeping your promise might be (part of) the outcome of taking the garbage out. Some find this unsatisfactory because they require the putative outcome to be in some way a causal outcome of the action, which, in this case, it is not. Taking the garbage out does not cause you to keep your promise, rather it just is for you to keep your promise. However the notion of an outcome is to be understood, then, such persons would require that it satisfies the condition: (1'') an event q would be (part of) the outcome of x's doing A only if were x to A then x's doing A would at least causally contribute to q's obtaining.

I shall not, however, strengthen the condition in this further way, and shall stick with (1') as the explication of the notion of an outcome for this thesis. Thus: if event q and action A fail to satisfy the necessary condition mentioned in (1'), I will say q would not be (part of) the outcome of A; if event q and action A do satisfy this necessary condition, then, as far as (1') is concerned, q may or may not be (part of) the outcome of A. I will stipulate in the second case - solely for the sake of argument - that q would be (part of) the outcome of A. While I believe that a condition stronger than the one mentioned in (1') does obtain, I shall not pursue the matter further. In the context of this thesis this stipulation is more than justified, for I shall be arguing that there are more considerations determining what we ought to do than just those mentioning outcomes, and so (as we will see) I only make things more difficult for myself by adopting this very inclusive notion of outcome.

[2] The expected-value of an action is, second, an agent-relative notion, and doubly so: it depends on expectations and values, both of which are in some way 'associated' with the agent in question. To illustrate

the forms such an 'association' may take, concentrate on the case of values. The values associated with an agent may be the ones they actually have concerning outcomes; the ones they subjectively ought to have concerning outcomes; or perhaps the degree to which the outcomes satisfy the objective well-being of the agent. However it is to be understood, though, the required 'association' can be captured formally as follows: (2) the values relevant to the rationality of x's doing A (denoted, recall, by $V_x(q)$) are two-argument functions of the pair consisting of the agent x and a possible outcome, q, of A.

Some may be inclined to think this condition too weak, and insist that it needs to be supplemented. To see what they have in mind, a distinction needs to be drawn between instrumental and intrinsic value. On the one hand, agent-relative value may be instrumental: a state-of-affairs, like having one's tooth removed, may derive whatever value it has from the value of its consequences. But, on pain of an infinite regress, not everything can have value solely on the basis of the value of its outcomes. On the other hand, then, value may be intrinsic: a state-of-affairs, such as enjoying oneself on a picnic, may derive its value not solely on the basis of the value of its consequences, but rather because it is an instance of a particular type. With this distinction in mind, some will insist that certain states-of-affairs, such as promise-keepings and intention-fulfillings, are not to be intrinsically valued by any agent deserving of the label 'consequentialist'. On this view, the consequentialist aspect of the Self-Interest Theory consists in more than just the view that the rationality of action depends on outcomes, but also in the agent's not holding certain values. The proper consequentialist, in particular, sees promise-keeping and intention-fulfilling as being valuable, if at all, only instrumentally. However value is otherwise to be understood, then,

---

17 See, for example, R. Brandt, 'The Concept of a Rational Action,' Soc Theor Pract 9 (1983), pp. 143-164, esp p. 161, and D. Gauthier, Morals by Agreement, Ch. II.
such persons would require that it satisfy the condition: (2') the values relevant to the rationality of x's doing A are two-argument functions (such as $V_x(q)$) of the pair consisting of the agent x and a possible outcome, q, of A, such that x assigns no intrinsic value to promise-keeping or intention-fulfilling.

I am inclined to think the Self-Interest Theory needs this further condition. It seems that the defender of Self-Interest Theory cannot avoid a charge of triviality if they are not at pains to distinguish their position from that of a deontologist such as Ross – who thinks that (some) promises are, in and of themselves, sufficient reasons for action – by insisting promise-keeping is not intrinsically valuable, and not intrinsically valued by a consequentalist agent. After all, Ross thinks we should keep promises simply because they are promises, or, as he might put the point, since promise-keeping is intrinsically valuable. If the consequentalist aspect of the Self-Interest Theory is to be distinguished from Ross's, then, it must be supposed that promise-keeping is not intrinsically valuable. I am not certain, however, the Self-Interest Theory needs this further condition. As a result, I will be careful to show, when relevant, that if the Self-Interest Theory is unacceptable when this requirement is made, then it is also unacceptable when it is not.

The expected-value of an action is a probability-weighted sum of values of outcomes, and its central features I take to be its consequentialism and its agent-relativism. The idea underlying this definition is simple and attractive: there are many things possibly resulting from any action, and the influence a certain possible outcome should have in determining the expected-value of the action which might produce it is directly proportional to the expectation that that outcome will in fact result from that action, and also to the value of that outcome. The idea behind the Self-Interest Theory, then, is equally simple and attractive: what one rationally ought to do is to be determined by the possible outcomes of one's actions, one's expectations concerning the likelihood of such outcomes, and the value one attaches to such outcomes.
§4 The Self-Interest Theory, Reinterpreted

In this thesis I will argue it is rational, for example, to cooperate in certain types of situations; but to do this, it turns out I will need to argue also that one has reason to cooperate in these situations. The formulation of the Self-Interest Theory I have just introduced is very close to the standard formulation, and is a theory of the rationality of action; therefore I will need to reformulate it also as a theory of reasons for action. This is what I shall do in this section.

[1] A plausible view about the nature of rationality is that what one ought to do is related both to expectations concerning, and strengths of, the reasons one has for acting. For example, I am at a loss about what to do this afternoon, and begin to deliberate about the matter. After some thought, I reason to myself:

(1) I promised to visit my mother.
(C) So I will visit my mother.

I have found a reason for me to act, and can use it to determine what I ought to do. One might thus be tempted by the suggestion that a person rationally ought to perform some action if and only if there is some reason for them to do so. This suggestion, though, would be mistaken, for it fails to allow for two complexities.

First, on many occasions, there are conflicting reasons for action, and determining what one rationally ought to do will involve considering the strength of one's reasons. On the way to my mother's

---

21 I shall say nothing about what reasons are, and shall take the notion as given. One thing it is important to note, however, is that I make no commitment about how we come to be guided by the reasons we have for action. In particular, I keep distinct the idea of an action's being guided by deliberation and its being guided by reasons. For example, E. J. Bond, in 'Reasons, Wants and Values,' Can J Phil 3 (1974), pp. 3, 15, distinguishes between acting for reasons and acting in accord with deliberation. See also fn. 6, above.

house, out in the country, I come across an overturned car on a deserted stretch of road. An accident has happened, and things look bad indeed for the victim unless I get him to the nearest hospital, many miles out of my way. He won't die, but is in much pain, and needs treatment soon. After some thought, I reason to myself:

(1) I promised to visit my mother; but if I do, this person would suffer needlessly.
(2) Preventing this person's suffering is in itself more important than my promise to my mother.
(C) So I will not visit my mother.

There are now two reasons for me to act. I promised to visit my mother — this is still a reason for me to do so. But if I visit my mother, then this accident-victim would suffer needlessly — this is a reason for me not to visit my mother. I have now two reasons to act, and must use them to determine what I ought to do. Plainly, I should go with the stronger reason. In the end, the strength of the reason that this person would needlessly suffer is greater than the strength of the reason that I promised my mother to visit her. Thus, I determine I ought not to visit my mother — by determining whether the strength of the reason for doing so it exceeds that for not doing so. What one rationally ought to do is logically determined by two types of facts: the reasons there are for or against certain actions, enumerated in premise (1) above; and the comparative (intrinsic) strengths of these reasons, enumerated in premise (2).

Second, this very popular conception of what one rationally ought to do is still, however, not quite right, for on some occasions, it is not certain that a particular reason obtains, and determining what to do will involve considering expectations concerning one's reasons. I am about to move the accident-victim into my car, to take him to the nearest hospital, when a local farmer comes along in his somewhat weather-beaten truck. He sees what has happened, asks me what I am doing, and, when I tell him, kindly offers to take the accident victim instead, since (he says) he was heading that way in any case. I cast an

---

24 This point is less remarked upon. One who does touch on this idea, however, is T. Nagel, The Possibility of Altruism, (Oxford: Clarendon Pr., 1970), p. 55.
eye over his truck, and come to the conclusion that there’s some chance it might in fact break down after I have gone on my own way, and if that happened the accident victim would be back in his original predicament. After some thought, though, I say to myself:

(1) I promised to visit my mother; but if I do, this person would suffer needlessly.

(2) Preventing this person’s suffering is in itself more important than my promise to my mother.

(3) But there is only a small chance this person would suffer needlessly, and a high chance I have in fact promised my mother.

(C) So I will visit my mother.

I still have the two original reasons: that I promised to visit my mother (if there is some chance this obtains) is still a reason to do so, and that the person would suffer needlessly were I to visit my mother (if there is some chance this obtains) is still a reason for me not to do so. I still retain my original ranking of the (intrinsic) strengths of these two reasons: the consideration that this person would suffer is stronger, in and of itself, than the consideration I promised to visit my mother. What has changed, however, is that there is only a chance that if I were to visit my mother then the person would suffer needlessly. The two reasons for me to act are not both certain to obtain, but I must still use them to determine what I ought to do. Plainly, I should go with the reason expected to be stronger. In the end, I decide that even though the consideration that this person would suffer is of greater (intrinsic) strength than the consideration that I promised my mother to visit her, the only slight chance the first in fact obtains means that the expected strength of the consideration that this person would needlessly suffer is less than the expected strength of the consideration that I promised my mother to visit her. Thus, I determine that I ought to visit my mother - by determining whether the expected strength of the reason for doing so it exceeds that for not doing so. What one rationally ought to do is thus logically determined by three types of facts: as before, the reasons there are for or against certain actions, enumerated in premise (1); as before, the comparative (intrinsic) strength of these reasons, enumerated in premise (2); and, in addition, the expectations the reasons in fact obtain, enumerated in premise (3).
[2] We see, then, that whether or not to perform an action is some function of the strengths and expectations of the reasons there are for or against that action. But what function in particular?

The practical deliberations above concerning whether to visit my mother can be generalised. There, we came to the conclusion that I ought to visit my mother if and only if the expected strength of the only reason I had for doing so (namely, that I promised) is greater than that of the only reason I had for not doing so (namely, that the accident victim would suffer needlessly). There, I had to deal with only two reasons, one for and one against. But when a person faces more conflicting reasons than just two, a more complex account is required. It is easy to see, though, 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.

Note that we are concerned with the expected strength of reasons for or against the action, just as before, when I introduced the consequentialist aspect of the Self-Interest Theory, we were concerned with the possible outcome, good or bad, of the action.

What is the 'expected strength of reasons' for or against some action? To answer this, suppose we symbolise by 'p.R.A' the fact that p is a reason for or against the agent's performing some action, A. Suppose further it is possible to assign some probability, P(p), to p's obtaining, and that it is possible to measure the strength of this reason for or against the agent's doing A by some number, SR(p,A). The natural suggestion for the expected strength of the reasons for or against doing A, denoted SR(A), is thus:

\[ SR(A) = \sum_{p.R.A} P(p).SR(p,A) \]

The summation is performed over the set of reasons, p, for or against doing A. The expected-strength of reasons for or against some action,
then, is a probability-weighted sum of the possible reasons for or against performing that action.\textsuperscript{25}

What should be clear, in light of all this, is that the dispute between theories of rationality should not hinge on whether or not reasons can be given for actions, or whether or not expectations concerning, and the (intrinsic) strengths of, these reasons need to be taken into account. Any adequate theory will admit as much. The debate between them hinges, rather, on the issues: (1) exactly what sorts of things give us reasons for action? (2) what sort of expectations are relevant in determining how much weight is to be given to a particular reason? and (3) how is one to go about determining the (intrinsic) strength of any particular reason for action? It is their responses to these questions that distinguishes one theory from another. In particular, it is in the responses to these questions in terms of which the Self-Interest Theory should be understood.

[3] What specific answers does the Self-Interest theory provide to these questions? Its answer to the first question makes it a consequentialist theory; its answer to the remaining two, an agent-relative theory:

\textbf{(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 \( q \).

On this specific theory of reasons for action, the expected strength of reasons for or against an action, \( A \), is given by:

\[
SR_x(A) = \sum_{q \in Q} P_x(O_x(A,q)).V_x(q)
\]

The general claim, (R), and the distinctly Self-Interested view about the nature of reasons, (S'), entail that the strength of the reasons for or

\textsuperscript{25} There are a number of questions this account raises: (a) how are we to specify the considerations, \( p \), over which the summation is performed? (b) how, if at all, does this account of expected-strength of reasons for action generalise for beliefs, desires, and so on? These are issues beyond the precincts of this thesis.
against some action is precisely the (agent-relative) expected-value of that action (i.e., \( \text{SR}_x(A) = \text{EV}_x(A) \)), and therefore entails the original formulation (S) of the Self-Interest Theory, that one rationally ought to do A when and only when it maximises expected-value to do so.

The central tenets of the standard and common formulation of the Self-Interest Theory are **consequentialism** and **agent-relativism**, and have their expression in statement (S). Yet this statement merges two central ideas it is best to keep separate. On the one hand, there is the general idea – expressed by (R) – 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. On the other hand, there is the specific idea – expressed by (S’) – that all and only considerations about outcomes provide reasons for action, and that the only relevant expectations and values are the agent’s own. The second idea is what captures, unmixed with others, the two central tenets of the Self-Interest Theory, and is therefore to be preferred as the formulation of the theory.

§5 Why the Self-Interest Theory?

Why, though, would anyone believe the Self-Interest Theory? In this last section, I will concentrate in turn on each of its central claims, and indicate very briefly (and with no claim to completeness) some independent arguments in support of each.

[1] The Self-Interest Theory’s first claim is that 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 \( A \)’. One half of this claim seems obvious: *all* considerations of this form are reasons. How could the fact that some valued or disvalued event would result from an action not be relevant to whether or not one should do it? The first half expresses what seems to me most plausible in consequentialism. The other half is not so obvious: *only* considerations of this form are reasons. Surely there are more reasons to act than just
those considerations concerning outcomes! The second half expresses something less plausible in consequentialism.\(^{26}\)

Promises seem to provide the clearest instances of non-consequentialist reasons for action.\(^{27}\) Suppose I promise to visit my mother. If the value of the outcome of keeping my promise matched precisely the value of the outcome of not keeping it, then, many would feel, I ought to keep the promise. Since the value of the outcomes balance, what extra reason could I have for keeping it? Simple – the fact I promised. The objector to the consequentialist supposes that my reasoning process is something like this:

(1) "If I visit my mother, then Q\(_1\) would obtain; if I do not, then Q\(_2\) would obtain; Q\(_1\) and Q\(_2\) are equally valuable; but I promised to visit my mother. So, I ought to visit my mother."

The objector supposes that the fact I promised to visit my mother is a different type of reason from that provided by listing the outcomes of visiting my mother, or not. This explains why we feel, when the values of the outcomes are equal, that I nevertheless ought to keep my promise.

How might the consequentialist respond? On the one hand, they may accept the intuition, but claim that, contrary to appearances, there is an extra consequentialist reason for keeping the promise. On this approach, the consequentialist accuses the objector of having left something out of the description of the outcome of my keeping (or breaking) the promise – namely, that a promise would be kept (or broken) – and supposes I value promise-keeping intrinsically. The consequentialist supposes my reasoning process is really something like this:

\(^{26}\) J. S. Mill, in 'Utilitarianism', reprinted in M. Warnock (ed.), *Utilitarianism*, *On Liberty*, *Essay on Bentham*, together with *Selected Writings of Jeremy Bentham and John Austin*, (London: Collins, 1962), opines that '[n]or is there any school of thought which refuses to admit that the influence of actions on happiness is a most material and even predominant consideration in many of the details of morals, however unwilling to acknowledge it as the fundamental principle of morals, and the source of moral obligation.' (p. 154)

(2) "If I visit my mother, then Q₁ would obtain and a promise kept, if I do not, then Q₂ would obtain but a promise broken; Q₁ and Q₂ are equally valuable, and, others things being equal, keeping a promise is better than breaking it. So, I ought to visit my mother."

Consequentialists suppose the fact that I promised to visit my mother is exactly the same type of reason to that provided by listing the (other) outcomes of visiting my mother. They will reject the objector's rather narrow understanding of the outcomes. If, other things equal, I intrinsically value states-of-affairs including promise-keepings to ones including promise-breakings, then I will have all the (consequentialist) reason I need for keeping my promise to my mother, which is what intuition counsels about this case.

On the other hand, they may just deny the intuition, and say I value promise-keeping, (or, at the least, I should value promise-keeping) only in an instrumental fashion. Recall I suggested above that, to avoid triviality, a consequentialist will need to insist that promise-keeping and intention-fulfilling are not intrinsically valuable. Other things being equal, they should say that keeping a promise is no more valuable that breaking it; in this case, other things are equal; so there is no more reason to keep the promise than there is to break it. Typically, of course, other things are not equal, since making a promise usually results in others expecting you to keep it; which expectations would be frustrated were you not to keep the promise. Either way, then, the consequentialist seems to have more than enough resources to deal with this objection.

Furthermore, it seems the consequentialist's way of understanding how promises provide one with reasons to act is superior. You value promise-keeping, and face a decision now about whether to keep a certain promise. Unfortunately, though, it turns out that if you now


29 This example is based on B. William's case of 'Jim and the Indians', in 'A Critique of Utilitarianism', pp. 98 ff. It differs, however, in two respects: where Williams is concerned with what it is moral to do, I am concerned rather with what it is rational to do; and where he stipulates that the unfortunate effects of action are due to the actions persons other than the agent would later perform, I stipulate
kept this promise then you would, freely and intentionally, later break many promises, and that if you did not now keep just this one promise, then you would later, freely and intentionally, keep these promises. The issue is whether one ought to keep this one, initial, promise, and it seems the consequentalist provides the correct answer. On the one hand, someone who thought promise-keeping was overridingly important, and who took the fact that they had made a promise into consideration in the way the objector above suggests, would reason as follows:

(1') "I promised to A now. So, I'll A."

They would keep their promise now, but (by hypothesis) would then later break many more promises. On the other hand, someone who thought promise-keeping was overridingly important, but who took the fact that they had made a promise into consideration in the way the consequentalist suggests, would reason as follows:

(2') "If I were to A, then I would keep my promise to A but break many promises later, if I were not to A, then I would break my promise to A but would keep many promises later. It is better to keep more promises than less. So I'll not A."

The objector, it seems, will have one keep the initial promise to do A (and thus break many more later promises), and the consequentialist will have one break the initial promise to do A (and thus keep many more promises later on). But if one really values promise-keeping, then it seems the second is, intuitively, what one ought to do.

What seems clear, then, is that insofar as considerations such as promise-keeping give one reasons for action, they do so in the outcome-orientated way the consequentialist advocates, rather than the direct way the objector advocates. Not only do all consequentialist considerations provide reasons for action, but they also seem to provide the only reasons for action.

that the unfortunate effects are due to the actions the agent themselves would later perform.

30 This is the way Williams suggests the non-consequentialist might reason. See 'A Critique of Utilitarianism', pp. 89 ff.
[2] The Self-Interest Theory's second claim is that the relevant expectations concerning what the outcome would be, were an action performed, are those in some way associated with the agent whose action it is, though I have not committed myself (and nor will I do so) to the precise form such an association might take.

One form the agent-relativity of expectations might take is that of subjectivism — the claim being that the expectations relevant to rational action are the degrees of belief, or credences, the agent actually has concerning the possible outcomes of their action.31 To see what reason one might have for thinking this, take an example of David Lewis's. There is an epidemic, all the antitoxin is locked in your safe, and you've lost the combination. You maximise subjectively expected-value by sending for the safecracker at once. Time is precious — don't waste it trying out combinations and hoping to be lucky! But this definitely is not what it is actually best to do, due to the time lost before the safecracker turns up. There is a certain combination — the correct one, which ever that might be — such that what it is actually best to do is to dial it immediately. Thus what one rationally ought to do depends, as the Self-Interest theory suggests, on the subjectively expected outcomes of an action, and not on its actual outcome.

A second form the agent-relativity of expectations might take is provided by a position called rationalised subjectivism — which suggests that the relevant expectations need themselves to be of a minimal level of acceptability if they are not to render irrational those actions performed on their basis.32 Where the subjectivist requires that action be rational according to the beliefs an agent actually has concerning the possible outcomes of that action, the rationalised subjectivist requires that action be rational according to the beliefs the agent rationally ought to have concerning those outcomes.

But surely the expectations relevant to rational action are the ones which actually obtain, and not the ones an agent believes obtain, or rationally ought to believe obtain! One might think — with the so-


called objectivists – that the expectations relevant to rational actions are all objective chances concerning the possible outcomes of the agent's actions. And one might reply to the example two paragraphs above by drawing a distinction between two types of 'ought'. The objectivist offers an account, on the one hand, of what one ought to do by the lights of someone with perfect information: this is an account of the objective-ought. The subjectivist offers, on the other hand, an account of what one ought to do by the lights of the information the agent has at the time of action: this is an account of the subjective-ought. The two views, one might suggest, thus talk past each other, since they present differing analyses of differing notions. This will not do, however, since even if the objectivist provides an adequate characterisation of what they call the objective-ought, it is clearly inadequate as a characterisation of central notion of concern – what I am calling the rational-ought. Consider again Lewis's example. You know that going for the locksmith does not actually have the best outcome, and so you know you objectively-ought not to go for the locksmith. Yet going for the locksmith is what you do. It is clear, though, that you are not irrational to both judge that you objectively-ought not to get the locksmith, and yet get the locksmith. The rational-ought, however, is – as we saw in section §2[1], above – strongly action-guiding: it is irrational to both judge that you rationally-ought not to A and yet A. Hence, the objective-ought, whatever else might be said for it, is not the rational-ought. The objectivist's reply is simply not to the point.

What seems clear, then, is that there are good reasons for supposing the expectations relevant to rational action are, as the Self-Interest Theory claims, in some way relative to the agent. An agent can act only on the basis of expectations they have at hand, or rationally ought to have at hand.

[3] The Self-Interest Theory's third, and final, claim is that the values attached to the possible outcomes of an action, like the expectations, are those in some way associated with the agent whose action it is, though I have also not committed myself (and nor will I do so) to the precise form such an association might take.

One form the agent-relativity of values might take is that of subjectivism (or of rationalised subjectivism). For just as one might think the expectations relevant to rational action are subjective, so one might think the relevant values are also. In the previous sub-section I introduced an example which gave strong reasons for supposing that expectations are subjective. But it would be strange – would it not? – if expectations needed to be subjective but that values did not. It is not right to expect agents to act on the basis of facts (namely, the objective ones) to which they may have no access at all, to which, even with deliberation and reflection, they cannot come to discover in time for them to act. It would be strange – would it not? – if the same did not apply to values. It is not right to expect agents to act on the basis of values (namely, the objective ones) to which they may have no access at all, to which, even with deliberation and reflection, they cannot come to discover in time for them to act.

A second form the agent-relativity of values might take is provided by a position sometimes called internalism. This is the view that there is some sort of necessary connection (different on different versions) between what the agent ought, or has reason, to do, and what actually motivates the agent. As Nagel puts the point: "[i]nternalism's appeal derives from the conviction that one cannot accept or assert sincerely any ethical proposition without accepting at least a prima facie motivation for action in accordance with it." Since, it may independently be argued, only the concerns or interests – generically, the desires – of the agent themselves could motivate them to act, it is only they, so the argument goes, that provide reasons for action, and determine what an agent rationally ought to do.

There is a third form of value agent-relativism. Subjectivism and internalism do not deny there could be such things as objective values, but claim merely that they would be irrelevant; the more radical theory, however, claims that there really is nothing valuable independently of agents. On this radical view, values are essentially

---

34 I am unaware of anyone adopting this line. For reasons that will become apparent in the text, though, I think it is a line worth pursuing.

agent-relative. In Gilbert Harman's version of the view, the best explanation for our actions and choices, so the story goes, will make no reference to either objective or agent-neutral value, and since there is reason for postulating the existence of an entity only when it is explanatorily useful to do so, there is therefore no need to postulate the existence of objective or agent-neutral values. All can be explained solely in terms of agent-relative value. In John Mackie's version of this type of argument, the claim is that such objective values would have to be understood as some sort of non-naturalistic, or moral, property of actions, and that such properties would be 'queer'. There would again be no reason for postulating them, since the presence of moral practices can be explained more parsimoniously on sociological grounds. On these views, objective value is not to be found among the metaphysical furniture of the universe.

I will not enter into the finer details of these arguments, but simply use their existence to suggest that there is justification for supposing the values relevant to rational action are, again as the Self-Interest Theory claims, in some way relative to the agent. An agent can act only on the basis of values they have at hand, or rationally ought to have at hand, or only on the basis of considerations which make a difference to them.

Conclusion

Though we shall spend most of our time discussing the standard formulation of what I have called the Self-Interest Theory, I believe the theory is best understood as the conjunction of three claims concerning reasons for action: (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; and (3) the (intrinsic) strength of such a consideration is given by the value to the agent of outcome q. It is, as we have just seen, an independently plausible theory. Yet it is

---

also a theory that, in the end, I wish to deny. I have claimed plausible
the view that the rational intentions are those maximising expected-
value, but – in denying the Self-Interest Theory – I claim false the view
that the rational actions are those maximising expected-value. The
Self-Interest Theorist thus possesses not only an independently
plausible theory, but as well a taunt: if the rational intentions are those
maximising expected-value, then why aren’t the rational actions also
those maximising expected value?
Chapter Two

Some Toxin Puzzle is a Rational Dilemma

If the rational intentions are those maximising expected-value, does this mean that the rational actions are also those maximising expected-value? Does it mean, in other words, that the Self-Interest Theory is true? I say no. In this chapter, I shall argue that the Self-Interest Theory is in fact false, and that given certain conditions (to be specified below), the actions resulting from rational 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 may very well be irrational absent those conditions).

§1 The Self-Interest Theory, and Rational Irrationality

The Self-Interest Theorist, of course, would deny this, and claim that if the actions resulting from expected-value maximising, and so plausibly rational, intentions are not themselves expected-value maximising, then those actions are irrational. The Self-Interest Theorist is committed to the claim, then, that it could be rational to intend to perform an irrational action.

[1] Some actions – the Self-Interest Theorist argues – are clearly irrational. One afternoon, as you are minding your own business, you are approached by an eccentric billionaire. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. If you do not drink it, then nothing untoward will happen, and you will never see the billionaire again. He asks you to drink it. Should you drink it or not?

This is not the world’s most difficult decision problem. You rationally ought not to drink the toxin. If you drink it, the outcome will be that you are painfully ill for a day; if you do not, the outcome will be that nothing untoward happens to you. It should be painfully
obvious that you rationally ought not to drink the toxin, and if it isn't, it will quickly become painfully obvious that you shouldn't have!

We can justify this evaluation of drinking the toxin 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.
(T1) You are free not to drink the toxin, and the expected-value of your not drinking the toxin exceeds that of your drinking it. You know this is the case.

The first assumption, (S), is, of course, the standard formulation of the Self-Interest Theory of rational action. The second assumption, (T1), is a description of the central features of the situation I have just described.

This little example is a paradigm of the application, and plausibility, of theory (S). Since were you to drink the toxin you would be severely ill the following day, and since this conditional is not purely backward-looking, then the severe illness on the following day would be (part of) the outcome of drinking the toxin. Since were you not to drink the toxin nothing untoward would happen, and since this conditional is also not purely backward-looking, then nothing untoward happening would be (part of) the outcome of not drinking the toxin. The value to you of that outcome is obviously much greater than your suffering a day's severe illness. Since the value to you of the outcome you expect of not drinking the toxin exceeds that of drinking it, you rationally ought not drink the toxin. This much seems clear.

The second claim, (T1), describes what I take to be the central feature of the situation I have just described - that you are free not to drink the toxin, and the expected-value of your not drinking the toxin exceeds that of your drinking it. Recall from Chapter One that when I say a person is free to do some action, I mean simply that they can perform it, and that they can be such as not to perform that action. This qualification needs firstly to be included in theory (S) since otherwise it would allow for the possibility that you rationally ought to do something you cannot do, and this would violate the doctrine that 'ought' implies 'can'. As a result, we need to assume, in (T1), that in the situation described you are free not to drink the toxin, since otherwise
the conclusion that seems to follow — that you ought not to drink the toxin — would not in fact follow.

[2] Even if some actions are clearly irrational, though, it doesn’t follow — the Self-Interest Theorist now argues — that forming the intention to perform them is also irrational. The idea you might be approached by an eccentric billionaire with a vial of toxin is not original. In a short 1983 article in *Analysis*, Gregory Kavka introduced just such a situation, though more detailed than the one above. Here is Kavka’s example:

You are feeling extremely lucky. You have just been approached by an eccentric billionaire who has offered you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. ... The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. He emphasizes that you need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. ... All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. (The presence or absence of the intention is to be determined by the latest ‘mind-reading’ brain scanner ....)

Confronted with this offer, you gleefully sign the contract, thinking ‘what an easy way to become a millionaire’. Not long afterwards, however, you begin to worry. You had been thinking that you could avoid drinking the toxin and just pocket the million. But you realise that if you are thinking in those terms when midnight rolls around, you will not be intending to drink the toxin tomorrow. So maybe you will actually have to drink the stuff to collect the money. It will not be pleasant, but it is surely worth a day of suffering to become a millionaire.

However, as occurs to you immediately, it cannot really be necessary to drink the toxin to pocket the money. That money will either be or not be in your bank account by 10 a.m. tomorrow, you will know then whether it is there or not, and your drinking or not drinking the toxin hours later cannot affect the completed financial transaction.¹

You face a choice about whether or not to adopt at midnight the
intention to drink the toxin tomorrow afternoon. Should you adopt
the intention to drink the toxin or not?\(^2\)

This decision problem is only a little less straightforward than the
previous one. You rationally ought to adopt the intention to drink the
toxin. The expected-value of your adopting the intention to drink the
toxin is greater than that of your not adopting this intention: if you
were to adopt the intention, then you would get a million dollars; if
not, not. Having a million dollars is better than not having it, even if it
turned out that you went ahead and drank the toxin the next
afternoon. All this is supposed to justify the claim that you rationally
ought to adopt the intention to drink the toxin. However, even though
you ought to adopt the intention of drinking the toxin, it is important
to realise it remains the case you ought not actually to drink it. The
expected-value of your not drinking the toxin still exceeds that of your
drinking it: if you were to drink it, then you would only get a day of
severe illness; if not, not. A day's severe illness is still worse than not,
and you do not have to drink the toxin to get the money, so the only
outcome it will in fact have is a day's severe illness. All this is
supposed still to justify the claim that you rationally ought not actually
to drink the toxin.

Our belief in these two claims - that it is rational to adopt the
intention to drink, though irrational to drink - depends, once again,
on theory (S), but also on an extended version of assumption (T1).
What we need to assume in this case is the following:

\[(T2) \text{ (a) You are free to adopt the intention to drink the toxin, and the expected-value of your adopting this intention exceeds that of your not adopting this intention; (b) and, whether or not you actually adopt the intention, you are free later not to drink the toxin, and the} \]

\(^2\) It is unclear, in 'The Toxin Puzzle', whether it is the having of the intention or the adopting (or forming) of the intention which Kavka is concerned to rationally evaluate. D. M. Farrell, in 'Intention, Action and Reason', p. 293, also finds this problem. In the first place, having an intention is (arguably) not an action, and so its rationality does not fall under the aegis of the Self-Interest Theory. Furthermore, there is some doubt that a consequentialist account can be given of the having of intentions. See R. Edgley, Reason in Theory and Practice, (London: Hutchinson, 1969), pp. 63 ff. Adopting an intention is (I shall assume) an action, and I shall concentrate solely on the rationality of adopting the intention.
expected-value of your not drinking it still exceeds that of your drinking it. You know this is the case.

I simply add clause (a) to (T1) to get this assumption, (T2). I will call situations satisfying (T2) Toxin Puzzle Cases. Though it is implicit, we need to assume that situations Kavka describes have a certain structure, and are in fact possible.

Toxin Puzzle Cases show it could indeed be rational to adopt an intention to perform an irrational action. It is rational to adopt the intention because of the million you would get for doing so; it is irrational actually to drink because of the day of pain it would cause you. Yet nothing crucially hangs on the fact it is one million dollars, or that it is one day of pain. The Self-Interest Theorist will want to generalise: whenever it maximises expected-value to adopt an intention to perform an action which does not maximise expected-value, it follows that it is rational to adopt an intention to perform an irrational action. The Self-Interest Theorist will simply deny my claim that, under some circumstances (which I have not yet introduced), if it is rational to intention to A then it is rational to A, even if one is free not to A, and it maximises expected-value not to A.

---

3 An important part of assumption (T2) is that you are free to adopt the intention to drink the toxin. Yet Kavka stipulates that 'arrangement of ... external incentives is ruled out, as are such alternative gimmicks as hiring a hypnotist to implant the intention, forgetting the main relevant facts of the situation, and so forth' (p. 34). It seems, then, he stipulates you cannot, even indirectly, adopt the relevant intention. A number of discussions of the Toxin Puzzle therefore centre on whether or not you can adopt the intention. See, for example, R. A. Sorenson, 'Blindspotting and Choice Variations of the Prediction Paradox,' Amer Phil Quart 23 (1986) pp. 338 ff., and A. Hajek, 'Nuke'em Problems,' Analysis 51 (1991), p. 256. This stipulated restriction by Kavka is, however, unmotivated. Kavka says (p. 36) the the 'Toxin Puzzle' paper broadens the application of an earlier discussion in 'Some Paradoxes of Deterrence,' J Phil 75 (1978), pp. 285-302, where he attempts to argue it might be morally (rationally) required to adopt a (conditional) intention to perform immoral (irrational) retaliatory actions. Yet the 'Toxin Puzzle' paper can broaden this discussion only if the conclusion it ends up with is that it could be rationally required to adopt a (non-conditional) intention to perform irrational toxin-drinking actions, and it will end up with this conclusion only if you can adopt such a (non-conditional) intention. After all, you can be required to adopt only those intentions you can adopt. This point is also noted by J. Thomas, 'The Toxin, the Blood Donor, and the Bomb,' Analysis 43 (1983): 207-210. Thus I will assume you can (if only indirectly) adopt the intention to drink the toxin.
§2 The Toxin Puzzle

It seems puzzling that it could be rational to adopt the intention to perform an irrational action. Yet, as we have just seen, this is precisely what the Self-Interest Theory entails. Some are inclined to see in this an objection to the theory. There are those who claim, roughly, that, in all circumstances, if it is rational to form an intention then it is rational to act upon it; others who claim, roughly, that agents cannot have an intention without thereby acting upon it; and yet others still who claim, roughly, that agents ought not have an intention without thereby acting upon it. In this section I will introduce more precisely the form these three objections take, and argue that none of them is plausible.

[1] The first objection begins with the suggestion that we should evaluate the rationality of an intention, and that of adopting an intention, not in terms of consequences, as we have already done, but rather always in terms of the rationality of the action which is the object of the intention. In the present case this means, presumably, something like:

(B1) If an agent rationally ought not to perform some action A, truly believes he is free not to intend (adopt the intention) to do A, and the considerations for or against doing A will not change by the time it comes to A or not, then they ought not to intend (adopt the intention) to do A.

One does not need to go far to find supporters of so-called bridging-principles like (B1), for they are principles with a long and distinguished background.4 The basic intuition motivating belief in a principle like (B1) is that what it is about an action we find irrational, or for that matter immoral, is not the actual carrying through of the motions constituting the action, but rather the will – the intention – behind these motions. Aside from the fact that the agent in question knew what they were doing was irrational, or immoral, what we find

---

objectionable is that they did it intentionally - that they intended to do it. The intention to do something irrational, or immoral, which is, for example, thwarted due to some external circumstance, we feel to be just as irrational, or immoral, as that which is not.

If this principle, or one like it, is correct, then it will be false to say that you ought to adopt the intention to drink the toxin, but ought not drink the toxin. For if you ought not to drink, and - as we have assumed - you truly believe you are free not to adopt the intention to drink, and the considerations for and against (but mostly against) drinking will not change by the time it comes to drink, then, according to (B1), you ought not adopt the intention to drink, and so it is false that you ought to adopt this intention. Since the Self-Interest Theory says you ought not drink the toxin, but you ought to adopt the intention, then it would be false.

It is plausible, however, to deny the rationality of intentions and that of the intended actions are in this way related. For in the very case we are now considering, where we do want to say that the intended action is irrational, we want to say this not because the intention behind the action is irrational (that intention is going to get us a cool million) but rather because it is the actual carrying through of the motions constituting the action which would cause all the trouble (those motions are only going to lead to a day of misery). You ought to adopt the intention to drink the toxin, since this is what will have the best outcome; and you ought not to drink the toxin, since this is also what will have the best outcome. It is consistent to say this - the adoption of the intention to act, and the action itself, are different things, though closely related they be. Since they are different things it is possible to make different judgements about them without inconsistency.

[2] The second form of the objection against the Self-Interest Theory centres on the sort of thing you are being required to do: namely, to be such that you intend to do something, but not do that thing. The

---

5 Note that I am not concerned in this section to identify precisely the form such a principle might take, but merely to indicate what a defender of the Self-Interest Theory might have to say about all such principles.

complaint is that people just cannot be like this. More precisely, the
claim is that

(I1) No agent can be such that they intend (adopt the intention) to A,
truly believe they are free to A, believe the considerations in favour
of intending (adopting the intention) to A have not changed by the
time it comes to A, believe that the time for action has arrived, and
yet not A.

On this view, the very function of intention is that, barring external
impediments and relevant changes in the situation, it leads to action.
The intuition behind principle (I1) is the idea that the intention to do
something is an internal state which is, so to speak, orientated to the
occurrence of the action which is its object. What would it be like for
you to have this intention, be free from external impediments, be
aware that none of the relevant considerations have changed, but still,
due to some solely internal factor, not perform the action when the
time comes?

If this principle, or one like it, is correct, then it will be false to say
that you ought to adopt the intention to drink, but ought not to drink.
For the considerations in (T2) relevant to the adoption of the intention
to drink the toxin, as we have seen, are: if you adopt this intention, you
get one million dollars; if not, not. These facts do not change from the
time of adopting the intention to the time of drinking. Hence, in (T2),
the considerations relevant to adopting the intention do not change.7
According to (I1), you cannot adopt the intention to drink the toxin,
truly believe (as you do) that you are free to drink, believe (as you do)
that the considerations in favour of adopting the intention have not
changed by the time it comes to drink, believe (as you do) that the time
for action has arrived, and yet not drink the toxin. Since the Self-
Interest Theory says you ought to adopt the intention, but ought not
drink – even though you can’t do both – then it would be false.

A principle such as this may plausibly be denied. For it is possible
that the explanation of the failure to carry through an intention lies
wholly within the individual. Such is the problem, for example, with

7 What might change, of course, is whether you have a million dollars. But the fact
that you (or do not) have a million dollars is neither a reason for nor against
adopting the intention, and so any change in your financial status will not be a
relevant change in circumstances.
fickle people, who are disposed to reconsider their decisions, even when the facts relevant to their original situation have not changed in any relevant way. Such persons may be said to suffer from what one might term weakness of intention—an internal failure to do what they have decided to do—in just the way others are said to suffer from weakness of will—an internal failure to intend to do what they judge they rationally ought to do. The possibility of weakness of will and weakness of intention should stand or fall together. In particular, these possibilities seem to stand, and so principle (II) is incorrect.

[3] The third form the objection against the Self-Interest Theory might take, and the final one I will consider, is a development of the previous one. It again centres on the sort of thing you are being required to do: namely, to be such that you intend to do something, but not do that thing. The complaint, in this case, is that people just ought not be like this. More precisely, the claim is that

(N1) No agent is rationally permitted to be such that they intend (adopt the intention) to A, truly believe they are free to A, believe that the considerations in favour of intending (adopting the intention to A) have not changed by the time it comes to A, believe that the time for action has arrived, and yet not A.

Such a principle certainly seems more plausible than (II). Even if we think that people can be such as to intend to do something, and yet fail for solely internal factors to do it, we might consider such people less than rational. Even if we concede, as I think we must, that people can suffer from weakness of intention, we must evaluate such a weakness as we would evaluate the analogous malady of weakness of will. In people suffering from weakness of intention, the intention fails to exert adequate control over their actions, to the detriment of their rationality.

Again, if this principle is correct, then it will be false to say that you ought to adopt the intention to drink the toxin, but ought not drink the toxin. According to (N1), you are not rationally permitted to adopt

---

8 D. Gauthier, in 'Deterrence, Maximization, and Rationality,' *Ethics* 94 (1984), pp. 482 ff. suggests that an agent is fully rational only if she is fully in control of what she does; that is, only if she does what she (conditionally) intends to do (if the condition arises).
the intention to drink the toxin, truly believe (as you do) that you are free to drink, believe (as you do) that the considerations in favour of adopting the intention have not changed by the time it comes to drink, believe (as you do) that the time for action has arrived, and yet not drink the toxin. Since the Self-Interest Theory says you ought to adopt the intention, but ought not drink – even though you ought not do both – then it would be false.

Those who are concerned to defend the Self-Interest theory have a ready response to all generalisations like (N1) about the dispositions people rationally ought to have, or adopt. Someone who is concerned to evaluate the rationality of actions on the basis of their outcomes, they will say, will not be at all convinced of the truth of claims such as (N1). Such a person will be suspicious of this principle, since they will think that whether or not one ought have a disposition to act in a certain way will depend on the outcome of having it, in just the same way that the rationality of actions is to be assessed on the basis of their outcomes. Certainly, they will admit, being disposed to do what you intend to do is a useful disposition in a wide range of situations. But will it necessarily be useful in all possible situations? Surely not. Indeed, the situation we are now discussing is precisely one in which the disposition to do what you have decided would be needlessly costly. In this situation it is clear that you rationally ought not to be such that if you intend to perform some action, are free to do so, and are aware that there has been no relevant change in circumstances, then you actually go through with action. In this situation, it would be best for you, after you have received the million, to reconsider. Defenders of the Self-Interest Theory will thus deny principles such as (N1), since it is unlikely the disposition it refers to is value-maximising in all possible circumstances.

I have considered three objections against the Self-Interest Theory, and they all fail to meet their mark. The situation Kavka describes leads, it seems, to the conclusion you rationally ought to adopt the

---

9 They are not the only ones. Someone who toys with the idea, for example, that a maximally rational agent might be to a certain extent akratic, is A. O. Rorty, ‘Self-Deception, Akrasia, and Irrationality,’ in J. Elster (ed.), The Multiple Self: Studies in Rationality and Social Change, (Cambridge: Cambridge Univ. Pr., 1986): 115-31. Furthermore, (some) consequentialists are famed for the insistence that all types of internal states of the agent are to be evaluated in terms of their consequences. See, for example, H. Sidgwick, The Methods of Ethics, (New York: Dover, 1966), p. 493, and, further on this point, R. M. Adams, ‘Motive Utilitarianism,’ J Phil 73 (1976): 467-481.
intention to perform an irrational action. Though puzzling, these situations give us no reason to reject the Self-Interest Theory, (S). Since the above three principles are false, Kavka's example is certainly no counterexample to their theory, the defenders of the Self-Interest Theory will say.\(^\text{10}\)

§3 The Self-Interest Theory, Refuted

Principles (II) and (N1) are obviously too strong: surely, for example, there are Toxin Puzzle Cases in which you can both have the intention and fail to carry it out; and surely there are Toxin Puzzle Cases in which you are rationally permitted both to have the intention and fail to carry it out. As objections to the Self-Interest Theory, however, these principles are on the right track. For just as surely are there Toxin Puzzle Cases in which you can not have the intention while failing to carry it out; and just as surely are there Toxin Puzzle Cases in which you are not rationally permitted to have the intention while failing to carry it out. In this section, I shall provide two counterexamples to the Self-Interest Theory.

[1] The first argument against the Self-Interest Theory is as follows. Imagine, again, the situation described by Kavka - what I called the Toxin Puzzle situation - but add to its details the following:

While contemplating that you do not need to drink the toxin in order to get the million, you realise that your resolute character stands in the way of this. For you were brought up to be such that if you intend to do something, then you do it. Furthermore, you realise, you cannot but be like this. Not only will you not, but you also cannot be such that both you intend to do something, and then fail to do it. You conclude, to your chagrin, that you cannot be such as to intend to drink the toxin without doing so.

It turns out, then, that you cannot be such that you adopt the intention at midnight to drink the toxin, and then not drink the toxin later that

\(^{10}\) This is the line Kavka himself takes in 'The Toxin Puzzle'. He concludes: 'when we have good reasons to intend but not to act, conflicting standards of evaluation come into play and something has to give way: either rational action, rational intention, or aspects of the agent's own rationality' (36). He, and others who would draw this conclusion, do not consider the possibility (as I shall in a moment) that it is the theory of rationality leading to this conclusion which has to give way.
afternoon. Not that this is a grand logical truth about the inevitable relation between intention and unhindered action – as was suggested by principle (II) – but just a humble truth about your own makeup.

The assumption we need to make if this is indeed to be a counterexample to the Self-Interest Theory is that situations of the following form are logically possible:

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

I simply add clause (x) to assumption (T2) to get this assumption, (T3a). That is, I simply add, to the Toxin Puzzle, the stipulation that you in fact cannot adopt the intention without later drinking the toxin. I will call (T3a) the First Counterexample.

If the stipulation I have just described can indeed be consistently added to assumption (T2), then it will no longer follow that the situation involves a rational intention to perform an irrational action. In the situation I have just sketched, you cannot both adopt the intention to drink the toxin and not perform this action. But, in every situation satisfying assumption (T3a), the Self-Interest Theory is committed to the claim both that you rationally ought to adopt the intention, and that you rationally ought not to drink the toxin. However, this is not possible:

(OC2) If an agent rationally ought to S₁, and rationally ought to S₂, then he can be such that he S₁'s and he S₂'s.

In other words, an agent rationally ought to S₁ and rationally ought to S₂ only if they can both S₁ and S₂. In the First Counterexample, the Self-Interest Theory is committed to the claim that you rationally
ought to adopt the intention, and you rationally ought not drink the toxin; in the First Counterexample, you cannot adopt the intention without drinking – the Self-Interest Theory is false.

[2] The second argument against the Self-Interest Theory goes as follows. Imagine, again, the original situation described by Kavka, but this time add to its details the following, which have to do with another very strange encounter you had this morning, with another eccentric billionaire. The story was as follows:

You didn’t believe your luck at the time. You were approached by an eccentric billionaire who made you the following deal. He would give you ten million dollars for *adopting the enduring disposition that if you intend to do something, then you do it*. He emphasised that he would give you the money only if the disposition to do what you intend was not a *compulsive* disposition. If you brought it about that the relevant conditional was true by ensuring that, whenever you have an intention, you are rendered incapable of refraining from the intended action, you would not get the money. It is only by freely doing what you intend to do that you will get the ten million. (The presence or absence of the disposition is to be determined by the latest mind-reading scanner, switched to disposition-detection mode, which this previous billionaire also happened to have on him at the time, and which you believe to be perfectly accurate.)

Again, confronted with the offer, you gleefully signed the contract, thinking ‘what an easy way to become a millionaire ten times over’. You acquired, right then, the relevant disposition, and so, for the next few days, you became a person such that if you intend to do something, then you actually (freely) do it. The ten million was in your bank account half an hour later.

It turns out, then, that you ought to have adopted this disposition. Again, not that this is a grand logical truth about the normative relation between intention and unhindered action – as was suggested by principle (N1) – but rather that this is a humble truth about your own situation.

In passing, it may be suggested that it is incoherent to suppose you can at a particular time make it enduringly that if you intend to A then you A. What it is coherent to suppose is only that you can at a particular time make it enduringly that you have the policy that if you intend to A then you A. I’m not so sure, though, there is any incoherence in the original supposition. It is certainly coherent to
suppose that, enduringly, if you intend to perform some action, then you (freely and intentionally) do so come the time for action. The enduring truth of this conditional, though, could presumably be explained by the presence of factors XYZ. It is coherent to suppose, then, that factors XYZ caused you to be such that, enduringly, if you intend to perform some action, then you (freely and intentionally) do so come the time for action. The existence of such factors, in turn, could presumably be explained by something you yourself did. If this is indeed the explanation, then it seems perfectly coherent to describe what you yourself did as ‘making it enduringly that if you intend to A then you A.’ Generally, any action at an earlier time $t_1$ causing action A at a later time $t_2$ can be properly described as ‘making it at $t_1$ that A is done at $t_2$’. It is such an action I have in mind.

The assumption we need if this example is indeed to be a counterexample to the Self-Interest Theory is that situations of the following form are logically possible:

(T3b) (x) You are free to adopt the enduring disposition that if you intend to do something then you do it, and the expected value of adopting this disposition exceeds that of your not adopting it; (a’) whether or not you adopt this disposition, you are free to adopt the intention to drink the toxin, and the expected value of your adopting this disposition exceeds that of your not adopting it; and (b’) whether or not you adopt this disposition, and whether or not you adopt the intention, you are free later still not to drink the toxin, and the expected value of your not drinking it still exceeds that of your drinking it. You know all this is the case.

This time I simply add (a different) clause (x) to (T2), and slightly modify the other two clauses, to get this assumption, (T3b). That is, I simply add, to the situation Kavka describes, the stipulation that you face an initial choice about whether to adopt the disposition to do what you intend to do, and that, in the situation you find yourself in, you get the best outcome by adopting just such a disposition. I call (T3b) the Second Counterexample.

If I understand him correctly, this is an objection Philip Pettit pressed against the Second Counterexample.
If the stipulation I have just described can be consistently added to assumption (T2), then again it will no longer follow that the situation involves rational irrationality. In the situation I have just sketched, it has the best outcome to adopt the disposition, has the best outcome to adopt the intention, and yet it also has the best outcome not to perform the action. The Self-Interest Theory is thus committed to the following claims: you rationally ought to make yourself such that if you intend to drink the toxin, then you drink it; you rationally ought to adopt the intention to drink the toxin; and, yet, that you rationally ought not drink the toxin. However, this is not possible:

(OP3) If an agent rationally ought to $S_1$, rationally ought to $S_2$, and rationally ought to $S_3$, then it is logically possible that he $S_1$'s, he $S_2$'s and he $S_3$'s.

In other words, an agent rationally ought to $S_1$, rationally ought to $S_2$, and rationally ought to $S_3$ only if it is logically possible they $S_1$, $S_2$ and $S_3$. In the Second Counterexample, the Self-Interest Theory is committed to the claim you rationally ought to adopt the disposition, rationally ought to adopt the intention, and yet you rationally ought not drink the toxin. But it is not logically possible for you to adopt the disposition, adopt the intention and yet not drink the toxin - the Self-Interest Theory is doubly false.

I have my own objections, then, to Self-Interest Theory. The three objections we considered above relied for their effectiveness on certain principles, such as (B1), (II) or (N1). They are principles defenders of the Self-Interest Theory was more than happy to deny. It turns out, however, that we can use other, less controversial, deontic principles, such as (OC2) and (OP3), to construct counterexamples to the Self-Interest Theory. It is easy to deny that rational intentions make for rational actions; it is harder to deny that one is rationally obliged to do only what one can do.\(^\text{12}\)

§4 Rational Intention, Rational Action

I say that the Self-Interest Theory has been refuted, but, of course, I will need to defend this claim against various objections. This I will do in the next chapter. For the rest of this chapter – and on the assumption I can meet these objections – I want to argue, firstly, in the First and Second Counterexamples, you rationally ought to adopt the intention but – contrary to the Self-Interest Theory – are rationally permitted to drink, and, secondly and more generally, that given certain conditions (to be specified below), the actions resulting from any rational 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 may very well be irrational absent those conditions).

[1] The first half of the task for this section consists in demonstrating that in the First and Second Counterexamples, you rationally ought to adopt the intention but are rationally permitted to drink. I concentrate only on the First Counterexample; the argument for the Second is similar.

In the First Counterexample it seemed that you ought, and could, adopt the intention to drink the toxin, that you ought, and could, not drink the toxin, even though you could not adopt the intention without carrying it out. Since, appearances notwithstanding, this is not possible, we must reject either the claim that you rationally ought to adopt the intention or the claim that you rationally ought not to drink the toxin. It is clear, though, we should reject the claim that you rationally ought not to drink the toxin, and this for two separate reasons.

First, in the First Counterexample, there are four ways you can be:

(a) I D
(b) I ~D
(c) ~I D
(d) ~I ~D

where: I=adopting the intention to drink the toxin; D=drinking the toxin. In the First Counterexample, you cannot adopt the intention without drinking – therefore (b) is not really a way you might be. Maximally rational agents are those who can expect to most promote
what they value. Of the three ways you might be: best is (a) - you get a million dollars, though drink the toxin ('It will not be pleasant, but it is surely worth a day of suffering to become a millionaire'); next best is (d) - since you don't have the intention, you don't get the million, so you would be foolish to drink; worst is (c). Hence, if you were maximally rational, then you would adopt the intention to drink, and therefore drink. But anything a maximally rational person would do is something ordinary mortals, like you, are rationally permitted to do: if you would A were you maximally rational, you are rationally permitted to A.13 You would drink the toxin were you maximally rational, and so you are rationally permitted to drink the toxin.

Some might be tempted to object, at this point, to my use of the principle that if A is what you would do were you rational, then you are rationally permitted to A. In particular, Kavka himself might object to this principle, which is closely related to what he calls the Right-Good Principle14 - that doing something is right if and only if a morally good person would do the same thing in the given situation. Such an objection, however, is misplaced. Assuming I can deal with objections to my above counterexamples (which assumption I shall discharge in the next chapter), it is perfectly legitimate for me to use the Right-Good Principle. Anyone inclined to make this objection at this point is asked to read the next chapter.

Second, consider a slightly altered version of Kavka's original Toxin Case. Suppose the adoption of the intention and drinking the toxin are actions which would occur at the same time, but all other details of the example remain the same, consistent with this change. The billionaire will pay you one million dollars now if, right now, you adopt the intention to drink the toxin, right now. It turns out, unsurprisingly, that you cannot now adopt the intention to drink the toxin now, without drinking the toxin now. In this variation, you face no puzzle. Clearly, Self-Interest Theorist or not, you rationally ought

---


adopt the intention to drink the toxin, and you are thus rationally permitted to drink the toxin. It would be wrong to think that you face, in this synchronic example, a dilemma about which action to choose. Why? The answer is plain: quite simply, you cannot adopt the intention without drinking the toxin. In general, the fact that a person cannot do both A and B is a conclusive reason to believe it false that they ought to do A and ought to do B.

Similarly, we should reject the claim, in the diachronic example I have been considering in this chapter, that you ought not drink the toxin. The only difference between this second, synchronic, case and the original, diachronic, case is the time at which you would drink the toxin. In this second synchronic case, it would be at the same time as that of your adopting the intention; in the original diachronic case, it would be at a later time. What we have to say about whether or not you face a dilemma should not be different in either of the two cases. But it is obvious that, in this second synchronic case I have just described, you do not face a dilemma at that time — it is obvious that you rationally ought to adopt the intention, and are rationally permitted to drink the toxin while you do so. Hence, in the original diachronic case you also do not face a dilemma — it should be just as obvious in that case that you rationally ought to adopt the intention, and are rationally permitted to drink the toxin after you have done so. The fact that a person cannot do both A and B is a conclusive reason to believe it false that they ought to do A and ought to do B, and this applies just as much to diachronic cases as it does to synchronic ones.

In either of these two ways, then, we see that in the First (and, similarly, in the Second) Counterexample you are rationally required to adopt the intention, and rationally permitted to drink the toxin, even though you are free not to drink the toxin (condition T3a(b)) and it has the best outcome not to drink the toxin (condition T3a(b)). This claim may be generalised.

[2] The second half of the task for this section is to demonstrate that, GIVEN that 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, 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). The demonstration relies on the deontic principles I introduced above.\textsuperscript{15}

The argument for this claim has a number of parts. On the one hand, deontic principle (OC\textsubscript{2}) above, says that one rationally ought to S\textsubscript{1} and rationally ought to S\textsubscript{2} only if one can S\textsubscript{1} and S\textsubscript{2}. Letting S\textsubscript{1} be ‘adopt the intention to perform some action A’ – for example, A might be ‘drink the toxin’ – and S\textsubscript{2} be ‘not A’, it follows from (OC\textsubscript{2}) that:

(1) You rationally ought to adopt the intention to A and rationally ought not to A only if you can be such that both you adopt the intention to A and not A.

On the basis of some propositional logic (in particular, the principle that I & A → C entails -C → (I → -A)), and the fact that it is false you rationally ought not to A if and only if you are rationally permitted to A (that is, that -Ought (-A) if and only if Permitted(A)), it follows from (1) that

(2) Given that you cannot both adopt the intention to A and not A, then if you rationally ought to adopt the intention to A, then you are rationally permitted to A.

On the other hand, deontic principle (OP\textsubscript{3}), above, says that one rationally ought to S\textsubscript{1}, rationally ought to S\textsubscript{2}, and rationally ought to S\textsubscript{3}

\textsuperscript{15} I know of only one other attempt to derive conclusions about the relations between the rationality of intention and that of action from deontic principles. J. Lango, in ‘Is it Wrong to Intend to do that which it is Wrong to do?’ Monist 70 (1987): 316-29, argues for the so-called Wrongful Intentions Principle (WIP) – that it is wrong to intend to do what it is wrong to do – on such a basis. His argument depends on: (1) the Requisites Principle, that if one intends to do X (or X if Y occurs), then one will do X, provided that the requisites (occasion, frame of mind, circumstances) for doing X are satisfied (p. 319); (2) the Wrongful Commitment Principle, that if an action X is morally wrong, then it is morally wrong that X will be done if the requisites for doing X are satisfied (p. 322); and (3) the Obligatory Implicant Principle, that if A implies B and if A is morally obligatory, then B is morally obligatory (p. 324). In my view: (1’) the Requisites Principle is not required at all to demonstrate the existence of a relation between the rationality of intention and that of action (though it may be necessary to demonstrate the relation Lango claims to obtain – namely, the WIP); (2’) the consequent of the Wrongful Commitment principle is ambiguous between ‘if the requisites are satisfied, then it is morally wrong that X will be done’ – which makes the principle true but the whole argument invalid – and ‘it is morally wrong that if the requisites are satisfied then X will be done’ – which seems to make the principle false; and (3’) the Obligatory Implicant Principle is stronger than the deontic principles I use – (OP\textsubscript{n}) and (OC\textsubscript{n}) – in my argument. I shall not discuss Lango’s argument further.
only if it is logically possible that one $S_1$, $S_2$, and $S_3$. Letting $S_1$ be 'adopt the enduring disposition that if you intend to $A$ then you $A'$, $S_2$ be 'adopt the intention to $A'$ and $S_3$ be 'not $A'$, it follows from (OP3) that

(3) You rationally ought to adopt the enduring disposition that if you intend to $A$ then you $A$, you rationally ought to adopt the intention to $A$, and you rationally ought not to $A$ only if it is logically possible for you to adopt the disposition, adopt the intention, and yet not $A$.

Yet this is not possible: it is not logically possible for you to adopt – and so have – the disposition that if you intend to $A$ then you $A$, adopt – and so have – the intention to $A$, and yet not $A$. On the basis of some propositional logic (in particular, the principle that $D \& I \& A \rightarrow P$ and $-P$ entail $D \rightarrow (I \rightarrow -A)$), it follows from (3) that

(4) Given that you rationally ought to adopt the enduring disposition that if you intend to $A$ then you $A$, then if you rationally ought to adopt the intention to $A$, then you are rationally permitted to $A$.

Finally, on the basis of yet some more propositional logic (in particular, the principle that $-C \rightarrow (I \rightarrow -A)$ and $D \rightarrow (I \rightarrow -A)$ entail $(-C \lor D) \rightarrow (I \rightarrow -A)$), and claims (2) and (4), we may conclude that GIVEN that 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, you are rationally permitted to perform that action. The above argument depends on no particular assumptions about the action in question – $A$. Hence, our conclusion obtains EVEN IF one is free not to $A$, and it has the best outcome not to $A$ (though, I should point out, it may not be rationally permitted to $A$ if the above conditions are not given).

There are a couple of deontic principles – generalisations of (OC2) and (OP3) – which will be important in the remainder of the thesis. They are concerned with the relation between the notion of rational obligation and that of possibility, and are closely related of Kant's doctrine that 'ought' implies 'can':

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

(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, ...)

These principles provide necessary conditions for being under a group of separate rational obligations. The formulation of my objection against the Self-Interest Theory thus needs to make explicit reference to generalisations of the doctrine that 'ought' implies 'can'. The strategy of my objection should now be clear: instead of relying on problematic principles such as (B1), (II), or (N1), I rely instead on the less problematic deontic principles (OP_n) and (OC_n) to make my argument work. To repeat, it is easy to deny these three principles; it is harder to deny that one is rationally obliged to do only what one can do.

Conclusion

The Self-Interest Theory is false, and under some (but not all) conditions, rational intentions make for rational actions. I started the thesis with the question: are the non-expected-value maximising actions resulting from expected-value maximising, and so plausibly rational, intentions also rational? I said there are two schools of thought on this issue. On the one hand, some say that since such actions are the result of intentions it is presumably rational to have, then they too must be rational; on this view of the matter, if it is rational to have or to adopt a certain intention, then it is always rational to act upon it. On the other hand, others say that since such actions are not expected-value maximising, then they must be irrational; on this view of the matter, since the rational intentions are those which maximise the expected-value, then surely the rational actions must also always be those maximising expected-value. The truth, as we now see, lies in the middle. Given the inevitability or rationality of certain dispositions, rational intentions do make for rational actions; lacking this inevitability or rationality, rational intentions may indeed have irrational actions as their objects. We may steer between the Scylla of accepting problematic bridging (and other) principles, and the Charybdis of endorsing wholesale the Self-Interest Theory.
I have argued that the Self-Interest Theory is false, and that given the inevitability or rationality of certain dispositions, rational intentions do indeed make for rational actions. My arguments for these claims, and particularly the first, depend (amongst other things) on two central assumptions:

(T3a) First Counterexample. (a) You are free to adopt the intention to drink the toxin, and the expected-value of your adopting this intention exceeds that of your not adopting this intention; and (b) whether or not you actually adopt the intention, you are free later not to drink the toxin, and the expected-value of your not drinking it still exceeds that of your drinking it; but (x) you cannot both adopt the intention to drink the toxin, and then, later, not drink the toxin. You know all this is the case.

(T3b) Second Counterexample. (x) You are free to adopt the enduring disposition that if you intend to do something then you do it, and the expected-value of adopting this disposition exceeds that of your not adopting it; (a') whether or not you adopt this disposition, you are free to adopt the intention to drink the toxin, and the expected-value of your adopting this intention exceeds that of your not adopting it; and (b') whether or not you adopt this disposition, and whether or not you adopt the intention, you are free later still not to drink the toxin, and the expected-value of your not drinking it still exceeds that of your drinking it. You know all this is the case.

The defender of the Self-Interest Theory, however, might claim that when I add each version of clause (x) to the Toxin Puzzle, (T2), to get (T3a) and (T3b) respectively, I change the situation in a crucial way. As we will see, this is to claim, in effect, that assumptions (T3a) and (T3b) are each inconsistent. There are two broad types of objections to these assumptions, and in this chapter I shall argue that neither of them succeeds.
§1 Is it really still possible for you not to drink the toxin?

I constructed the First Counterexample by adding to the Toxin Puzzle, (T2), the stipulation that you cannot adopt the intention to drink the toxin without later drinking it, to get (T3a). I constructed the Second by adding the stipulation that it maximises expected-value for you to adopt the enduring disposition of always doing what you intend, to get (T3b). I claimed that, in these modified situations, you are rationally permitted to drink the toxin, even though you are free not to do so, and it maximises expected-value for you not to do so. The Self-Interest Theorist, however, may object that these two stipulative additions mean you are no longer free not to drink the toxin. As I shall argue in this section, though, that this objection is mistaken, and in the First and Second Counterexamples, that you may indeed still be free not to drink the toxin.¹

[1] What form does this objection take to the First Counterexample? The original description of the Toxin Case included the assumption you are free, before midnight, to adopt the intention to drink the toxin the following afternoon, and the assumption that, whether or not you adopt this intention at midnight, you would still be free the following afternoon to drink the toxin or not. I assumed, amongst other things, that

(T3a) (a) You can adopt the intention to drink the toxin; (b) if you were to adopt the intention to drink the toxin, then you still could later not drink the toxin.

This is why we could both evaluate the rationality of adopting the intention, and – whether or not you actually adopted the intention –

¹ There are forms this objection may take different from the one I consider in the text. For instance, in each of the counterexamples, the billionaire must be aware of your disposition to do what you intend. Some suggest, though, that free actions are essentially unpredictable. See G. Schlesinger, 'The Unpredictability of Free Choices,' Brit J Phil Sci 25 (1974): 209-21, and R. Hardin, 'Pragmatic Intuitions and Rational Choice,' in A. Diekmann & P. Mitter (ed.s), Paradoxical Effects of Social Behaviour: Essays in Honour of Anatol Rapoport, (Hiedelberg: Physica Verlag, 1986): 27-36. Space prevents me from dealing with this objection, or other forms it might take.
THE TOXIN PUZZLE – OBJECTIONS

evaluate the rationality of drinking the toxin. However, in the First Counterexample, I made the additional assumption that, in spite of this,

(T3a) (x) You cannot both: adopt the intention to drink the toxin, and later not drink the toxin.

Even though you are free with respect to the adoption of the intention, and free with respect to carrying it out, I wanted to add that you nevertheless could not have the intention without carrying it out. The objection is that this is just inconsistent – T3a(a), T3a(b) and T3a(x) are jointly inconsistent.²

[1.1] Denote by ‘Cₓ(Fₓ)’ the claim that x can be such as to F, ³ and by ‘p ⊨ q’ the subjunctive claim that if p were the case then q would be the case. The issue is whether or not the following inference is valid:

\[
\frac{Cₓ(Fₓ), Fₓ ⊨ Cₓ(Gₓ)}{Cₓ(Fₓ & Gₓ)}
\]

(C)

My opponent insists that if you can adopt the intention to drink the toxin (T3a(a)), and, if you were to adopt the intention you would still be able later not to drink the toxin (T3a(b)), it logically follows that you can both adopt the intention and then not drink the toxin (not T3a(x)).⁴ My opponent endorses the inference, I do not. Which of us is right?

² I would like to thank Peter Menzies and Andrew Gleeson for convincing me of the force of this objection, and for their helpful comments regarding it. This section is a product of my discussion with them.


⁴ There are other forms an inference such as this might take. B. A. O. Williams, in ‘Ethical Consistency’, Proc Aris Soc 39 (1965), p. 121, and R. Trigg, ‘Moral Conflict,’ Mind 80 (1971), pp. 44, both seem to endorse the closely related inference ‘Cₓ(Fₓ),
A simple counterexample shows inference (C) is invalid. Suppose that 'Fx' is you drink the toxin, that 'Gx' is you do not drink the toxin (equivalent to '~Fx'), and suppose you are free to drink the toxin or not, and your ability not to drink the toxin is causally independent of your drinking it. Substituting into (C), we get an inference of the form:

\[
\begin{align*}
\text{C}_x(Fx), \ Fx &\rightarrow \text{C}_x(\neg Fx) \\
\hline
\text{C}_x(Fx &\rightarrow \neg Fx)
\end{align*}
\]

The premises are true, though the conclusion false. Since you are free to drink the toxin, the first premise – Cx(Fx) – is true. Since you are free not to drink the toxin – Cx(¬Fx) – and this ability is (we have assumed) causally independent of your drinking it, then the second premise – Fx \rightarrow Cx(¬Fx) – is also true. However, the conclusion – Cx(Fx & ¬Fx) – is plainly false.

My opponent may not be at all impressed with the supposition that your ability not to drink the toxin could be causally independent of your drinking the toxin. They will be inclined to deny that the second premise – that Fx \rightarrow Cx(¬Fx) – is logically possible. What reason might they have for this denial? There seems, in fact, to be only one. It is false that, possibly, if you were to drink the toxin then you would still be able not to drink the toxin, because, necessarily, if you were to drink the toxin then you would not be able not to drink it. In other words, your ability not to drink the toxin is not causally independent of your drinking it, because the very fact that you are drinking the toxin entails that you cannot at the very same time refrain from drinking it.

[1.2] Denote by 'p \Rightarrow q' the claim that p entails q. The validity of the above inference, (C), therefore hinges on whether or not the following is a correct general principle:

\[
\neg \text{C}_x(Fx \& Gx) / \therefore \ Fx \rightarrow \neg \text{C}_x(Gx)'. \quad \text{Since this is an inference logically stronger than it needs to be to make the relevant objection, I will only consider the one I do in the text, (C).}
\]

\text{On the standard definition, p is causally independent of q if and only if p} \equiv (q \rightarrow p) \& (\neg q \rightarrow p). \quad \text{Hence, your being able not to drink the toxin – Cx(¬Fx) – is causally independent of your drinking it – Fx – if and only if Cx(¬Fx) \equiv (Fx \rightarrow Cx(¬Fx)) \& (¬Fx \rightarrow Cx(¬Fx)). \quad \text{It follows (as I shall claim in a moment) that, if you are able not to drink the toxin, and this ability is causally independent of your drinking it, then Fx \rightarrow Cx(¬Fx).}
(D) \[ Fx \Rightarrow \neg C_x(\neg Fx) \]

My opponent thinks it is true, and that my counterexample (C) fails. I think this principle is false, and thus that my counterexample to inference (C) stands. Which of us is right?

It is easy to see that it is my opponent, rather than myself, who is mistaken. Principle (D) entails, by substitution and rearrangement, that \[ C_x(Fx) \Rightarrow Fx. \] It entails, in other words, that the only things we can do are the things we actually do. This is implausible. ‘Ought’ implies ‘can’:

(0) \[ [= (OC_1)] \]

\[ O_x(Fx) \Rightarrow C_x(Fx) \]

That is to say, an agent x ought to F only if they can be such as to F. It would follow, were my opponent correct, that an agent ought to perform an action only if they are in fact performing that action. If my opponent were right, this would mean agents never perform irrational actions, but this, surely, is not the case.

My opponent will still not be satisfied. We need to be careful, they will insist, about the exact temporal indexing of actions, abilities, and ‘ought’s. In particular, the claim they wish to endorse is that

(Dt) \[ F_{tX} \Rightarrow \neg C_{x,t}((\neg Fx)) \]

That is, if x F's at time t, then x cannot at time t not F at time t. This, they will insist, is obviously true – you cannot at some time drink, and

---

6 There are other forms a principle like (D) might take. Philip Pettit, in private conversation, suggests, for example, that if an agent believes they will F, then not-F no longer remains an option. This claim will be relevant to the discussion only if it is to be interpreted as the claim that if an agent believes they will F, then it is false they can not-F. A claim such as this may be refuted by employing arguments similar to those below I use to refute (D), and so I will not consider this suggestion separately.

7 Interestingly, Gauthier at one time seems to have endorsed a principle such as this. In The Logic of the Leviathan, (Oxford: Clarendon Pr., 1969), he says: ‘Deliberation terminates in action. But when we act, then we are no longer at liberty to act or not to act. Although our capacities are unchanged, yet the ‘external circumstances’ of our acting prevent us from choosing between acting or not’ (64). I do not know whether he still holds this view; but, even so, this is precisely the sort of position I am concerned to deny. In essence, I claim that our own actions do not constitute ‘external circumstances’.

8 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 necessary.
at the very same time not drink, the toxin. Things are slightly more complex, though, when it comes to the interpretation of the doctrine that 'ought' implies 'can'. They will, if reasonable, agree that something like this doctrine is correct, but will want to dispute its interpretation. In particular, they will reject the interpretation which I gave to principle (O), that the doctrine that 'ought' implies 'can' is to be given by:

\[(O^a) \quad O_{x,t}(F_t|x) \Rightarrow C_{x,t}(F_t|x)\]

In other words, that x ought at time t to F at t only if x can at time t F at time t. As we saw, (D^b) and (O^a) entail the implausible conclusion that, necessarily, agents always do the rational thing. My opponent will urge a slightly different interpretation of the doctrine that 'ought' implies 'can' and will insist the correct interpretation is in fact:

\[(O^b) \quad O_{x,t}(F_t|x) \Rightarrow C_{x,t-F_t}(F_t|x)\]

In other words, that x ought at time t to F at t only if x can at some time slightly before t (say time t-) F at t. On this view, 'ought' implies 'could have', and my opponent will have been delivered, it seems, from the problem.

This modified interpretation of the doctrine that 'ought' implies 'can' will not do. I am in jail, and coming up to the end of a long sentence. There are no phones in the jail for prisoners, and the first thing that I should do when I get out, next Monday, is to call my mother. What is true is that I ought on Monday to call my mother, and that I can call my mother on Monday. What is also true is that there is no time before Monday at which I can call her. The first time at which I obtain the relevant ability is on Monday, and this is the time at which I ought to call my mother. The antecedent of the proposed reinterpretation of the doctrine that 'ought' implies 'can' is true – I ought on Monday to call my mother, and that I can call my mother on Monday. What is also true is that there is no time before Monday at which I can call my mother. The first time at which I obtain the relevant ability is on Monday, and this is the time at which I ought to call my mother. The antecedent of the proposed reinterpretation of the doctrine that 'ought' implies 'can' is true – I ought on Monday to call my mother – and yet the consequent is false – there is no time before Monday at which I can call her. The proposed reinterpretation, (O^b), of the doctrine that ought implies can is incorrect. As a consequence, my objection to the problematic inference, (C), stands.

At this point, my opponent may claim that counterexample I have provided to (C), even if it were correct, is not to the point, because it is
put in terms of conflicting states at the same time as one another, rather than, as it needs to be, in terms of states at different times. Inference (C), they will insist, is really of the form:

\[
(C^1) \quad \frac{C_{x,t1}(F_{t1}x), \ F_{t1}x \rightarrow C_{x,t2}(G_{t2}x), \ t_1 < t_2}{C_{x,t1}(F_{t1}x \& G_{t2}x)}
\]

In other words, given the premises that x can F at time t1, and, even if x F's at t1 then he can G at time t2, where t1 is before t2, the conclusion follows that x can at t1 both F at t1 and G at t2. My opponent will complain that the Jail Counterexample, even if otherwise correct, would only be a counterexample to the principle if it allowed that t1 = t2. The relevant example, though, has you adopting an intention (or not) at midnight, and acting (or not) the following afternoon.

The considerations enunciated suggest that inference (C) is invalid, and that at the very least it is my opponent’s responsibility to justify it. My opponent, in response to this point, has not discharged this responsibility. In any case, it is even easier to see that this inference is invalid, if we consider the form this type of objection would have to take against the Second Counterexample.

[2] What form does this objection take to the Second Counterexample? I assumed you can, in fact, both choose to be disposed to doing what you intend to do, and then later, choose what you would intend to do, and then added that, even if you chose both of these options, you would still be free not to drink the toxin. Thus, I assumed that:

\[(T4(b')) \quad \text{Even if you adopt this disposition, and even if you adopt the intention later, you would still be free not to drink the toxin.}\]

The objection in this second case is that this is inconsistent. The issue between my opponent and I is whether or not the following statement is logically possible:

\[(Fx \rightarrow Gx) \& Fx \rightarrow C_x(\sim Gx)\]
My opponent thinks it is not logically possible, and that the Second Counterexample does not work. I think this statement is perfectly possible. Which one of us is right?

My opponent denies it is logically possible that \((Fx \rightarrow O \rightarrow Gx) \& Fx \rightarrow Cx(-Gx)\). But what reason might they have for making this denial? There seems, in fact, only to be one. It is not logically possible that \((Fx \rightarrow Gx) \& Fx \rightarrow Cx(-Gx)\), since it is necessarily true that \((Fx \rightarrow Gx) \& Fx \rightarrow -Cx(-Gx)\). In other words, your ability not to drink the toxin is not causally independent of whether or not you adopt the relevant disposition and intention, because if you adopt the relevant disposition and adopt the intention, it is false you can be such as not to drink the toxin. The very fact that you (a) are disposed to doing what you intend, and (b) intend to drink the toxin, means that (c) you cannot be such as not to drink the toxin. The validity of the relevant inference thus hinges on whether or not the following is a valid principle:

\[
(E) \quad (Fx \rightarrow Gx) \& Fx \Rightarrow -Cx(-Gx)
\]

My opponent thinks it is valid, and that the Second Counterexample is inconsistent. I think this principle is invalid, and thus that it stands. Which of us is right?

Here is a simple counterexample to \((E)\). If my butcher were to believe I asked him for a pound of ground beef, then he would give me a pound of ground beef \((Fx \rightarrow Gx)\). I have asked him for a pound of ground beef and, aware of me, he believes I have asked him \((Fx)\). He gives me a pound of ground beef \((Gx)\), but this is not to say that he was compelled to do so. Not at all. He was free not to give me a pound of ground beef \((Cx(-Gx))\), but, thankfully, he is disposed not to exercise this freedom. It makes perfect sense to suppose one has a certain capacity one is disposed not to exercise (eg, my ability to put out my cigarette on this young child’s arm). Principle \((E)\) is false, and so my Second Counterexample stands.

All of this is, of course, very familiar. Above, I discussed the closely related principle, \((D)\), which says that \(Fx \Rightarrow -Cx(-Fx)\), arguing that it is false. Exactly the same type of argument shows that the more complex form of the principle, \((E)\), is also false. Above, my opponent responded with the claim that, even if the counterexample were correct, it would not be to the point since it is put in terms of internal states supposed to occur at the same time. But, in the case I have just discussed, such a
response is not available. The time of the adoption of the butcher’s disposition (Fx $\rightarrow$ Gx) is presumably long since past, the time he comes to believe I have asked for some meat (Fx) has just past, and the time of his giving me the meat (Gx) is right now. Once again we see that my arguments against the truth of the Self-Interest Theory stand.

As a consequence, we can come up with yet another counterexample to the original inference, (C). Making relevant substitutions (namely, ‘((Fx$\rightarrow$Gt2x) & Fx)” for ‘Fx’, and ‘~Gx’ for ‘Gx”) in inference (C), we get:

\[
\begin{align*}
C_{x,t1}(\text{(Fx} \rightarrow \text{Gt2x)} & \text{& Fx), } \text{(Fx} \rightarrow \text{Gt2x)} & \text{& Fx) $\rightarrow$ C_{x,t2}(\sim\text{Gt2x})} \\
(\text{C”}) && \therefore C_{x,t1}(\text{(Fx} \rightarrow \text{Gt2x)} & \text{& Fx & } \sim\text{Gt2x})
\end{align*}
\]

Presumably it is possible that one can both adopt a disposition to perform certain sorts of actions (Gt2x) upon a certain condition (Fx) and then, afterwards, have the disposition triggered, and so the first premise, $C_{x,t1}(\text{(Fx} \rightarrow \text{Gt2x)} & \text{& Fx), is true;}$ and, if the argument I have just presented is right, then even if this is what happens, it is possible that one can do otherwise, so the second premise, ((Fx $\rightarrow$ Gt2x) & Fx) $\rightarrow$ C_{x,t2}(\sim Gt2x), is true; but, obviously, the conclusion, that $C_{x,t1}(\text{(Fx} \rightarrow \text{Gt2x)} & \text{& Fx & } \sim\text{Gt2x}), is false.

The Self-Interest Theorist might claim that the stipulative additions I made to the Toxin Puzzle, (T2), to get the First and Second Counterexamples mean that you are no longer free not to drink the toxin, but, as we have now seen, such a claim is mistaken. The considerations above show inference (C) is not valid, and that to suppose so is to make a simple modal error. For all my opponent has said, in the First and Second Counterexamples you may indeed still be free not to drink the toxin.

§2 Does it really still have the best outcome for you not to drink?

I constructed the First Counterexample by adding to the Toxin Puzzle, (T2), the stipulation that you cannot adopt the intention to drink the toxin without later drinking it, to get (T3a). I constructed the Second by adding the stipulation that it maximises expected-value for you to adopt the enduring disposition of always doing what you intend, to get (T3b). I claimed that, in these modified situations, you are rationally
permitted to drink the toxin, even though you are free not to do so, and it maximises expected-value for you not to do so. The Self-Interest Theorist, however, may now object that these two stipulative additions mean *it no longer has the best outcome for you* not to drink the toxin. As I shall argue in this section, though, this objection is also mistaken; and in the First and Second Counterexamples it may indeed still have the best outcome for you not to drink the toxin.

[1] The problem is that your drinking the toxin seems unintelligible, unless we suppose that, somehow, it now has the best outcome for you to drink the toxin. If it is still the case that the only relevant outcome of drinking the toxin is day’s severe illness, it seems there is no way you could rationalise drinking it, and so there is no way you can be said to do it intentionally. This objection may take a number of forms.

First, the objection might be that, since you believe the effects of drinking the toxin will be a day’s severe illness, then presumably you do not believe there is good which will come of drinking the toxin. This means that you would not drink the toxin intentionally, since

(G1) In the Toxin Puzzle Case, you drink the toxin intentionally only if you believe there is good which will come of it.9

This objection fails. On the one hand, (G1) is not true. Consider Sophie, famous of William Styron’s *Sophie’s Choice*, who must decide which of her children she is to give to the guard, to be sent to the gas chambers. Will it be her girl or her boy? In the end, she gives her girl to the guard. In particular, Sophie’s giving the girl to the guard is an intentional act, even though Sophie did not believe (we may assume) that good would come of it. On the other hand, even if (G1) were true, the objection would still fail. What you know is that the effects of drinking the toxin include a day’s severe illness, but it doesn’t follow from this that you do not believe there is good to come of drinking the toxin. Suppose, for example, you are mildly curious about the taste of

---

9 See W. Quinn, ‘The Right to Threaten and the Right to Punish,’ *Phil Pub Affairs* 14 (1985): 327-73, who asks: 'how an agent who follows this policy [i.e., intending to drink] is to think of his reasons. Is N [the potential toxin-drinker] at the later time to think that he has a good reason to drink the toxin despite the fact that no good will come of it? And if so, what does this good reason amount to? Or is N to think that this kind of choice can be rational in the absence of any reasons to make it? Neither option seems to me inviting' (371, fn. 54).
the toxin, though your curiosity is not so strong that its satisfaction is more valuable to you than the ill-effects of drinking it. Then, for all that principle (G1) says, you could drink the toxin intentionally, believing that there is good to come of it (namely, that you would satisfy your curiosity), even though drinking the toxin still does not maximise expected-value.

Second, the objection might be that, since you believe the effects of drinking the toxin will be a day's severe illness, then presumably you value not drinking more than not drinking it. This means that you would not drink the toxin intentionally, since

(G2) In the Toxin Puzzle Case, you drink the toxin intentionally only if
the value to you of drinking it is no less than that of not drinking it.10

This objection also fails. On the one hand, (G2) is not true. Here one may (controversially) cite the possibility of weakness of will, which is intentional action performed against one's better judgement. One does some act A intentionally, even though there is another, B, one took to be better.11 On the other hand, even if (G2) were true, the objection would still fail. What you know is that the only outcome of drinking the toxin includes a day's severe illness, but it does not follow from this that the value of drinking it is less than that of not drinking it. What may be true is that an action is intentional only if the agent values the action most; but it is not true, and does not follow from this, that an action is intentional only if the agent values the outcome of action most. The agent may think – and, I shall argue, may be right to think – that bestness of actions consists in more than just bestness of outcomes of actions. I reiterate my claim – first made in Chapter One, §3 – that the notion of the expected-value of an action is conceptually distinct

10 See J. Mendola, 'Gauthier's Morals by Agreement, and Two Kinds of Rationality,' *Ethics* 97 (1987): 765-74, who says that: 'the conception of rationality as individual utility maximization does little more than capture the truism that we do what we most want' (767). As we will see in a moment, it may be a truism that we do what we most want, but this is no support for the consequentalism underlying the Self-Interest Theory, since the truism fails to refer to consequences.

11 Indeed, J. Thomas, in 'The Toxin, the Blood Donor, and the Bomb,' *Analysis* 43 (1983): 207-210, suggests that one can get the million by becoming weak-willed. This is not the way I am inclined to argue. For the agents I am considering – who, within a certain time period, always do what they intend to do – are best described as resolute, even bloody-minded, rather than weak.
from that of the *value* of an action. Then, for all principle (G2) says, you could drink the toxin intentionally, value doing so to not doing so, even though drinking the toxin does *not* maximise expected-value.\(^\text{12}\)

Third, and finally, the objection might be that your action in not an intentional one, unless we now assume it maximises expected-value for you to drink the toxin. The issue is whether or not the following claim is true:

\[(G3) \text{ In the Toxin Puzzle Case, you drink the toxin intentionally only if the expected-value of drinking the toxin is no less than that of not drinking.}\]

If statement (G3) were true, then my opponent would have established that my counterexamples are, indeed, incoherent. It is this statement which lies at the heart of the dispute.

Statement (G3), though, is not true. Suppose you are aware of all the relevant outcomes: if you were to drink the toxin, then the outcome would be that you suffer a day's severe illness; if not, not. Suppose further you value the outcome of not drinking to that of drinking. We ask you, and this is what you say. Realising the benefits attaching to adopting the intention to drink, however, you (somehow) adopt this intention. (Perhaps you have a special pill, close at hand.) You intend to drink the toxin, and are aware this is what you intend. Suppose, finally, you believe (rightly or wrongly) that the fact you now intend to drink the toxin is a sufficient\(^\text{13}\) reason, in and of itself, to drink the toxin. Come the time for drinking, you drink. In this case, (G3) is false. On the one hand, the antecedent is true: you drink the toxin, and do so intentionally, since (roughly) your drinking the toxin is caused, presumably in the right way, by there being a consideration (namely, that you intend to drink the toxin) you took to be the case, and took to be (rightly or wrongly) a sufficient reason for you to drink

\(^{12}\)This is a move E. F. McClennen, in 'Constrained Maximization and Resolute Choice,' *Soc Phil Pol* 5 (1988): 95-118, is also inclined to make. Discussing the Prisoner's Dilemma, McClennen says: 'I start with the assumption that the agents' preferences for outcomes conform to the pattern of a classic Prisoners' Dilemma situation. I simply move from there to challenge the distinct assumption that preferences for outcomes, abstractly considered, must be taken as controlling for preferences over actions and, hence, for choice' (115).

\(^{13}\)I shall say that \(p\) is a *sufficient* (conclusive) reason to \(A\) when the expected-strength of \(p\) as a reason to \(A\) is no less (greater) than the expected strength of reasons for any alternative to \(A\). If there is a sufficient (conclusive) reason to \(A\), then you are rationally permitted (obliged) to \(A\).
the toxin. From your perspective, you had all the reason you needed to drink the toxin. On the other hand, the consequent is false: the expected-value of drinking the toxin is less than that of not drinking, since you are completely aware of the fact that the only outcome of drinking the toxin will be a day's severe illness, and you would rather not suffer this outcome. From your own perspective, however, this reason against drinking is not of greater strength than the reason you take yourself to have for drinking – namely, that you intend to do so. Intentional action, then, is not necessarily action which maximises (agent-relative) expected-value, but is, if anything at all, action which maximises (agent-relative) expected-strength-of-reasons.

My opponent may not be impressed. I claimed you drink the toxin intentionally, even though this is not value-maximising, if the fact you intend to drink the toxin gives you (at least by your own lights) an extra, non-consequentialist, reason to drink. My opponent may claim, though that while this intention does really give you a reason, it is in fact a consequentialist reason.

[2] In short, the suggestion is if you drink the toxin, your action manifests the value you place on fulfilling intentions. Unconvinced by the above counterexample, the Self-Interest Theorist might offer an argument for the principle (G3).

Here is one such argument. Clearly, if you are aware of the details of the situation, then you will be aware that the outcome of drinking

---


15 This is a response the so-called revealed preference theorists will want to press. There may be other forms the response could take; I shall concentrate on the one appearing to me to be the most plausible. R. Sugden briefly considers such a response in his 'Rational Choice: A Survey of Contributions from Economics and Philosophy,' Econ J 101 (1991), p. 782. He replies, in essence, that fulfilling or failing to fulfill an intention should not be considered part of the outcome of drinking the toxin on standard formulations of decision theory, such as that of L. J. Savage, The Foundations of Statistics, (New York: Dover, 1967). Only because I do not want to enter into debates about the nature of outcomes, I shall provide a different response.
the toxin is that you fulfil your intention but suffer a day's severe illness, and that the outcome of not drinking that you fail to fulfil your intention but avoid the suffering. However, if you are aware that these are the relevant outcomes, and yet you drink the toxin, then you must (implicitly or explicitly) value fulfilling your intentions no less than the suffering that may be involved in doing so. Furthermore, in a Toxin Puzzle Case you are aware of the details of the situation. Thus, in a Toxin Puzzle Case, if you drink the toxin then (a) you must be aware that the outcome of drinking is the fulfilment of intention plus suffering and the outcome of not drinking is the non-fulfilment of intention without suffering, and (b) you must (implicitly or explicitly) value fulfilling your intentions no less than the suffering involved in doing so. However, if (a) and (b) are true, then, by definition, the expected-value of drinking the toxin is in fact no less than that of not drinking. It follows that, in the Toxin Puzzle Case, if you drink the toxin then the expected-value of doing so must be no less than that of not drinking. Statement (G3) is true.

In response to this argument, I deny its second premise – that if you are aware of the relevant outcomes, and yet you drink the toxin, then you must (implicitly or explicitly) value fulfilling your intentions no less than the suffering that may be involved in doing so. To see how to deny this premise, consider the question: is the relevant value instrumental, or intrinsic?

On the one hand, my opponent may think your drinking the toxin reveals that you value fulfilling your intentions instrumentally, as one may value keeping promises instrumentally. There will, after all, be a cost to changing one's mind, and this cost may well typically mean that the instrumental value of making a decision and sticking to it will be greater than that of making a decision and not sticking to it.

However, it is clear in the case at hand that the costs of changing the intention to drink the toxin need not – and probably will not – mean that the instrumental value of fulfilling the intention will be greater than that of not fulfilling it. After all, the effect of drinking the toxin is a day's severe illness, and the inconvenience caused by changing one's mind about whether to drink are likely to be more than outweighed by this. The defender of the Self-Interest Theory would be wrong to insist that, if you drink the toxin, you must instrumentally value fulfilling your intentions to not fulfilling them.
On the other hand, the defender may think it reveals you value fulfilling your intentions *intrinsically*, as one may value keeping promises intrinsically. The suggestion is that at the time for drinking the toxin, you weigh (implicitly or explicitly) the outcome of drinking against that of not drinking it, and, since you intend to drink the toxin and *intrinsically* value fulfilling your intentions, you end up drinking the toxin.

The quick response to this suggestion is to claim that, if the Self-Interest Theorist needs to make the descriptive supposition that people intrinsically value fulfilling their intentions in order to defend his normative claims concerning rationality, then his position is rendered trivial. As we saw in Chapter One, in order to avoid the charge of triviality, consequentialists should be at pains to distinguish their position from that of a deontologist such as Ross—who thinks that (some) promises are, in and of themselves, sufficient reasons for action—by insisting that promise-keeping is not intrinsically valuable. In order to avoid a similar charge of triviality, defenders of the Self-Interest Theory should be at pains to distinguish their position from that, for example, of Micheal Bratman and Joseph Raz—who think that (some) intentions are, in and of themselves, sufficient reasons for action—by insisting that intention-fulfilling is not intrinsically valuable.

The not-so-quick response to this suggestion is to claim the action-guiding role it attributes to intentions is, in any case, mistaken. The role suggested is that the fact one intends to A is just one more fact to be weighed up in (implicit or explicit) deliberation about whether to A. If one values fulfilling one’s intentions enough, then the expected-value of drinking will exceed that of not drinking, and, weakness of will to one side, this will guide one to drink the toxin; if not, not. This suggestion is mistaken, though, for anyone who weighed (whether implicitly or explicitly) the outcome of drinking the toxin against that of not drinking it would, *ipso facto*, be someone who at that time did not have the intention to drink the toxin. The only persons inclined to

---

16 The picture of the causal role of intention I introduce in this paragraph depends heavily on Joseph Raz’s notion of an exclusionary reason, and Michael Bratman’s notion of a framework reason. See J. Raz, *Practical Reason and Norms*, (London: Hutchinson, 1975), and M. Bratman, *Intention, Plans, and Practical Reasoning*, (Cambridge: Harvard Univ. Pr., 1987). See also J. Kilcullen, ‘Utilitarianism and Virtue,’ *Ethics* 93 (1983): 451-466 for similar claims. I do not mean to suggest these authors would endorse the interpretations I put on their positions.
reason (whether implicitly or explicitly) in the way the Self-Interest Theorist suggests are exactly those who have not yet decided what to do. The role intentions actually play is in pre-empting any (implicit or explicit) consideration of actions inconsistent with the fulfilment of that intention. If an agent intends to A, then any further (implicit or explicit) deliberation about whether or not to A will typically be pre-empted, even though such actions may be ones the agent is perfectly free to do and even though the agent believes they may be ones it has the best outcome for the agent to do. On this view, the agent places priority on fulfilling intentions, rather than a premium. Thus, an intention to A is a reason guiding the performance of A, but not through (implicit or explicit) deliberation concerning whether to A. The defender of the Self-Interest Theory would thus be wrong to insist that, if you drink the toxin, you must intrinsically value fulfilling your intentions to not fulfilling them.

The Self-Interest Theorist might claim that the stipulative additions I made to the Toxin Puzzle, (T2), to get the First and Second Counterexamples mean that it no longer maximises expected-value for you not to drink the toxin, but, as we have now seen, such a claim is false. The considerations above show one can indeed act intentionally against what maximises expected-value, and that to suppose that this is not possible is to have an overly simplistic view of the relation between values, intention and action. For all my opponent has said, in the First and Second Counterexamples it may indeed still have the best outcome for you not to drink the toxin.

§3 What reason is there for you to drink the toxin?

The conclusion of the previous chapter stands: you are rationally permitted to drink the toxin, even though you are free not to do so, and even though it has the best outcome for you not to do so. This, though, raises a question: if you really are rationally permitted to drink the toxin, then what reason could you possibly have to do so? I shall argue in this section that in the First and Second Counterexamples, and given some conditions, the fact you intend to drink the toxin is reason enough for you to drink.

[1] My claim is that if you adopt the intention to drink the toxin in the First Counterexample, or you adopt both the disposition and the
intention in the Second, then the fact you intend to drink the toxin is (sufficient) reason for you to drink it. It is important to note two things about this claim. First, I am making a claim about the reasons you have only in those situations described in the First and Second Counterexamples, and am not claiming that, in all cases, intending to do something is a sufficient reason to do that thing. Second, even in this restricted class of cases, I claim only that you have reason to drink the toxin if the specified conditions are given, and am not claiming you have reason to drink, even if you do not have the relevant intention, or (in the Second Counterexample) you do not have the relevant disposition.

Here is an argument that there must be some non-consequentialist reason for drinking the toxin. To this point I have been using what I have called the standard formulation of the Self-Interest Theory, which (upon suitable substitution), is that

(S) If you are free not to drink the toxin, then you rationally ought not drink the toxin if and only if the (agent-relative) expected-value of not drinking the toxin exceeds that of doing any alternative.

Statement (S) is false: you are free not to drink the toxin, it maximises value for you not to drink the toxin, and yet you are rationally permitted to drink the toxin. Statement (S) follows (upon suitable substitution) from two others:

(R) If you free not to drink the toxin, then you rationally ought not drink the toxin if and only if the expected-strength of reasons for not drinking the toxin exceeds that of doing any alternative.

---

17 Due to discussions with Andre Gallois, Peter Menzies, and Robert Dunn, I need to signal two complications. I assume the reason you have for drinking the toxin is that you intend to drink it. (a) Typically, though, you do not say that the reason for doing A is that you believe that p, but, rather, simply that p. That is, typically, the reason to act is not the fact that you have some intentional state (eg, belief, desire, intention, ...) but is rather the propositional object of that state. See, for example, J. Rachels, 'Reasons for Action,' Can J Phil 1 (1972), pp. 174-6, and, R. Abelson, 'Doing, Causing, and Causing to Do,' J Phil 66 (1969): 178-92. Similarly, you might argue, the reason for you to drink the toxin is not that you intend to drink the toxin, but, rather, simply that you are going to drink the toxin. (b) The fact you intend to A provides a reason, it may be suggested, only for those actions which are a means to A, and is not a reason, strictly speaking, for A itself. These complications, if correct, would require only minor adjustments to the argument, and so I shall ignore than henceforth.
(S') A consideration \( p \) is a reason for or against drinking the toxin if and only if it takes the form ‘\( q \) would be (part of) the outcome of your drinking the toxin’; the weight of such a consideration is to be given by your expectation that \( q \) would be (part of) the outcome of your drinking the toxin; and the (intrinsic) strength of such a consideration is given by the value to you of outcome \( q \).

Hence, one of these latter statements must also be false. I won’t repeat the reasons I gave for believing \((R)\); I conclude that \((S’)\) — the Self-Interest Theory’s view concerning the reasons there are in the present case — should be abandoned. This statement itself consists of basically two claims: one about the sorts of things which are reasons (namely, statements about outcomes of drinking), and the other about the sorts values and expectations which are relevant (namely, yours). One of these claims must be false. It should be clear, though, that the two Counterexamples do not really touch the agent-relative aspect of the Self-Interest Theory, since they can obviously be reworked to involve only agent-neutral expectations and values. The two Counterexamples refute, rather, the consequentialist aspect of the theory — that all reasons for drinking the toxin need to mention possible outcomes of drinking the toxin. There must be some non-consequentialist reason for drinking the toxin.

My further speculation, then, concerning this non-consequentialist reason for drinking the toxin is that the reason is simply the fact you intend to drink the toxin. The idea that intentions provide one with extra, non-consequentialist, reasons for action is not a new one, and I want to add my own examples in support of it.\(^{18}\)

It is important to be clear about the form of my argument. It does not depend on an inference from the rationality, or the fact, of your believing that your intention is a reason to drink the toxin to that of its actually being a reason. This would be to invoke just another bridging principle, which Parfit claims\(^{19}\) (and I agree) is false. Rather, I am arguing that (a) we may use your belief that it is a reason to defeat an objection to my position, allowing me (b) to reassert my claim that you

---


are rationally permitted to drink the toxin, and thus leading to my speculation (c) that your intention is in fact a reason to drink the toxin. There are, however, those who would object to this speculation. In the remainder of this chapter, I will consider what they have to say.

[2] Daniel M. Farrell, in a discussion of the Toxin Puzzle case,\(^\text{20}\) comes at one point in his paper to consider the position of the person who – like myself – suggests that ‘while drinking the toxin would be irrational if one had not previously adopted and maintained an intention to drink it, drinking it would not be irrational in light of a prior (and rational) intention to drink it’ (291). Farrell canvasses the following argument in support of this suggestion:

(1) Intention-adoption is itself an intentional action (or, at any rate, is something one does “at will”).

(2) In circumstances of the sort that interest us, it would be rational to adopt the relevant intention (if one could), given the desirability of having it, despite the fact that it is an intention to do something it appears it will not be rational for one to do when the time comes to do it.

(3) Intentions it is rational to adopt are intentions it is rational to have.

(4) Intentions it is rational to have are intentions on which it is rational to act, provided the conditions under which one is called upon to act are the conditions one anticipated in one’s intention so to act. (292)

Farrell thinks (and spends earlier sections of his paper arguing) that premise (1) is false – that intention-adoption is not itself an intentional action – but concedes the premise for the sake of argument. Premise (2), of course, is just the description of the type of situation with which we are concerned, but note that to avoid begging the question Farrell needs to say (as he does in fact say), that the intention is an intention to ‘do something it appears it will not be rational for one to do when the time comes to do it’ (292, emphasis added). And Farrell says premise (4) ‘can be made good, though I shall not attempt to defend it here’ (292).

This leaves only premise (3), which Farrell denies. This premise seems to licence the inference from the claim it is rational to adopt the intention (with which Farrell, for argument’s sake, agrees) to the claim

that it is rational to have the intention (with which he does not). 'But
why should we suppose this is so? Why not suppose, instead, that in
intentionally (and, let us suppose, rationally) adopting the desired
intention, the agent has brought it about that she is less than fully
rational? After all, we know from elementary game-theory that there
are circumstances in which it would be rational to do just that' (292).
Farrell refers to the type of situation, described in Thomas Schelling's
*The Strategy of Conflict*, where it might be rational for a smaller nation
to intend to retaliate — irrationally — if a large nation threatens its
security. Farrell concludes there is nothing in the above argument to
suggest that anything other than this is what is involved in the Toxin
Puzzle Case.

Fortunately, my position is not susceptible to this challenge. On the
one hand, I am inclined to agree with Farrell in his rejection of the
third premise — that intentions it is rational to adopt are always
intentions it is rational to have. This, it seems to me, is yet another
bridging principle the defenders the Self-Interest Theory may be
inclined to reject. I, certainly, would not be concerned to defend it. On
the other hand, though, I can accept this point with equanimity, since
the argument above is not the argument I am inclined to give for the
rationality of drinking the toxin. In the previous chapter, I eschewed
use of any bridging principles in my attempt to establish the possible
rationality of drinking the toxin, and employed in their stead certain
deontic principles. Further, I must emphasise again that I make no
general claim about the situations Kavka originally described, but I
claim only that it is rational to drink the toxin in the First and Second
Counterexamples. Farrell's refutation of the argument above, then,
passes me by.

There is, however, a tangential point of relevance to my argument.
Farrell thinks false, though for the sake of argument accepts as true,
premise (1) — as he puts it, that 'intention-adoption is itself an
intentional action (or, at any rate, is something one does “at will”)
'(292, emphasis added). In this, Farrell suggests that one can defend the
rationality of the intention only if one defends the rationality of
adopting the intention, and that one can do this only if one
understands the adoption of an intention as an act ‘of will’. But it is

I shall have more to say about such situations in Chapter Nine.
clearly implausible, as Farrell points out, that adopting the intention to
drink the toxin could be an act done 'at will'. In response, though, what
reason do we have for supposing that adopting the intention needs to
be understood as an act done 'at will'? None at all, I suggest. There are
many actions one performs, and rationally evaluable ones at that, not
plausibly understood as actions done 'at will'. Going to the cinema,
driving a car, and posting a letter, are examples. Adopting the
intention to drink the toxin need not be understood as an act 'at will',
but rather as a more complex action which is, nonetheless, rationally evaluable.

[3] Michael E. Bratman, in the course of his discussion of the nature of
intention,\(^{22}\) comes eventually to a discussion of the Toxin Puzzle. He
does so because it seems to follow from principles he endorses that it is
rational to drink the toxin, and this is a claim he is most concerned not
to have to endorse. My discussion of Bratman will proceed as follows.
[3.1] I describe the argument he introduces purporting to show it is
rational to drink the toxin, and introduce Bratman's response. [3.2] I
then examine the implications of his response for the argument I have
given for the rationality of drinking the toxin.

[3.1] First, I present the argument Bratman introduces and which
purports to show it is rational to drink the toxin. Bratman says the
most powerful way of developing the Toxin example as an objection to
his theory is on the basis of the following argument, which I quote at
length:

...we have the following quartet of premises:

(1) It is rational of you on Tuesday deliberatively to intend to drink the
toxin on Wednesday, given your strong desire-belief reasons on Tuesday for so
intending then.

(2) If it is rational of you on Tuesday deliberatively to intend to drink the
toxin on Wednesday, and if you find the relevant conditions on Wednesday to

\(^{22}\) In his *Intention, Plans, and Practical Reasoning*, (Cambridge: Harvard Univ. Pr.,
1987). Pages references in this subsection will be to this work. Even though
Bratman thinks intentions can be, in and of themselves, reasons for action, he
would likely object to my argument on the basis that it allowed actions -- such as
drinking the toxin -- to be 'bootstrapped' into rationality. One implication of the
argument in this thesis is that Bratman would be wrong to make this point.
be just as on Tuesday you expected them to be, then it is rational of you on
Wednesday not to reconsider that intention then.

(3) Conditions as you find them on Wednesday are just as on Tuesday you
expected them to be, and that is why on Wednesday you do not reconsider
your prior intention to drink the toxin.

(4) Both the historical principle of nondeliberative rationality for the
basic case and the intention-action principle are correct.

From (1) through (3) we can infer that it is rational of you on Wednesday not to
reconsider your prior intention to drink the toxin. Since your nondeliberative
intention on Wednesday is (we may assume) an instance of the basic case, we can
apply the historical principle of nondeliberative rationality. This tells us that
it is rational for you on Wednesday to intend to drink the toxin then. The
intention-action principle then tells us that it is rational of you intentionally to
drink the toxin. So from (1) through (4) we can infer:

(5) It will be rational of you on Wednesday intentionally to drink the toxin.

Since this condition seems seriously wrong and the argument seems valid, one of
the premises must be rejected. ... (102-3).

I do not have the space, and nor do I need, to introduce what, in
premise (4), Bratman calls the 'historical principle of nondeliberative
rationality for the basic case' and the 'intention-action principle'23 – my
discussion is comprehensible without detailed introduction to these
principles. My interest in Bratman's discussion stems from the fact
that, if he is right, then there will be no justification for supposing in
any Toxin Puzzle Case that your intention to drink is a reason for you
to drink.

How does Bratman respond to this argument? He begins by
claiming that premise (1) is false as it stands. In Bratman's parlance, it
is rational for you on Tuesday deliberatively to intend to drink the
toxin only if you can reasonably suppose that drinking the toxin is at
least as well supported by your desire-belief (or what I call conse­
quentialist) reasons as its admissible alternatives.24 In general, it is

23 For an explanation, see Bratman, Intention, Plans, and Practical Reasoning, pp. 79-
83 ('historical principle of nondeliberative rationality for the basic case'), and pp.
53-5 ('intention-action principle').

24 Admissible alternatives, on Bratman's theory, are those things one can do and the
doing of which is not inconsistent with the satisfaction of any of the intentions one
already has. Pre-existing intentions provide a filter (as it were) through which
available actions must pass if they are to be 'admissible' for deliberation. See,
rational of an agent deliberatively to intend to A only if it is rational for them to form that intention on the basis of deliberation concerning whether to A, and such deliberation will refer only to desire-belief reasons. Hence, it is rational for you on Tuesday deliberatively to intend to drink the toxin only if it maximises expected-value for you to drink the toxin. But it does not maximise expected-value for you to drink the toxin, so premise (1) is false.

Premise (1) therefore needs to be modified, if the argument is to be sound. It is simplest just to omit the word 'deliberatively'. If we do so, we get a new first premise:

(1') It is rational of you on Tuesday to intend to drink the toxin on Wednesday, given your strong desire-belief reasons on Tuesday for so intending then to do so.

This premise seems true, and Bratman does not dispute it. But, if the argument is to remain valid, there must be further changes in the other premises. In particular, we will also need to excise the word 'deliberatively' from premise (2), to get (as Bratman puts it):

(2') If you acquire on Tuesday the intention to drink the toxin on Wednesday, and if on Tuesday this intention is rational of you to have, and if you find the conditions on Wednesday to be just as on Tuesday you expected them to be, then it is rational of you on Wednesday not to reconsider that intention then.

The problem, now, according to Bratman, is that this necessarily altered premise is also not true. This premise says, in effect, that if it is rational to adopt an intention to drink, and nothing has changed when the time for drinking comes, then it is rational not to reconsider the intention.

In the sense required for the argument to be valid, to 'adopt' an intention to A is to perform certain actions (be they deliberations or not) which lead one to intend to A. Even if it is rational of you to have adopted the intention to drink the toxin (via some non-deliberative strategy), and even if the circumstances develop as you anticipated they would (as we have assumed they do), then it does not follow, Bratman claims, that you should not reconsider your intention:
Even if it is rational of you on Tuesday to acquire and to have the intention to drink the toxin on Wednesday, when Wednesday arrives you will have ample opportunity to reconsider and to deliberate about whether to drink. It will be obvious to you then that you have nothing to gain and much to lose from drinking the toxin. And it is does not seem very plausible that reasonable habits of reconsideration — habits whose expected long-term impact on the agent’s interest in getting what she wants exceeds an appropriate threshold — would inhibit reconsideration in such cases. (106)

Bratman’s strategy here is this. Bratman wants to argue that (2’) is false: that it is possible you acquire on Tuesday the intention to drink the toxin on Wednesday, on Tuesday this intention is rational of you to have, you find the relevant conditions on Wednesday to be just as on Tuesday you expected them to be, and yet it is not rational of you on Wednesday not to reconsider that intention then. To do this, he introduces (and defends) a particular theory concerning the rationality of reconsideration, and then applies the theory to the case at hand.25

The theory is, in effect, a rule-consequentialist theory, and has two parts. First, it is rational for you not to reconsider your intention to drink if reasonable habits of (non) reconsideration would inhibit reconsideration in this case. Thus, the rationality of actions of reconsideration is to be determined by the rationality of habits of reconsideration. Second, such a habit is reasonable if and only if it is one whose ‘expected long-term impact on the agent’s interest in getting what she wants exceeds an appropriate threshold’. Thus, reasonable habits are, roughly, those which ‘satisfice’ the agent’s expected-value, amongst those habits of reconsideration which are options for the agent. In effect, then, Bratman’s view is that it is rational for you not to reconsider your intention to drink in this case if and only if value-satisficing habits of reconsideration would inhibit reconsideration of your intention in such cases.

With this theory in hand, Bratman can show why (2’) is false in this case, even though it seems quite plausible. On the one hand, (2’) seems quite plausible, since in the vast majority of situations, it is value-satisficing to have the habit that if you acquire on Tuesday the

25 It is, actually, a theory of the rationality of what Bratman calls non-reflective (non) reconsideration. Bratman has more to say about this theory in ch. 5 of Intentions, Plans, and Practical Reasoning. I introduce this theory in more detail in Chapter Nine.
intention to A on Wednesday, and if on Tuesday this intention is rational of you to have, and if you find the conditions on Wednesday to be just as on Tuesday you expected them to be, then you do not on Wednesday reconsider whether to A. In short, in the vast majority of situations, it is value-satisficing to have the habit that you do not reconsider decisions when things have not changed from the time you made the decision, because it will avoid unnecessary decision costs. On the other hand, though, (2') is false in the sorts of cases we are considering, since the habit of reconsidering, after midnight, your intention to drink the toxin (even if nothing has changed) is clearly the one it is value-satisficing to have. Such a habit of reconsideration would come into play after you adopted the intention — and received the million — to prevent you from actually drinking the toxin — and receiving a day’s severe illness. It is not rational of you on Wednesday not to reconsider your intention to drink the toxin. Principle (2') is false, and Bratman is not committed to the claim it is rational to drink the toxin.

[3.2] What implications, if any, does Bratman’s response have for my suggestion that your intention gives you a reason to drink the toxin? It implies the following objection: ‘You claim that in some Toxin Puzzle Cases the fact you intend to drink the toxin is sufficient reason for you to drink the toxin. But this is false. Bratman has shown that in all Toxin Puzzle Cases, if you adopt the intention, then you later rationally ought to reconsider the intention. And, presumably, in all Toxin Cases, if you were to reconsider the intention then you would change your mind and no longer intend to drink the toxin. However, an intention to A which you ought to reconsider and which you would lose if you did reconsider is surely one that provides no reason at all to A. Hence, in all Toxin Puzzle Cases, the fact you intend to drink the toxin is no reason at all to drink.’

How do I respond?26 I shall concede that if you were to reconsider your intention to drink the toxin, then you would change your mind.

26 Note that in responding to this objection I have teased from Bratman’s discussion, I am more concerned with the threat such an objection poses to my own position, than with the questions of whether Bratman himself would put the objection. There are two points, in particular, where the principles he mentions may not be directly relevant to my discussion. First, Bratman formulates his principles in terms of the locution ‘it is rational for x to reconsider whether to A’, rather than, as I am inclined, in terms of ‘x rationally ought (or is rationally permitted) to reconsider
Reconsidering an intention to A is, after all, coming to deliberate about whether or not to A, and, as I have already admitted, there are no facts you could use in deliberation to come to the conclusion you ought to drink the toxin. Further, I shall concede that an intention to A which you ought to reconsider and which you would lose if you did reconsider is one that provides no reason at all to A. In the absence of a general theory of reasons for action, I do not see what could count as an argument for this claim, but it nevertheless seems correct. I concede, then, the last two premises of my objector’s argument.

I deny, though, that in all Toxin Cases, if you adopt the intention, then you later rationally ought to reconsider the intention. Bratman claims – and I agree – that you rationally ought not to reconsider your intention to drink if reasonable habits of (non-) reconsideration would inhibit reconsideration in your case; and, that a habit is a reasonable if and only if it is value-satisficing. (Actually, as a first approximation, I would claim that a habit is reasonable if and only if it is expected-value-maximising, but I shall let this point pass here, and come back to it in Chapter Nine.) Bratman, however, also claims that in all Toxin Puzzle Cases the habit ‘if you initially intend to drink the toxin, then reconsider this intention after midnight’ is value-satisficing, and so one you rationally ought to have.

But this is false. In the First and Second Counterexamples, it is value-satisficing to have the different habit that ‘if you initially intend to drink the toxin, then do not later reconsider this intention’. To see this, consider each Counterexample.

(1) In the First Counterexample, you cannot but be disposed to not reconsidering this intention. Recall that, in this example, you cannot but be disposed such that if you intend to drink the toxin, then you drink the toxin. We have assumed (two paragraphs above), though, that were you to reconsider your intention you would change your mind, and not drink the toxin. Hence, were you to drink the toxin,
then you would not have reconsidered. Hence, presumably, you also cannot but be disposed such that if you intend to drink the toxin, then you do not reconsider. Since, then, there are no alternatives to the disposition not to reconsider the intention in this example, it follows trivially that the expected-value to you of this disposition is greater than that of all its alternatives. It follows trivially, in other words, that it maximises expected-value for you to have this disposition.

(2) In the Second Counterexample, it follows non-trivially that it maximises expected-value for you to have this disposition. Recall that, in this example, the second eccentric billionaire is concerned with your dispositions, and it turns out that in order for you to get the ten million, you must adopt the disposition not to reconsider your intention to drink the toxin.

In the First and Second Counterexamples, and given some conditions, the fact you intend to drink the toxin is reason enough for you to drink. My argument for this claim has been that, in these situations, the deontic principles \((OC_n)\) and \((OP_n)\) entail that there is some non-consequentialist reason for you to drink the toxin, and I simply make the further speculation that this reason is the fact you intend to drink the toxin. This speculation, however, receives independent (if reluctant) support from Bratman’s theory of rational reconsideration. It turns out that, in the First and Second Counterexamples, you rationally ought not reconsider the intention to drink the toxin, and, this being the case, it seems clear you are rationally permitted to act on this intention, and drink the toxin.

Conclusion

The Self-Interest Theorist might claim the two counterexamples I provide to their theory are, in effect, incoherent, but if they do they are mistaken. To be sure, Farrell and Bratman identify defective arguments for the view that it is rational to drink the toxin, but these are not my arguments. We end the first part of the thesis, then, with the conclusions of the previous chapter standing. First, that the Self-Interest Theory’s claim that only consequentialist considerations provide reasons for action is mistaken. Second, 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, you are then rationally permitted to
perform that action, EVEN IF one is free to do otherwise and it has the best outcome for one to do otherwise (though this action may very well be irrational absent this condition). There are indeed more reasons for doing things than are dreamt of in the consequentialist's philosophy, and thus there is a space into which, if we are lucky, we may be able to insert moral reasons.
RATIONAL COOPERATION

PART II
Chapter Four

David Gauthier on Rational Cooperation

The most troubling form of moral deflationism, I claimed in the Introduction, is one granting that moral utterances have truth values, and even granting that some of them are true, but denying that they make any difference. Why should I be moral? - or so asks this most troubling of moral deflationists. There are (at least) two ways of answering this challenge. On the one hand, one may simply repudiate the demand that morality answer to each and every person's reason, and insist that this an inappropriate way of attempting to justify morality. On the other hand, one may accept the deflationist's demand, and try to argue for the possible rationality of morality within the bounds of each person's reason. David Gauthier, in his *Morals by Agreement*, attempts to take moral deflationism seriously, and to show why, under certain conditions, each person's reason does indeed counsel the way of morality. His argument has two basic parts. First is a contractarian analysis of morality: one morally ought to perform some action when it is what one would agree to do were one to employ a rational bargaining procedure, from a rational initial bargaining position, in a situation of perfect information. Second is a rationalistic justification for cooperation: when others are sufficiently cooperatively disposed and sufficiently knowledgeable about how each is disposed to behave, it is rational to do what one would in this manner rationally agree to do. In this chapter I will concentrate on the second part of Gauthier's project. We shall see that his argument is not without problems.

§1 The Self-Interest Theory, and rational cooperation

It seems clear that irrational persons may do things which fail to promote their interests, or what they value, and - worse still - things actually contrary to their interests; and equally as clear that rational
persons may expect to do the best they could to promote their interests, or what they value. It seems, then, that a group of rational persons may expect to do better than a group of irrational persons in promoting their interests, or what they value. Yet this is not always the case. There are some situations in which rational persons can expect to do very poorly indeed, and certainly very much worse than their irrational cousins.

[1] You find yourself in Hobbes’s state-of-nature. There is a group of you, each armed to the hilt and suspicious of the others. None of you draws a distinction between what you morally may or may not do, and, in particular, each has the right, capability, and also no compunction against, using force upon others, if each perceives that this will benefit the pursuit of what they value most. There are clear advantages to be had by the possession of such a right, whether in the exploitation of others or in the defence against just such exploitation. You face a choice whether or not to relinquish your right to use force against the others. Should you do so or not?

This is not the world’s most difficult decision problem. Pick any other person in the state-of-nature. Me, for instance. Either I will relinquish my right to use force against you, or I will not. On the one hand, if I will not relinquish my right, then the outcome of your also not doing so (namely, a nervous and somewhat tenuous balance of terror) would be much better for you than the outcome (namely, my ability to exploit you with threats of force) of your unilaterally relinquishing your right. On the other hand, if I will relinquish my right, then the outcome of your not doing so (namely, your ability to exploit me with threats of force) would again be much better for you than the outcome (namely, a calmer peace) of relinquishing your own right. Whatever I will do, then, it is obvious you rationally ought not to relinquish your right. And this is so even though the outcome of joint non-relinquishment of the right to use force (namely, a tenuous balance of terror) would be much worse for each of us than the outcome (namely, a somewhat calmer peace) of joint relinquishment of this right.

---

1 I shall assume – solely for the sake of illustration – that Hobbes’s state-of-nature is to be interpreted as a Prisoner’s Dilemma. This is, by now, a common interpretation. See D. Gauthier, The Logic of the Leviathan, (Oxford: Clarendon Pr., 1969), pp. 14 ff., and pp. 76 ff.
We can justify all of these statements on the basis of the following two claims:

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

(P1) Each of us faces an independent choice between the actions of relinquishing our right, or not; whatever the other person does, each values the outcome of not relinquishing to that of relinquishing; however, each of us values joint relinquishment of the right to joint non-relinquishment.

The first assumption, (S), is, of course, what I have been calling the standard formulation of the Self-Interest Theory. The second, (P1), is a description of the central features of the situation I have just described. Since it will be important later, it needs some explaining.

[2] Assumption (P1) defines what is in modern parlance is referred to as the Prisoner's Dilemma. The three clauses of the assumption may be represented diagrammatically.

---

2 The name comes from the original example used to illustrate situations described by assumption (P1). For a detailed introduction to this example, and to the Prisoner's Dilemma more generally, see R. D. Luce & H. Raiffa, Games and Decisions, (New York: Wiley, 1957), pp. 94-102, and R. Campbell, 'Background for the Uninitiated,' in R. Campbell & L. Sowden, Paradoxes of Rationality and Cooperation, (Vancouver: Univ. Brit. Columbia Pr., 1985): 3-41.

3 The three clauses are explicitly identified, amongst others, by G. Harman, 'Rationality in Agreement: A Commentary on Gauthier' "Morals by Agreement," Soc Phil Pol 5 (1988): 1-16. Since the sort of situation I will be examining is, in a number of ways, simplified, I will say a few words in justification.


(b) There is clearly only one outcome it would be rational for us to agree on: mutual cooperation. Typically, though, cooperating involves (at least) two separate problems: firstly, coming to some agreement concerning a joint course of
The first clause says that each of us faces an independent choice between the actions of relinquishing our own right or not. It will save words if, henceforth, we say that you cooperate (and denote this by ‘C’) if you relinquish this right to use force against me; and defect (‘D’), if you do not. In short, the first clause says that each agent faces an independent choice between cooperating and defecting. Each pair of actions we may perform has an outcome, and each of us assigns values to these outcomes. Since both of us face a choice between two actions, there are four (= 2 x 2) possible outcomes. These outcomes, and the values each of us attaches to them, may be depicted as follows:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>You</td>
<td>Civil</td>
<td>You exploit</td>
</tr>
<tr>
<td></td>
<td>Society</td>
<td>Me</td>
</tr>
<tr>
<td></td>
<td>c,c</td>
<td>s,t</td>
</tr>
<tr>
<td>Me</td>
<td>I exploit</td>
<td>War of all</td>
</tr>
<tr>
<td></td>
<td>You</td>
<td>against all</td>
</tr>
<tr>
<td></td>
<td>t,s</td>
<td>d,d</td>
</tr>
</tbody>
</table>


(c) The outcome it would be rational for us to agree on—mutual cooperation—is very unstable: whatever the other person does, it is better for each not to cooperate. Typically, though, the stability of an agreement is thrown into doubt by much less truculent factors (See M. Taylor, *Possibility of Cooperation*, (New York: Cambridge Univ. Pr.), Ch. 2, and A. K. Sen, ‘Choice, Orderings and Morality,’ in S. Korner (ed.), *Practical Reason*, (New Haven: Yale Univ. Pr., 1974): 54-66). In response it need only to be noted that if it can be shown cooperating is rational, even in conditions most hostile to the emergence of cooperation, then we may assume it can be rational under these less hostile conditions.

(d) Hobbes’s state-of-nature is a particularly violent form of conflict, in which the concerns of each are simply with survival. Typically, though, the problems of cooperation are not as violent, and do not involve such a narrow conception of self-interest. In response, it is important to note that conflicts of the sort described in the Prisoner’s Dilemma arise, strictly speaking, from the presence of what I have called agent-relative value, and not just from the presence of narrower forms of self-interest. As such, the type of situations addressed by the text is broader than just Hobbes’s admittedly violent state-of-nature, which I use solely for the purposes of illustration. (D. Parfit, *Reasons and Persons*, (Oxford: Clarendon Pr., 1984), pp. 95 ff., sec. 36; F. C. T. Moore, ‘The Martyr’s Dilemma,’ *Analysis* 45 (1984): 29-33.)
The values I assign to possible outcomes are listed first, and are: my exploiting you (=t, Temptation payoff); mutual cooperation and the possibility of civil society (=c, Cooperation payoff); mutual defection and the war of all against all (=d, Defection payoff); and, your exploiting me (=s, Sucker payoff). Clearly, \( t > c > d > s \), since I value most the outcome of my exploiting you (t), next mutual cooperation and the possibility of civil society (c), third the war of all against all (d), and, worst of all, your exploiting me (s).

This first clause means there can be no literally binding agreement between us.\(^4\) Neither of us can perform any actions which will render us literally incapable of defecting. On the one hand, neither can force the other to cooperate, for the difference between the one of us and the other is not so considerable that the one can claim a benefit from the use of force over the other, that the other cannot also claim against the one. On the other, neither can force themselves to cooperate, for, unlike Odysseus neither has the rope and strong sailors to physically restrain themselves from not cooperating,\(^5\) and cannot literally bind themselves to the cooperative action.

The second clause says that whatever the other person does, each values the outcome of not relinquishing to that of relinquishing their right. It will save words if, henceforth, we say that action A dominates action B for me if, whatever anyone else does, I value the outcome of doing A to that of doing B. Hence, the second clause says that for each agent, defection dominates cooperation. In terms of the diagram, if you were to cooperate, then defecting would get me an outcome (namely, my being able to exploit you with threats of force) I value most at t, which is greater than the value, c, of the outcome (namely, a calmer peace and the possibility of civil society) I would get were I to cooperate. On the other hand, if you were to defect, then defecting would get me an outcome (namely, a somewhat tenuous balance of terror) I value at

---


\(^5\) See Homer, The Odyssey, 12.154-200
d, which is greater than the value, s, of the outcome (namely, your being able to exploit me) I would get were I to cooperate. Either way, the outcome is better for me if I defect, and the same applies, mutatis mutandis, to you.

This second clause means that there can be no binding agreement (in a different sense) between us. What is implied in particular is that there are no actions available to us which could set up incentives such that each player ends up valuing the outcome of their own cooperation to that of their own defection, whatever the other person does. In particular, it entails that coercive enforcement of cooperation, benevolence towards the other person, and mutual similarity of goals are either non-existent or limited to an extent consistent with still valuing the outcome of defection to that of cooperation, whatever others do.

The final clause of (P1) says that each of us, however, values joint relinquishment of the right to joint non-relinquishment. It will save words if, henceforth, we say that an outcome O₁ is optimal when there is no other outcome which everyone values at least as much as O₁, and someone values more than O₁. Hence, the final clause says that the outcome of joint defection is not optimal. In terms of the diagram, if both of us cooperate, then there would be an outcome – namely, the possibility civil society – which each of us values second-best, at c; if both of us defect, then there would be an outcome – namely, the war of all against all – which each of us values only third-best, at d. The possession of the right to use force in the state-of-nature leads, as Hobbes believed, and as I shall assume, to a war of all against all. There is the universal realisation that the right to use force might advantageously be employed in pre-emptive attacks upon others, and in this way is there universal uncertainty about whether or not one will be the object of just such an attack. Such a situation is a state of war, a state which ‘consisteth not in battle only, or the act of fighting; but in a tract of time, wherein the will to contend to battle is sufficiently known’. For Hobbes, it is a state in which peace of mind is hardly possible, where the conveniences and securities of civil society surely lacking, and in which life is ‘solitary, poor, nasty, brutish, and short’.6

In Hobbes's state-of-nature, a group of irrational agents—who cast off their right to use force against the others, and obtain as a result the possibility of civil society—would do much better than a group of rational agents—who retain their right, and with it the war of all against all. Many find this disturbing. Hobbes himself was certainly disturbed by this conclusion, and his solution to the problem involved our mutual relinquishment of the right to use force, and the setting up of a sovereign to enforce it. But is Hobbes, or anyone else for that matter, correct to suppose that the state of nature, as disturbing as it is, contains within it a justification for the rejection of the Self-Interest Theorist's conception of rationality?

§2 The Prisoner's Dilemma, and David Gauthier

Notoriously, Hobbes has problems explaining why persons initially disposed to act solely to produce outcomes best for themselves could, or would, come to relinquish their right to use force. For everyone to agree to do so is, of course, not to solve the problem, since, whatever others do, any particular person will do better by reneging on such an agreement. However rational it might be for me to agree to cooperate, and cast off my right to use force against you, it remains irrational for me actually to do so. We are left with the residual problem of explaining why anyone would have reason to keep such an agreement. In this section I will introduce David Gauthier's attempt to show how it might be rational to keep just such agreements.

[1] Gauthier tries to show that, under certain conditions, it is rational to make such an agreement, and, if it is, then it is rational to keep it. To be able do this, Gauthier changes the focus of discussion, and considers the question of the type of agent it is best to be. He distinguishes, in particular, between two types of agents: what he calls straightforward, and constrained, maximisers.

---


The first, straightforward, maximisers are just those agents who are disposed to do what the Self-Interest Theory tells them to do. They are such that if the expected value to them of the outcome of doing A exceeds that of doing any alternative action, they do A. In particular, they keep a promise, or honour an agreement, if and only if it maximises expected-value for them to keep the promise, or honour the agreement.

The second, constrained, maximisers are disposed, in certain situations, to keep agreements for their own sakes. Says Gauthier:

We shall therefore identify a constrained maximiser thus: (i) someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies, and approach what she would expect from the co-operative outcome determined by minimax relative concession; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies.

This is quite a general (and somewhat complex) definition, more so than I need for my discussion. It is amenable to useful simplification in the sorts of situations we are considering.

Take the first clause, which is concerned with the type of joint practices a constrained maximiser is conditionally prepared to adopt. These are ones (to take the second condition mentioned, first) where the utility she expects will "approach what she would expect from the co-operative outcome determined by minimax relative concession". Minimax relative concession is Gauthier's preferred rational bargaining procedure, but we can replace any reference to it in the simple sorts of situations we are considering, since the advice it gives (and the advice any adequate bargaining procedure would give) is that we both cooperate, and cast off our right. These practices are, further,

---

9 See Gauthier, RM, pp. 428 ff, and MA, pp. 167 f.
10 Gauthier, MA, p. 167. Note that Gauthier uses the term 'utility' in the way I use the term 'expected-value'.
12 This is not, strictly speaking, correct. Suppose $t+s>2d$, so that each of us values a half chance of being exploited and half chance of exploiting to the certainty of mutual defection. If, further, joint mixed strategies are introduced (Gauthier, MA,
ones such that "the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies." Since the outcome to be expected were everyone to employ individual strategies (that is, were everyone to follow the dictates of the Self-Interest Theory) is the war of all against all, it is plain that the joint strategy of mutual cooperation is (to say the least) no worse than this. In the sorts of situations we are considering, then, a constrained maximiser is conditionally disposed to cooperate.

Take now the second clause of the definition, which is concerned with the conditions under which a constrained maximiser will actually base her action on the rationally acceptable joint strategy. She acts in this way "should her expected utility be greater than what she would expect were everyone to employ individual strategies". What, exactly, does this mean? Gauthier makes this clear later:

Her disposition to co-operate is conditional on her expectation that she will benefit in comparison with the utility she could expect were no one to co-operate. Thus she must estimate the likelihood that others involved in the prospective practice or interaction will act co-operatively, and calculate, not the utility she would expect were all to co-operate, but the utility she would expect if she co-operates, given her estimate of the degree to which others will co-operate. Only if this exceeds what she would expect from universal non-co-operation, does her conditional disposition to constraint actually manifest itself in a decision to base her actions on the co-operative joint strategy.13

It is clear Gauthier intends the cooperation of a constrained maximiser to be dependant upon her expectations regarding whether or not the others will cooperate. I make an estimate of the likelihood you will do your part of the joint strategy — casting off your right to use force — and then calculate the expected-value to me of my doing similarly. If this exceeds the expected-value to me of our war of one against other (not the expected-value to me of my doing otherwise), will I do similarly. Furthermore, when a constrained maximiser meets uncooperative persons, she 'does not play into their hands by basing

---

13 Gauthier, MA, p.169, emphasis added.
her actions on the joint strategy she would like everyone to accept, but rather, to avoid being exploited, she behaves as a straightforward maximizer. In the simple sort of situations we are considering, then, Gauthier's definition amounts to the following:

(CM) An agent is a constrained maximiser ('CM') if and only if: the agent cooperates if and only if the expected value to the agent of doing so exceeds that of mutual non-cooperation. The expected value of an action is calculated on an estimation of the probability that the other will cooperate.

An important implication of this definition, which we will need later, is that a constrained maximiser will cooperate if they believe the other person will cooperate, and will not if they believe the other will not.

[2] Gauthier then argues that, given certain conditions, one rationally ought to be a constrained, rather than a straightforward, maximiser. He imagines a situation, occurring before the issue of cooperation or not has arisen, in which an agent is to make a choice between becoming a straightforward, or a constrained, maximiser. He denotes by $u$ the expected utility the agent could expect were each person to act on the basis of an individual strategy - in the case we are considering, this is the value, $d$, the agent assigns to the war of all against all. He denotes by $u'$ the expected utility should all act on the cooperative joint strategy - in our case, the value, $c$, the agent assigns to civil society. Clearly, $u'$ exceeds $u$. Gauthier then has the agent argue as follows. (For simplicity I will replace Gauthier's $u$ by my $d$, and his $u'$ by my $c$).

Suppose I adopt straightforward maximization. Then I must expect the others to employ maximizing individual strategies in interacting with me; so do I, and expect a utility, $[d]$.

Suppose that I adopt constrained maximization. Then if the others are conditionally disposed to constrained maximization, I may expect them to base their actions on a cooperative joint strategy in interacting with me; so do I, and expect a utility $[c]$. If they are not so disposed, I employ a maximizing strategy and expect $[d]$ as before. If the probability that others are disposed to constrained maximization is $p$, then my overall expected utility is $[pc + (1 -pd)]$.

---

14 Gauthier, MA, p. 169
Since $c$ is greater than $d$, $[pc + (1 - p)d]$ is greater than $d$ for any value of $p$ other than 0 (and for $p = 0$, the two are equal). Therefore, to maximize my overall expectation of utility, I should adopt constrained maximization.\(^{15}\)

Thus we see that Hobbes was correct to insist that rational persons in the state of nature could come to reason differently, and not act solely to produce outcomes the best for themselves. They would cease to reason as straightforward maximisers, and come instead to reason like constrained maximisers, cooperating when they have an expectation that others, too, will cooperate.

This particular argument, though, depends on a few quite strong assumptions, unstated but nevertheless very important. The first unstated assumption is that the agents, in choosing how they are to be disposed, face a choice only between constrained, and straightforward, maximisation, or that all other choices are inferior. What the argument shows though, if anything, is that under certain conditions it is better to be a constrained maximiser than a straightforward one. It follows from this that you \textit{ought} to be a constrained maximiser only if these are the only choices or only if all other choices are clearly inferior. The second unstated assumption of this argument is that both the constrained and the straightforward maximiser will appear in their true colours.\(^{16}\)

On the supposition I have adopted straightforward maximisation, I must expect others not to cooperate only if I expect others to \textit{know} I have adopted straightforward maximisation. And, on the supposition I have adopted constrained maximisation, I must expect other constrained maximisers to cooperate only if I expect them to \textit{know} I have adopted constrained maximisation. Thus, it needs to be assumed that, after having adopted a disposition, the other will be able to tell what disposition I have adopted. It needs to be assumed, in other words, that dispositions are transparent.

The fact Gauthier's argument depends on such strong assumptions might (and has) provoked objections to the relevance of his argument.

\(^{15}\) Gauthier, MA, p. 172

\(^{16}\) Gauthier presents the argument in the text, and then himself points out that it depends on this second assumption, which he calls the assumption of \textit{transparency} (MA, p. 173). In the following pages, he introduces a weaker form of the assumption, that of \textit{translucency}, and proceeds to a second argument for the claim that, under certain conditions, it is rational to be a constrained maximizer. Since, in this thesis, I will not be concerned with this more complex second argument, or with the objection making it necessary, I will not introduce it.
After all, the dispositions amongst which one can choose presumably include others than just the two Gauthier considers,¹⁷ and such dispositions will, in any case, be nowhere near as transparent as his argument requires.¹⁸ Important as they are, I will not in this thesis pursue these points. In the first place, I believe (but will not in this thesis argue) that even when one does consider more realistic types of situations – in which there are more dispositions than just the two Gauthier introduces, and when these dispositions are not transparent – then it will still remain the case that some non-expected-value maximising disposition is rational (though, I should point out, it will not be that of constrained maximisation).¹⁹ In the second place, and more importantly, even when we do restrict our attention to the sort of unrealistic case required for this argument to work, its conclusion is still sufficiently threatening to some. In particular, it is threatening to those who would defend the Self-Interest Theory, who would want, even in the unrealistic case we are considering, to deny it is rational to cooperate.


¹⁹ I have in mind, here, the discussion of the so-called ‘iterated’ Prisoner’s Dilemma. In an environment of many different types of agent who face the iterated Prisoner’s Dilemma, it may very well be rational to adopt a disposition – so-called TIT-FOR-TAT – which reacts to previous cooperation of others with cooperation, and previous defection with defection, whether or not it is expected-value-maximising to do so. I shall not here argue that this is so, but see R. Axelrod, The Evolution of Cooperation, (New York: Basic Books, 1984) for details of the iterated Prisoner’s Dilemma.
Such is Gauthier’s attempt to argue it might be rational to cooperate, and keep agreements, even if one is free to do otherwise, and it has the best outcome to do otherwise. The strategy is an indirect one: first to argue for the rationality of cooperative dispositions, and then to move from there to the rationality of cooperative actions. There are thus two general counter-strategies available for those who would wish to deny his conclusion: first, one may deny he has shown it is rational to be disposed to cooperation; and second, one may claim that, even if he had shown this, he would not have shown it was rational actually to cooperate. I shall discuss these objections in turn.

§3 Rational Cooperative Dispositions?

There has been much discussion of Gauthier’s argument and some have not been totally convinced. Some objections centre of the first half of Gauthier’s strategy: his argument for the rationality of being disposed to cooperating. We shall see it is an argument with a number of problems.

[1] The first problem concerns Gauthier’s assumption that there is a fixed probability, p, to be assigned to the other agent’s being a constrained maximiser. In order to understand the objection, we need to introduce the distinction between parametric and strategic choice - a distinction Gauthier himself endorses. He says parametric choice occurs when ‘the actor takes his behaviour to be the sole variable in a fixed environment. In parametric choice the actor regards himself as the sole centre of action.’ Contrasted with this is strategic choice, in which ‘the actor takes his behaviour to be but one variable amongst others, so that his choice must be responsive to his

---

20 There are a number of books and journal issues devoted to a discussion of Gauthier's views, as well as numerous reviews of Morals by Agreement. These include: Ethics 97 (1988); Can J Phil 18 (1987); Soc Phil Pol 5 (1988); P. Vallentyne (ed.), Contractarianism and Rational Choice: Essays on Gauthier's Morals by Agreement, (New York: Cambridge Univ. Pr., 1991).

expectations of others' choices, while their choices are similarly responsive to their expectations.\footnote{22}

The problem is then as follows. The situation Gauthier is concerned to address is one in which there are a number of agents interacting, in the knowledge they are doing so. In the first instance, they face a choice about whether to cooperate or not. Gauthier understands\textit{ this} choice as one in which each agent takes his behaviour to be but one variable amongst others, so that his choice must be responsive to his expectations of others' choices. He understands the choice of cooperation, or not, to be a strategic one. If this is the only choice then, as we have seen, it seems that mutual non-cooperation and the war of all against all is the rational outcome. To handle this problem, Gauthier introduces dispositions, and he asks: which is the best disposition to choose? In the second instance, then, the agents in question face the question about whether or not to choose to be disposed to cooperating. Gauthier understands\textit{ this} issue, strangely, as one in which each agent regards himself as the sole centre of action. He understands the choice of a disposition of cooperation to be a parametric one. The problem, then, is simply this: it is inconsistent to suppose (as Gauthier does) that, when agents are choosing to cooperate, they should assume the choices of others are not fixed, but that, when they are choosing dispositions, they should assume the choices of dispositions of others are fixed.

Certain passages suggest that Gauthier may be able to respond to this problem. For the role that a\textit{ choice} of disposition plays in his argument is not, it turns out, a central one:

the idea of a choice among dispositions to choose is a heuristic device to express the underlying requirement, that a rational disposition to choose be utility maximizing. In parametric contexts, the disposition to make straightforwardly maximizing choices is uncontroversially utility-maximizing. We may therefore employ the device of a parametric choice among dispositions to choose to show that in strategic contexts, the disposition to make constrained choices, rather than straightforwardly maximizing choices, is utility-maximizing.\footnote{23}

\footnote{22 Both quotes are from Gauthier, MA, p. 21. For a more detailed explanation of the distinction, he refers the reader to J. Elster, \textit{Ulysses and the Sirens: Studies in Rationality and Irrationality}, (Cambridge: Cambridge Univ. Pr., 1979), pp. 18-9, 117-23.}

\footnote{23 Gauthier, MA, p. 183}
There is a distinction between the rationality of *choosing* to be a CM, and that of *being* a CM, and it is clear from this passage that Gauthier places most importance on the second. Gauthier does not explicitly address the objection we are now considering, but one may speculate that he could use this passage to claim that, since the idea of a choice of disposition is in any case only a heuristic one, he is not committed to any problems there may be with it.

If this is the way Gauthier would reply to the objection, then it is not clear he would have said enough to dispel the problem. For even though Gauthier's position does not, strictly speaking, commit him to a view about how *choosing dispositions* is to be understood, whether parametrically or strategically, it turns out that neither option would be congenial. He faces a dilemma: choice of disposition is to be understood as a parametric, or a strategic, choice. On the one hand, if it is to be understood as a parametric choice, then Gauthier encounters the objection with which we started: it seemed inconsistent to suppose that, when agents are choosing to *cooperate*, they should assume the choices of others are not fixed, but that when they are choosing *dispositions*, they should assume the choices (of dispositions) of others are fixed. These choices are either both strategic or both parametric. On the other hand, if the choice of disposition is to be understood as a strategic one, then Gauthier encounters a very similar objection: it seems inconsistent to suppose that, when agents are *choosing* dispositions, they should assume the choices of dispositions of others are not fixed, but when they are considering the rationality of *having* dispositions, they should assume the dispositions that others have are fixed. Choosing dispositions and having dispositions are either both parametric or both strategic.

[2] The second problem with Gauthier's argument for the rationality of constrained maximisation centres on the following crucial part of the above passage: "Suppose I adopt constrained maximization. Then if the others are conditionally disposed to constrained maximization, I may expect them to base their actions on a co-operative joint strategy in interacting with me." 24

Even if we assume transparency, this is not quite right. Suppose both of us adopt constrained maximisation, and are both transparent.

24 Gauthier, MA, p. 172.
Since I am transparent, you come to believe that I am a constrained maximiser. You come to believe I will cooperate if and only if I expect that you are sufficiently likely to cooperate. Similarly, since you are transparent, I come to believe that you are a constrained maximiser, and thus come to believe you will cooperate if and only if you expect that I am sufficiently likely to cooperate. How does it follow from this that each of us may expect that the other is sufficiently likely to cooperate? In fact, this does not follow.\textsuperscript{25} From the mutually known facts that (a) I will cooperate if and only if I expect that you will, and that (b) you will cooperate if and only if you expect that I will, it does not follow that either of us will, in fact, come to expect the other to cooperate, and so does not follow that we both will, in fact, cooperate. It is as if each of us is waiting for the other to make the first move, but, simply because we are waiting, we may each fail to make any move at all. How might we solve this problem?

[2.1] The simplest solution is the following. If CM is defined in such a way that CMs are disposed to act on the basis of an estimation of the likelihood the other \textit{will} cooperate, then it does not follow two CMs would cooperate. Since the agent's dispositions are assumed to be transparent, it would be better to define a CM in such a way that they are disposed to act on the basis of an estimation of the likelihood the other \textit{is a constrained maximiser}. In particular, it would be simplest to define a constrained maximiser thus:

\begin{enumerate}
\item[(CM2)] A person is a constrained maximiser if and only if: they cooperate if and only if they believe the other person \textit{is a constrained maximiser}.
\end{enumerate}

If this is our definition, then the required inferences will indeed be valid. On the one hand, a CM will cooperate with a CM. When a CM faces another CM, she comes to believe the other is a CM (since they are transparent), and since she herself is a CM (and so cooperates with

all those she believes are CMs), then she will cooperate. On the other hand, a CM will not cooperate with an SM. When a CM faces an SM, she comes to believe the other is an SM (since they are transparent), and since she is a CM (and so cooperates only with those she believe are CMs), then she will not cooperate.

The problem with this solution, though, is that it renders the notion of constrained maximisation circular. Of course, circularity is not necessarily a problem. It may, for example, be a conceptual truth that something is red if and only if it appears red to standard observers under standard conditions. This is not a problem - even if circular - since most sighted people have a grasp of the concept independent of this truth. Sometimes, though, circularity is a problem. In particular, (CM2) claims that someone is a constrained maximiser if, and only if, they cooperate with all and only those they believe are also constrained maximisers. This is a problem - because circular - since no-one has a grasp of this concept independent of this statement. We need to be provided with an independent handle on constrained maximisation, which this circular redefinition does not provide.

[2.2] Richmond Campbell argues that Gauthier does have a way out of the problem we have discovered, and proposes his own redefinition of constrained maximisation. It goes like this:

(CM3) A person is a constrained maximiser if and only if: (i) they have property R, and (ii) they cooperate if and only if they believe that the other person has property R.

The task for this definition is to find a substitution instance for 'R' that will give the intuitively right answers.

First, one could take 'R' to be 'will cooperate.' If so, then a person is a constrained maximiser if and only if: (i) they will cooperate, and (ii) they cooperate if, and only if, they believe that the other person will cooperate. This form of the definition will not do. Its first clause is defective, since it is perfectly possible to be a CM without cooperating, when, for example, one is the sole CM in a population of transparent SMs. Sensibly, one does not cooperate, but this does not mean one

26 H. Smith, in 'Deriving Morality from Rationality,' p. 242 fn. 18, rejects a very similar suggestion of Danielson's for this reason.
would not be prepared to cooperate with those who are themselves more cooperative, and so this does not mean one is not a CM. Its second clause is also defective, since it suffers from the same problems as the original definition, (CM).

Second, one could take 'R' to be 'is a CM'. If so, then a person is a constrained maximiser if and only if: (i) they are a CM, and (ii) cooperate if, and only if, they believe that the other person is a CM. This form of the definition will also not do. It is, of course, equivalent to (CM2), and thus shares its circularity. What we need, then, is some property different from that of being a CM which nevertheless no SMs will have. If we can find such an 'R', then it will turn out that CMs will cooperate with one another, but not with SMs, just as the argument requires.

Third, one could adopt one of Campbell's own suggestions, and take 'R' to be the property of 'being ready to reciprocate cooperation when making the second move in sequential PDs'. A sequential Prisoner's Dilemma ('PD') has the same structure as the Prisoner's Dilemma we have been considering in this chapter, except that one agent knows what the other did, typically because the other agent acted first. If this suggestion is adopted, then a person is a constrained maximiser if and only if: (i) they are ready to reciprocate cooperation when making the second move in sequential PD, and (ii) they cooperate if and only if they believe that the other person is ready to reciprocate cooperation when making the second move in sequential PD. This form of the definition will still not do. The term 'ready' is troublesome. If it is as non-committal as it sounds, then we will run into the same problem that generated the need for a redefinition of constrained maximisation in the first place. For, surely, our original constrained maximisers – those who would cooperate if they expected the other to cooperate – were ready to cooperate, but as ready as they were, they kept each other waiting and did not necessarily end up cooperating.

Fourth, and finally, let us slightly modify Campbell's definition, and take 'R' to be 'would reciprocate when making the second move in sequential PD'. If so, then a person is a constrained maximiser if and only if: (i) they would reciprocate when making the second move in sequential PD, and (ii) they cooperate if and only if they believe that the

---

other person would reciprocate when making the second move in sequential PD. On this suggestion, a CM would thus use the fact that the other person would reciprocate in a sequential Prisoner’s Dilemma to ground his own cooperation.

This suggestion also has its difficulties. For what a person would do in some counterfactual situation is sometimes not a good indicator of what they will do in a similar, but actual, situation. Suppose, for example, you know I am a trustworthy, but untrusting, person. Since I am trustworthy, I myself will not attempt to exploit anyone, and would reciprocate when making the second choice in a sequential Prisoner’s Dilemma. I satisfy property R, and you know it. Since I am not trusting, though, I fear that you might try to exploit me, and will not cooperate in simultaneous Prisoner’s Dilemmas which are the ones we are actually concentrating on in this chapter. I will not cooperate, and you know it. This shows that Campbell’s definition is inadequate. Suppose you are a CM. It follows, on Campbell’s definition, that you will cooperate with anyone you believe has property R. You know (and so believe) that I have property R. Therefore, CM that you are, you cooperate with me in the simultaneous Prisoner’s Dilemma which is the focus of concern in this chapter. You cooperate, while knowing (and so believing) that I will not. The CM, as Gauthier meant to define her, would however not cooperate were she to believe, as you do, that I will not cooperate. Recall Gauthier says that when a CM has reason to suppose the other agents would not cooperate, then she “does not play into their hands by basing her actions on the joint strategy she would like everyone to accept, but rather, to avoid being exploited, she behaves as a straightforward maximizer”. This final substitution for ‘R’ is inadequate.

Gauthier, then, faces problems with his argument for the rationality of constrained maximisation. On the one hand, he supposes that the choice between the CM and SM dispositions is to be a parametric one; but it seems that, on pain of inconsistency, it should be understood as a strategic choice. On the other hand, the argument he provides is not valid, given his official definition of constrained maximisation; and, it seems, there need be no definition adequate to

29 See H. Smith, ‘Deriving Morality from Rationality,’ p. 242 fn. 18
30 Gauthier, MA, p. 169
the argument. Gauthier's indirect strategy seems to have fallen at the first hurdle.

§4 Rational Cooperative Actions?

Other objections centre of the second half of Gauthier's strategy: his move from the rationality of cooperative dispositions to the rationality of cooperative action. For even if we were to grant that Gauthier had shown each of us rationally ought to be disposed to cooperating, he still needs to conclude that it is rational actually to cooperate. We shall see it is a move open to a number of objections.

[1] It is a move Gauthier believes he is entitled to make. He states, in Morals by Agreement, that '[i]mperfect actors find it rational to dispose themselves to make less than rational choices. No lesson can be drawn from this about the dispositions and choices of the perfect actor. If her dispositions to choose are rational, then surely her choices are also rational.' Gauthier, it seems, endorses a principle like the following:

(B2) If you rationally ought to adopt the disposition that if you believe that $p$ then you perform $A$, and if you do believe that $p$, and if nothing relevant to the adoption of the disposition has changed save what must be changed with the coming about of $p$, then you are rationally permitted to perform action $A$.

This principle is prominent in practical reasoning, and it is surely plausible. If Gauthier's argument for the rationality of constrained maximisation is correct, then you rationally ought to adopt the disposition that if you believe the other person will cooperate then you cooperate as well. If, as seems likely, the other person becomes similarly disposed, then you will come to believe they will cooperate, and that (presumably) nothing relevant to the adoption of the disposition has changed. In this case you are rationally permitted to cooperate, even if you are free not to, and it has the best outcome for you not to.

Alas, like other bridging principles, there are more than enough people prepared to deny the validity of the principle, (B2), Gauthier needs. One nay-sayer is Gregory Kavka himself:

[Gauthier says CM] is more rational that the disposition to maximize expected utility, because those who possess the former disposition will have more opportunities for mutually beneficial interaction. This may well be true, but it hardly follows, as Gauthier believes, that particular acts of constrained maximization are rational. It may be rational to dispose oneself to perform irrational acts, as Thomas Schelling has shown with examples like that of the small country which, for the purposes of deterrence, rationally disposes itself to resist—irrationally—any invasion by its much large neighbour. Is constrained maximization an instance of a (possibly) rational disposition to perform irrational actions? I believe that it is.32

This is a serious objection, for it is an important part of Gauthier's argument that he can move from the rationality of the disposition to cooperate to the rationality of the cooperative action itself.

Consistent with his discussion of the rationality of constrained maximisation, Gauthier is prepared to admit that being disposed, for example, to nuclear retaliation could be rational, when it maximises utility to be so disposed, and, if it is, then it is rational for the nation to carry out its failed threat, even if only a nuclear holocaust will result. These claims are, to say the least, far from plausible. Perhaps sensing in *Morals by Agreement* that the onus is now on him, Gauthier directs us to other arguments: "Deterrence, we have argued elsewhere, may be a rational policy, and non-maximising choices are then rational."33 Since this is the crux of the issue before us, let's look at these other arguments.

[2] The reference in the above quote is to Gauthier's discussion of the rationality of deterrence, in his 1984 paper 'Deterrence, Maximisation,

---


33 Gauthier, MA, p. 186.
and Rationality.' I will start with this paper, and, in particular, in section V, where Gauthier considers various objections to his claim that, if it is rational to adopt deterrent policies, then, if they fail, it is still rational to act on them.

Gauthier considers four objectors; it is the second who is of present concern. We are considering, remember, the person who admits 'the rationality of some deterrent policies, but nevertheless insists that these policies, although fully rational, involve the performance of irrational actions if certain conditions are satisfied.'34 Gauthier's response to this person as follows:

How then does his position differ from mine, in which I claim that deterrent policies may be rational, and if rational, involve performance of actions which, in themselves and apart from the context of deterrence, would be irrational, but which, in that context, result from rational intentions and so are rational? Surely he grants the substance of my argument but expresses his agreement in a misleading and even paradoxical way, insisting that actions necessary to a rational policy may themselves be irrational. To assess an action as irrational is, in my view, to claim that it should not be, or have been, performed.35

Gauthier suggests here that there is little difference between himself and his objector's position. And in this he is right: both agree, in the circumstances, that it is rational to adopt the deterrent policy, and both agree that, were such a threat not met with the required action, anyone who adopted this policy would carry out the threat.

As a response, however, this will not do. Gauthier asks how his objector's position differs from his own, and, as we have just seen, the two positions are very similar. Where they differ, though, is that Gauthier thinks the resulting action would be rational, whereas his opponent thinks it would be irrational. David Lewis puts the point as follows:

I am the second objector, the one who says that "it may be rational to adopt an intention even though it would be, and one knows that it would be, irrational to act on it"; I claim that it may be "rational to commit oneself to irrational behaviour" (and also that it may be good to commit oneself to evil behaviour).

34 Gauthier, DMR, p. 487.
35 Gauthier, DMR, p. 487.
Gauthier claims that my position is no different from his own. Not so; I deny what he firmly asserts, that there may be actions which "in themselves and apart from the context of deterrence would be irrational, but which in that context result from rational intentions and so are rational."\(^{36}\)

Gauthier believes the objector to be confused, and to set him right, he makes it clear that when he says that an action is rational, he means that it ought to be performed. But it seems Gauthier is confused, for what the objector says is exactly that you ought to have the disposition, but you ought not perform the action.

Gauthier has responded to Lewis's objection—he claims that Lewis adopts a position he had failed to consider. Gauthier understands Lewis to be saying that since the disposition and the retaliatory action expressing it are different things, opposed judgements about them are consistent. It's a view he finds 'schizophrenic',

> [b]ut suppose I accept it. ... Suppose that I am a rational actor, considering now what to do should I find myself faced with ADVANT [that is, faced with someone who has ignored my threat]. If I know, as Lewis supposes that I do, that it would be irrational for me to RETAL [that is, retaliate] given ADVANT, then is it possible for me to form the intention to RETAL? It seems clear to me that it is not possible. If Lewis were to say that it would be rational for me to form the intention to RETAL, if I could, then I could understand, although not accept, his position. But I find that I do not understand it.\(^{37}\)

The claim here seems to be that if I believe that it would be irrational for me to form the intention to RETAL, then it is not possible for me to form that intention, and this is inconsistent with the assumption that I can form this intention.

This claim is ambiguous, but on either interpretation of no help to Gauthier.\(^{38}\) On the one hand, it might mean that if I believe it would be irrational for me to form the intention to RETAL, then it is not not

---

36 D. Lewis, 'Devil's Bargains and the Real World,' p. 143.
37 D. Gauthier, 'Afterthoughts,' in D. Maclean (ed.), The security gamble: deterrence dilemmas in a nuclear age, (Totowa, NJ: Rowman & Allanheld, 1984), p. 160. In a similar vein, S. I. Benn, in 'Deterrence or Appeasement? Or, on Trying to be Rational about Nuclear War,' J Applied Phil 1 (1984): 5-20, suggests that if one thinks retaliation is grossly immoral then one cannot form the intention to retaliate.
possible for me, if I am to remain rational, to form that intention. That is, it is not possible for me to remain rational and to form the intention. On this interpretation, what Gauthier says is true, but of no use to him. It is of no use to him since Kavka claims that adopting the deterrent intention is precisely to conditionally intend to perform an irrational action – when one forms the intention, Kavka will say, one thereby becomes irrational. On the other hand, Gauthier’s claim might mean that if I believe it would be irrational for me to form the intention to RETAL, then it is not possible for me, whether I remain rational or not, to form that intention. On this interpretation, what Gauthier says is false, and still of no use to him. It is false because agents can promote the formation of the intention by exposing themselves to external influences which will render them irrational in the relevant respects. It is of no use to Gauthier, since what is to stop his objectors from claiming that, similarly, it is not possible for agents to adopt constrained maximisation once they realise it is a disposition to perform non-maximising, and so presumably irrational, actions? Gauthier’s claim in response to Kavka is either false or not to the point.

[3] Gauthier has another answer to this so-called paradox of deterrence. In response to the third objector in ‘Deterrence, Maximisation, and Rationality,’ Gauthier claims that 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.39

The objector can almost, but not completely, agree with this claim. To see why, note that people such as Lewis and Kavka are claiming that since there are two things under discussion – the disposition to retaliate, and the action of retaliating – then there are two evaluations to be made. The objector can almost agree with this claim because they do not deny that an evaluation can be made of the larger segments of

39 Gauthier, DMR, p. 488.
her activity, such as dispositions. On the rationality of the disposition to retaliate Gauthier and his opponents are not necessarily in disagreement. The objector, though, cannot completely agree with this claim, since Gauthier insists that the evaluation of the smaller segments of activity are to be determined by the evaluation of the larger, while the objectors insist that they require separate evaluation. If the rational dispositions are those which maximise expected-value, then why aren't the rational actions also the ones which maximise expected-value? Gauthier, it seems, has no answer. On the rationality of actually retaliating, then, Gauthier and his opponents are still in disagreement.

This is not the only objection one might have. Kavka admits that 'there may be something to' this wider segments view, and that there are clear advantages of agents acting according to rules, plans, and policies, than on a case-by-case basis. The advantages includes lower decision costs, and more efficient coordination and cooperation. Even so, Kavka 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 outcome is available. This is precisely the situation that arises when deterrence fails in [a Special Deterrence Situation]. There is much harm to be done by retaliation, and the benefit that motivated formation of the intention to retaliate – prevention of the offence – is now unobtainable.  

Gauthier supposes, in bridging principle (B2), that if it is initially rational to adopt a disposition to A when p, and if p, then it is rational A, unless something relevant to the adoption of the disposition has changed except what must have changed with the coming about of p. Thus, for Gauthier, the fact the one's deterrence has failed is no reason to reconsider one's newly formed intention to destroy the world. Kavka offers a different picture, and suggests that if it is initially rational to adopt a disposition to A when p, and if p, then it is rational A, unless there are significant stakes involved and it is clear that the disposition cannot now do the job for which it was adopted. Thus, for Kavka the fact that one's deterrence has failed is more than enough

reason to reconsider one's newly formed intention to destroy the world. The objector can, then, accommodate the intuition towards which Gauthier gestures, but can do so without having to admit - what seems totally implausible - that if it is rational to be disposed to nuclear retaliation, then, in the unlikely case that such deterrence fails, it is also rational to retaliate.

Gauthier, then, also faces problems with his move from the rationality of the cooperative disposition to that of cooperative actions. Even if we suppose, then, that Gauthier has managed to show that it is rational to be disposed to cooperating, he still will have failed to show that it is rational actually to cooperate. This is because there are situations - those involving deterrence - which clearly may very well involve rational intentions to perform irrational actions. Gauthier's indirect strategy seems also to have fallen at the second hurdle, and he joins the advocates of the other bridging principles which we met in Chapter Two.

Conclusion

Gauthier's argument for the rationality of cooperation in the state of nature is not without its problems. In the state of nature, the Self-Interest Theory unconditionally counsels agents to retain the right to use force against others, even though the mutual retention of this right leads to the war of all against all, and short and miserable lives for each. Though many, including now David Gauthier, think this fact can provide some sort of justification for rejecting the voice of such reason, it still seems they are mistaken. Even if we were to accept the contractarian analysis of morality - that one morally ought to perform some action when it is what one would agree to do in certain circumstances - Gauthier cannot show it is rational to be moral. It seems there can be no rational morals by agreement.
The precise details of Gauthier's argument that it is rational to cooperate in some Prisoner's Dilemmas are, as we have seen, problematic. His argument for the rationality of cooperative dispositions, on the one hand, depended illegitimately on the assumption that disposition choice is parametric, and, on the other, employed a defective notion of constrained maximization. Furthermore, his use of a bridging principle connecting rational cooperative dispositions to rational cooperative action was also unjustified, as shown by the so-called paradox of deterrence. Rational morals by agreement seem not to be possible. But only seems not possible, I say. The idea behind the argument is, I believe, essentially sound. In this chapter I shall argue that given certain conditions (to be specified below), the actions resulting from rational 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).

§1 Dealing with Two Objections against Gauthier

Gauthier's argument for the rationality of cooperative dispositions, on the one hand, depended illegitimately on the assumption that disposition choice is parametric, and, on the other, employed a defective notion of constrained maximization. The argument can, however, be reformulated to avoid both of these problems.

[1] Gauthier's argument for the rationality of constrained maximization suffered, as we saw, from the problem of assuming that each agent could, before decision-making commenced, assign a fixed probability to the other's disposition. In other words, that each agent takes their choice of disposition to be the sole variable in a fixed
environment – that choice of disposition is a parametric one. This problem can be remedied.

A better argument for the rationality of cooperative dispositions would assume, rather, that each agent needs to determine, and not take as given, the likelihood of particular choices of dispositions on the part of the others. In other words, that each agent takes their choice of disposition to be but one variable amongst others, so that his choice must be responsive to his expectations of others’ choices, and so on. A better argument would assume that our choice of disposition is a strategic one. Indeed, we need to assume that all choice – and not just that of dispositions – is strategic. We need to assume each agent takes all their behaviour to be but one variable amongst others.

The assumption that choice is strategic leads immediately to the question: how am I, for example, to derive an expectation concerning your behaviour, given that I do not have such an expectation to start with? The key to answering this question is that I will be able to come to some expectation about what you will do if I can put myself in your place, and can assume that you will act rationally. I will be able to come to some expectation about what you will do, that is, if, first, I can determine what you rationally ought to do, and, second, I have conclusive evidence to believe that you would do what (I see) you rationally ought to do. These assumptions need to be explained in more detail.¹

each has conclusive evidence to be acquainted with our decision situation. Continuing this line of reasoning, we see that we also need to assume that we have conclusive evidence to believe that we have conclusive evidence to believe that we are acquainted with our decision situation. And so on. What needs to be assumed, then, is that:

(P2a) (a) Each of us faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, and (d1) each of us has conclusive evidence to believe so, the other has conclusive evidence to believe that we have conclusive evidence to believe so, and so on.

I simply add clause (d1) to assumption (P1) in the previous chapter to get this assumption, (P2a). Clause (a) of (P2a) merely reintroduces the assumptions underlying the basic decision situation we face – that of the state of nature. Clause (d1) introduces the fact, just discussed, that there is common knowledge of the circumstances we face. I say that there is common knowledge that p amongst a population S if and only if (a) everyone in S has conclusive evidence to believe that p, (b) everyone in S has conclusive evidence to believe that everyone in S has conclusive evidence to believe that p, (c) and so on...

To be able to figure out what the other will do, it needs to be assumed, second, that each of us is rational. The first assumption, (P2a), entails that each of us has enough information to be able to figure out what it is rational for the other to do. But such figurings will be useless in predicting the other’s actions unless each has some assurance the other is the sort of person who will come to the right conclusion about what to do, and will then do it. Such figurings will be useless unless it is assumed that

(P2b) (b) Each is doxastically and practically rational, and (d2) there is common knowledge between you and me this is so.

---

Note I have assumed each has conclusive evidence to believe this information, and not that each actually believes it, or that each has reason to believe it. This definition differs from the one offered in D. Lewis, Convention, (Cambridge, Mass.: Harvard Univ. Pr., 1969), p. 56, in that where Lewis talks of reason to believe, I talk of conclusive evidence to believe.
Clause (b) of (P2b) introduces the assumption of complete rationality. This second assumption, (P2b), differs in two ways from the type of assumptions usually made.

First, when most authors assume the agents they are discussing are rational, they in fact explicitly assume that they are maximisers of (consequentialist) value. I have chosen, however, to divide this usual assumption in two: (i) that each of us is practically rational and (ii) that an agent ought to do what maximises (consequentialist) value. I say that an agent is practically rational at some time if and only if they do at that time what they rationally ought to do at that time. What these actions are, of course, will depend on the circumstances, and the particular theory of rational action one most favours. According to the Self-Interest Theory – (ii) – assuming that each of us is practically rational amounts to making the usual assumption that each of us will perform that action with the greatest expected-value to ourselves.

Second, when most authors assume the agents they are discussing are rational, they in fact explicitly assume, in addition, that they are perfect reasoners, and that none of the arguments the authors in question provide in their texts would escape the attention of the agents themselves. Again, I have chosen to split this assumption in two: (i') that each of us is doxastically rational, and that (ii') rational belief is determined by evidence:

(B) An agent rationally ought to believe that \( p \) if and only if they have conclusive evidence to believe that \( p \).

I say that an agent is doxastically rational at some time if and only if they believe at that time what they rationally ought to believe at that time. What these beliefs are, of course, will depend on the

---


5 This is what might be considered the standard theory of rational belief. As a theory, though, it needs further work: (i) what does 'conclusive evidence' mean? (ii) There are problems with sets of infinite beliefs. Given any particular prior belief set, there are infinite propositions one will have conclusive evidence to believe, and many of these it would serve no purpose at all to believe, and so would be propositions one was rationally permitted not to believe. Space prevents me from canvassing how one might deal with these problems.
circumstances, and the particular theory of rational belief one most favours. According to this theory of rational belief -- (B) -- assuming that each of us is doxastically rational amounts to assuming that each of us will believe what we have conclusive evidence to believe. The Self-Interest Theory is a theory of rational action; theory (B) one of rational belief.

What do I mean by saying that an agent rationally ought to believe that p? As with action, it is an 'ought' satisfying three requirements: (a) it is deliberative -- it would be irrational to judge I ought, in this sense, believe that p and yet keep deliberating about whether or not to believe that p; (b) it is (strongly) belief-guiding -- it would be irrational to judge I ought, in this sense, believe that p and yet not believe it. And, finally (c) it is absolute, rather than relative -- to make a judgement I ought, in this sense, to believe that p is to make a judgement not relativised, for example, to evidence (though perhaps made on the basis of evidence).

We will henceforth be dealing with two theories of rationality -- one of rational action, (S), the other of rational belief, (B). I shall argue, however, that whenever we run into problems as a result of assuming both the Self-Interest Theory of rational action, and Theory (B) of rational belief, it is the first rather than the second which should be rejected. Assumptions (P2a) and (P2b), then, define the conditions of strategic choice in the state of nature.

Gauthier need not have assumed that disposition choice was parametric. As we shall see presently, the conclusion he desires will still follow from the above assumptions that such choice is strategic.

[2] Gauthier's notion of constrained maximization, we also have seen, is not unproblematic. In particular, he has difficulties identifying the grounds on which a constrained maximizer would actually cooperate. This problem can also be remedied.

What we need to do is to build the circularity we need into a promise, rather than into a disposition. First, we need to introduce the possibility that each of us can make such a promise to the other:

(P2c) (c) Each one of us, before we cooperate or defect, has the option of promising to cooperate or not; each choice is causally independent of that of the other; and each of us, after the other has agreed to or not, will have conclusive evidence to believe they have in fact done so
or not. (d3) There is common knowledge between you and me that this is so.

Clause (c) of (P2c) introduces the possibility – consistent with all we have assumed to this point – that you and me can communicate, and, in particular, that each of us can make promises to the other before we do anything about relinquishing our right to use force.

Second, we need to determine the form a promise needs to take if it is to be effective. There are a number of possibilities. (1) I, for example, might simply say: ‘I promise to lay down my right to use force’. But this is no good, since if I am disposed to keeping my promises, then this will only make me prey for you – I will lay down my right unconditionally, and you will then be able to exploit me. I need to make the promise conditional. (2) I might make such a promise, and say: ‘I promise to lay down my right to use force, if you would lay down your right.’ But this is also no good, since even if both of us make this promise and are disposed to keep it, then we still will not have ensured that we will lay down our rights to use force – we might each be waiting for the other to go first. I need to make the conditional promise circular. (3) The proper form of the promise is in fact (P) ‘I promise that: I will lay down my right to use force against the other if and only if the other says (P).’ If each of us were promise keepers, and each made this promise, then each of us would cooperate.

We have thus built the circularity we require into a promise, rather than into a disposition. And note that while statement (P) is self-referring, the circularity is not in this case vicious, since such self-reference is not unusual in language. The trick to surmount the second objection to Gauthier, therefore, is to replace a problematically circular notion of constrained maximization with an unproblematically circular promise of cooperation.

---

This seems to be the form the promise takes in Hobbes. See his English Works, vol. ii, pp. 91-2, and D. Gauthier, The Logic of the Leviathan, (Oxford: Clarendon Pr., 1969), pp. 103-4. As we have also noted, it is the type of promise a constrained maximiser would make.

This sort of promise is discussed by J. H. Sobel, 'The Need for Coercion,' in J. Pennock & J. Chapman (eds.), Coercion, (Chicago: Aldine-Atherton, 1972), pp. 171-176. In his discussion, Sobel does not consider (as I shall, later in the chapter) any type of person other than the ones I will call Agreement-Keeper, and does not argue (as I shall) that if it is rational to become an Agreement-Keeper, it is rational to actually keep one's agreements. My argument is an extension of Sobel's, itself partly inspired by Gauthier's discussion in 'Morality and Advantage,' Phil Rev 76 (1967): 460-75.
The first two objections I considered to Gauthier's position can, then, be adequately met. I will deal with the third – the paradox of deterrence – in the final chapter of the thesis. For the rest of this chapter, though, I want to continue with the reformulation of Gauthier's argument that, under certain conditions, it might be rational to cooperate in the Prisoner's Dilemma.

[3] Suppose, then that we find ourselves in the state-of-nature, are well-informed about our situation, are completely rational and can make promises. A situation such as this involves rational irrationality, for it is possible to argue that each of us rationally ought promise to cooperate, even though each of us rationally ought not actually to cooperate.

First, one can argue that we rationally ought to agree to cooperate.\(^8\) (a) If I, for example, promise to cooperate then there are two possibilities: you will believe that I will carry out my promise, or you will not. If you believe that I will carry out my promise then there is a (small) chance that you (foolishly) will cooperate by laying down your arms. If you do cooperate, I will be able to exploit you, since even though I promised to cooperate I will not actually be so foolish as to do so. If you do not believe that I will carry out my promise, then you will presumably not cooperate, so we merely remain in the state of nature. Thus, on the one hand, if I promise to cooperate, then there is a (small) chance I will be able to exploit you, and a (large) chance it won't make any difference. (b) If, however, I do not promise to cooperate, there is presumably only one possibility: you will not cooperate. Thus, on the other hand, if I do not promise to cooperate, then it also won't make any difference. Considering (a) and (b) together, I might conclude I rationally ought to promise to cooperate.

Second, one can also argue that, even if we rationally ought to agree, it will still not be rational for us to keep the agreement.\(^9\) (a) If you were to cooperate, then I would do better by not cooperating, since

---

\(^8\) I should point out, though, that I do not necessarily endorse the argument – I introduce it simply for the sake of illustration.

I would then be able to exploit you. (b) If you were not to cooperate, then I would again still do better by not cooperating, by preventing the possibility that you will exploit me. Considering (a) and (b) together, I (once again) conclude I rationally ought not to cooperate.

The belief in this claim – that even if it is rational for us to agree to cooperate, it is irrational for us actually to cooperate – depends on the Self-Interest Theory (S), Theory (B) of rational belief, as well as an extended version of the conditions, (P1) in the previous chapter, which defined the Prisoner’s Dilemma:

(P2) (a) Each of us faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, (b) each of us is completely rational, (c) each of us, before we cooperate or defect, can make an agreement with the other, and (d) there is common knowledge between us that all this is so,

I simply add, then, clauses (b), (c), and (d) to assumption (P1), in the previous chapter, to get this assumption, (P2). That is, I simply add, to the Prisoner’s Dilemma, the stipulation that each of us is well-informed, completely rational, and can make each other promises. I shall call (P2) the *Promise Puzzle*.

In these situations, rational irrationality again rears its ugly head. It is rational irrationality, though, which any sensible person would not find conceptually troubling, since it is a commonplace it might be rational to make false promises, particularly if one has no assurance that others will do their part.

§2 Rational Agreement-Keeping Dispositions!

It may seem that promises, even if correctly formulated, are useless to induce cooperation in the state-of-nature. But this need not be so. In particular, if [1] each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react, then [2] it is rational to be the sort of person who keeps agreements.
[1] In this section, I want to introduce and motivate an extra assumption: that each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react.

As we have seen, the state of nature is, for Hobbes, a state of war of all against all, in which there can be no security for anyone (be they however strong or wise). As a consequence,

\[ \text{it is a precept, or general rule of Reason, } \text{That every man, ought to endeavour Peace, as farre as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre. The first branch of which Rule, containeth the first, and Fundamentall Law of Nature; which is, to seek Peace, and follow it. The Second; the summe of the Right of Nature; which is, By all means we can, to defend our selves.} \]

Hobbes claims two things in the passage. On the one hand, he claims that a rational agent would cooperate, and renounce their right to use force, if they have the assurance that others would keep their side of the bargain, and renounce their right to use force. This Hobbes calls the Fundamental Law of Nature. On the other hand, he claims a rational agent would not cooperate, and renounce their right to use force, if they do not have the assurance others would keep their side of the bargain, and renounce their right to use force. This is what Hobbes calls the Right of Nature.

[1.1] If Hobbes is right, then any adequate theory of rationality will have to meet two requirements. On the one hand, an adequate theory will imply that rational agents would cooperate when they have the assurance that others would keep their side of the bargain. On the other hand, an adequate theory will imply that rational agents would not cooperate when they do not. Does the Self-Interest Theory meet these requirements?

First, the Self-Interest Theory implies rational agents would not cooperate when they have no assurance that others would keep their side of the bargain. Whatever the other does, it has the best outcome for one not to cooperate, and, according to the Self-Interest Theory, not 10

---

cooperating is what one rationally ought to do. This is so particularly if there is no assurance the other would cooperate. If one is a rational agent without such an assurance, then, the Self-Interest Theory correctly implies one would not cooperate.

Second, the Self-Interest Theory does not, however, imply rational agents would cooperate when they do have the assurance others would cooperate. Suppose, for example, we are two agents who can assure the other that we would cooperate. Suppose, in particular, that each of us, by taking a certain pill, can become transparently trust-worthy. We would cooperate with anyone similarly disposed, even if it did not have the best outcome for us to do so, and this would be obvious to anyone who met us. If you, for example, are like this, then you can give me an assurance that you would cooperate. To do so you would simply have to take the pill: you would become trustworthy; I would come to believe that you were (since you would be transparent); and I would have the assurance I need.

It is not clear, however, the Self-Interest Theory implies that we rational agents would cooperate when we have this type of assurance. To see why, consider two possibilities: that our mutual rationality will last until the time for cooperating, or that it will not.

(i) Suppose, on the one hand, that our mutual rationality lasts until the time comes for cooperating. Come the time for cooperating, it still remains that each of us would only obtain a loss from doing so, and thus, if the Self-Interest Theory is to be believed, each rationally ought not to cooperate. If our mutual rationality will endure to the time for cooperating, then, according to the Self-Interest Theory, each of us will not cooperate. In this first case, the Self-Interest Theory implies that initially rational agents would not cooperate, even if there is the indicated possibility of assuring the other that they will not be exploited.

(ii) Suppose, on the other hand, that our mutual rationality will not necessarily endure until the time for cooperation. In this case, the Self-Interest Theory would not have the implication described in case (i). But it would also fail to imply that two initially rational agents would cooperate. If we drop the assumption that we remain mutually rational up to and including the time for cooperation, there are no grounds for making any predictions about what we will do come this time, and thus no grounds for supposing that we, as two rational agents, would in fact cooperate. In this second case, the Self-Interest
Theory fails to imply that initially rational agents would cooperate, if there is the indicated possibility of assuring the other they will not be exploited.

On these grounds, the Self-Interest Theory should be rejected. On the one hand, it implies – reasonably enough – that rational agents would not cooperate when they have no assurance that others would keep their side of the bargain; but, on the other, it does not imply – surely unreasonably – that rational agents would (and may in fact imply that they would not) cooperate when – as above – they do have the requisite assurance. If Hobbes is right, then, an adequate theory implies that rational agents would cooperate when they have the assurance others would keep their side of the bargain, and so the Self-Interest Theory is inadequate. But even if Hobbes is not in general right, it should be clear in the above example that rational agents would indeed become transparently trustworthy, and thus avoid for themselves the ravages of the state of nature. Whether or not Hobbes is in general right, the Self-Interest Theorist will want to argue two initially rational agents will come to cooperate in the situation I described above.

[1.2] The best response for the Self-interest theorist at this point is to reject the assumption that two initially rational agents – you and I – will necessarily remain rational come the time for cooperation, and to assume instead we are rational up to, but not necessarily including, this time.\footnote{On this response, see D. Parfit, Reasons and Persons, (Oxford: Clarendon Pr., 1984), pp. 45-49, where he discusses whether being rational could ever be a mere means.} The Self-Interest Theorist need not be enamoured with rationality per se, and might suppose agents will hold onto rationality only up to the point it begins to get in the way of producing the best outcomes. The Self-Interest Theory advises one to perform those actions which produce the best outcomes for oneself, and if such actions involve making yourself disposed to perform later irrational actions (as it seems they will in the case I have introduced), then the initially rational agent is happy to treat their own rationality as a mere means.

To drop the assumption that our initial rationality will endure puts the Self-Interest Theorist on the other horn of the dilemma, since it then seems that their theory lacks any resources for showing that we
will come to cooperate. This is not a problem, however. For, rather than assuming we will make our decisions about cooperating when the time comes for action, and assuming with this that we will be rational when this time comes, it would be better to assume instead that we make our decisions concerning cooperation beforehand, while we are still assumed to be rational. The Self-Interest Theorist may assume, as I shall follow them in assuming, that while we are rational beforehand we can commit ourselves to various plans of action, plans making later actions of cooperation contingent upon our believing that certain events have occurred.

What plans of action are available to us? There are two actions we will face later on: cooperating or defecting. There are two contingencies upon which each of us can base our action: our having agreed to cooperate, or our not having so agreed. There are thus four (=2×2) complete ways of reacting you, for example, might adopt:

(C|C) you make it now that you cooperate regardless of what (you believe that) we agreed to beforehand: if (you believe that) we agreed to cooperate then you cooperate, and if (you believe that) we did not agree then you cooperate.¹²

(C|D) you make it now that you cooperate if and only if we agreed to cooperate: if we agreed to cooperate then you cooperate, and if we didn't then you defect,

(D|C) you make it now that you cooperate if and only if we did not agree to cooperate: if we agreed to cooperate then you defect, and if we didn't then you cooperate,

(D|D) you make it now that you defect regardless of what we agreed to do beforehand: if we agreed to cooperate then you defect, and if we did not then you cooperate.

¹² These different plans are also called strategies, and are denoted by '(x|y)', which indicates that if (one believes) we have agreed to cooperate, then one will do x, and if (one believes) we have not agreed to cooperate, then one will do y. The notation is that of N. Howard, The Paradoxes of Rationality: the Theory of Metagames and Political Behaviour, (Cambridge, Mass.: MIT Pr., 1971). Note that since these strategies are concerned with internalised conditional commitments to act, they should, strictly speaking, be understood as conditionals relating two internal states of an agent. Thus, strictly speaking, if you adopt (CID), then you are disposed to cooperate if you believe we have agreed, and not simply if in fact we have agreed. However, it gets tedious to include such a qualification, and I shall do so only when necessary.
Three of these strategies deserve particular mention. The first, (C!C), is the disposition someone we could call a 'Kantian' would have. This person performs that action they would will to become a universal law. This, of course, is cooperation (whether or not it was agreed to), since universal cooperation is valued higher than universal defection. The second, (C!D), is the disposition of the person I will call the Agreement-Keeper. Such a person cooperates with those with whom they have agreed, but defects otherwise. Unlike the 'Kantian', they respond to a failure to agree with defection. The third, (D!D), is the disposition of the enduringly Self-Interested agent, who at all times performs that action which produces the best outcomes for themselves. Unlike the previous two agents, they respond to an agreement with defection.

Hobbes is right to suggest that without an assurance that the other is trustworthy, a rational agent would not cooperate, and may be right also to suggest that with such an assurance, a rational agent would cooperate. What he would be wrong to suggest, though, is that a person could have such an assurance from another only if there were some coercive power standing over both, enforcing the agreements made between them. If you, for example, can choose how to be disposed, and if you are transparent, you would be able to give me the requisite assurance without having to defer to some coercive power. To do so, you would simply have to become either a (C!C) or a (C!D): you would become the sort of person who would cooperate if we had agreed; I would come to believe you were (since you would be transparent); and I would thus have the assurance I need.

Any adequate theory of rationality will imply that two agents will cooperate if they find themselves in a situation where each can offer the other some assurance that they will not be exploited. A situation in which our dispositions are transparent is just such a situation, and the Self-Interest Theory can imply that we two initially rational agents would cooperate only if, it seems, we assume that:

---

(P3) (x) Each of us, before we promise to cooperate or not, has the option of adopting one of the strategies (CIC), (CID), (DIC) or (DID); the choice each makes is causally independent of that of the other; and each of us, after the other has adopted one of their possible strategies, will have conclusive evidence to believe they have in fact done so; (a) each faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection; (b) each of us is completely rational; (c) each of us, before we cooperate or defect, can make an agreement with the other; and (d') there is common knowledge between us that all this is so.14

I simply add clause (x) to assumption (P2), introduced above, to get this assumption, (P3). That is, I simply add, to the Promise Puzzle, the stipulation that each can choose how they would react to an agreement, though, after the fact, such a choice will be transparent to the other. I will call (P3) the Third Counterexample (the first two, of course, occurring in the discussion of the Toxin Puzzle in Chapter Two).

[2] I now want to argue that if each of us can choose how we are disposed to react to any agreement made between us, and if, after we have made our choice, it is possible for the other to tell how we have chosen to be disposed to react – that is, if assumption (P3) obtains – then it is rational to become an Agreement-Keeper. In showing this, three times are important. They are indicated on the following line:

\[
\begin{align*}
\text{t1} & \quad (CIC)...(DID) \\
\text{disposition} & \quad P, \sim P \\
\text{t2} & \quad \text{promise} \\
\text{t3} & \quad \text{cooperate} \\
\text{t1} & \quad \text{cooperate (or not)} \\
\end{align*}
\]

Initially, at time t1, each is free to adopt one of the dispositions (CIC) to (DID). At a time later than this, t2, each is free to make a promise or

---

14 This is, of course, a very restrictive assumption. The contingencies upon which one might base the action of cooperating are presumably numerous, and the plans or dispositions one has adopted will almost never be transparent. In response, I reiterate my earlier (unsubstantiated) claim that even when one does consider more realistic types of situation, then it will still remain the case that some non-expected-value maximizing cooperative disposition is rational. Substantiating this claim is work for another time.
not. And at a time still later, t₃, each is free to keep the promise, and cooperate, or not, and defect. The situation we both initially face at time t₁ (while we are still well-informed and completely rational) can be depicted as follows:

<table>
<thead>
<tr>
<th></th>
<th>You</th>
<th>Me</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>P</td>
<td>~P</td>
<td>P</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>P</td>
</tr>
<tr>
<td>~P</td>
<td>c,c</td>
<td>s,t</td>
</tr>
<tr>
<td>C</td>
<td>D</td>
<td>P</td>
</tr>
<tr>
<td>~P</td>
<td>t,s</td>
<td>t,s</td>
</tr>
<tr>
<td>D</td>
<td>C</td>
<td>P</td>
</tr>
<tr>
<td>~P</td>
<td>c,c</td>
<td>s,t</td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td>P</td>
</tr>
<tr>
<td>~P</td>
<td>t,s</td>
<td>t,s</td>
</tr>
</tbody>
</table>

Key: x/y = making it at t₁ that: one does x at t₃ if we agree at t₂ and y if not; P = promising at t₂ to cooperate at t₃; t = the value of exploiting the other at t₃; c = the value of mutual cooperation at t₃; d = the value of mutual defection at t₃; s = the value of being exploited by the other at t₃. (NB: t > c > d > s.)

Each of us is free to adopt one of the strategies (C|C) to (D|D), and the outcomes of all combinations of these are as shown. The immediate outcome of a pair of strategies is, itself, a situation in which both of us are free later with respect to another action, this time about whether or not we are going to make the promise, (P), to cooperate.

Take, for example, the outcome if I adopt (C|D), and you also adopt (C|D). Whether or not cooperation will immediately result is not yet given. What is given is that we will face a later situation of the form:

<table>
<thead>
<tr>
<th></th>
<th>You</th>
<th>Me</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C</td>
</tr>
</tbody>
</table>
It is straightforward to see this is so. Suppose that each has made themselves an Agreement-Keeper – you and I have each performed action (C | D). Then I (ditto: you) am such that if I promise to cooperate, and believe that you have promised, then I will cooperate. Later, each of us is free to promise or not: P or ~P, and the outcomes of all combinations of these are shown. On the one hand, if each of us promises (I do P, and you do P), then I will believe that we have done so, and I will cooperate. Since, similarly, you will also cooperate, the result is an outcome (mutual renunciation of our right to use force) which each values at 'c'. And this is what is in the top left-hand cell. On the other hand, if one of us does not promise (I do ~P or you do ~P or both), then each will know, and neither will cooperate; the result being an outcome (mutual retention of the right to use force) which each values at 'd'. And this is what is in the remaining cells. Performing this sort of reasoning with all other possible combinations of strategies gives the complex decision matrix indicated above.

The situation we both initially face (while we are still well-informed and completely rational) may be simplified, to obtain:

|       | CIC | CID | DIC | D|D |
|-------|-----|-----|-----|---|
| CIC   | c,c | s,t | c,c | s,t|
| CID   | t,s | c,c | t,s | d,d|
| Me    |     |     |     |   |
| DIC   | c,c | s,t | c,c | s,t|
| D|D   | t,s | d,d | t,s | d,d|

For even though the immediate outcome of a pair of strategies is, as we have seen, a situation in which each of us is free to promise or not, we can determine now what it would be rational then to do, and so determine (if we assume that we will at the time of promising still be rational) what we will then do, and so determine the outcome of our adoption now of various strategies.

Take again, for example, the resulting situation if I adopt (C | D), and you also adopt (C | D). The Self-Interest Theory says to each of us, in this case, to promise, for each knows that the other will adopt either P or ~P, and know that if they choose P in this case then they will never
do any worse, and may very well do better, than if they were to adopt any other strategy. Of either choice, in this case, P is the only one they have nothing to lose by performing. Each is initially uncertain about what the other will do. Therefore, each rationally ought to do P, and (if we assume that we will then be rational), it follows that each of us will do P, and the result will be mutual renunciation of our right to use force, and an outcome we value at c,c. Performing this process with the remaining fifteen possible combinations of strategies gives the simplified matrix indicated.

This inference employs the following dominance, or ‘sure-thing’, principle: if action A weakly dominates action B over partition P, and one is uncertain concerning P, then one rationally ought not to B. This principle employs two key notions: (a) A weakly dominates B over P when, for all p in P, the expected value of A&p is no less than that of B&p, and, for at least one p in P, it is greater; (b) one is uncertain concerning P when, for all p in P, one assigns no probability to p. I am initially uncertain about what you will do, since ours is a strategic situation, in which we do not take as given the likelihood of particular behaviour on the part of the other.

Note, we have to be careful how we apply dominance reasoning with regard to the complex matrix with which we started. I may not, for example, eliminate from consideration the line (C ⇔ D) | P – where I become an Agreement-Keeper and then promise to cooperate – because it is dominated, as it is, by the line (D ⇔ D) | P – where I remain a Straightforward Maximiser, but then agree to cooperate. The reason is simple. I may use dominance reasoning amongst my actions A₁, A₂, ..., Aₙ, relative to your actions B₁, B₂, ..., Bₙ, only if your choice of Bᵢ is causally independent of my choice of Aᵢ. However, your choice, for example, of (C ⇔ D) | P is not causally independant of my choice, for example, of (D ⇔ D) | P. If I were to choose (D ⇔ D) | P you would not be so foolish as to choose P, and, a fortiori, would not be choosing (C ⇔ D) | P. Hence, I may not use dominance reasoning amongst my actions (C ⇔ C) | P, ..., (D ⇔ D) | ~P relative to your actions (C ⇔ C) | P, ..., (D ⇔ D) | ~P. I may, however, use dominance reasoning (as I will) amongst my (C ⇔ C), ..., (D ⇔ D) relative to your (C ⇔ C), ..., (D ⇔ D), since these are causally independent. (This is assumption P3(x).) And I may use dominance reasoning amongst my actions (C ⇔ C) | P, ..., (D ⇔ D) | ~P relative to your actions (C ⇔ C) | P, ..., (D ⇔ D) | ~P.

---

reasoning (as I have) amongst my $P, \sim P$ relative to your $P, \sim P$ since these are also causally independent. (This is assumption P3(a).)

On the basis of this simplified matrix, we can see that each of us rationally ought to become Agreement-Keeper. Each knows that the other will adopt one of the strategies (C₁C) to (D₁D), and know that if they choose (C₁D) then they will never do any worse, and may very well do better, than if they were to adopt any other of their own strategies. Strategy (C₁D) is the only one they have nothing to lose by adopting, and so each ought to adopt (C₁D). Each is initially uncertain about what the other will do. Each of us ought to make ourselves Agreement-Keeper, and, if we do, then we each ought to make the promise, thus agreeing to cooperate, and thus eventually cooperating by casting off our right to use force against the other, and escaping from the war of the one against the other.

If we are well-informed and completely rational agents in the state-of-nature who can make agreements with each other, and who can choose how to be disposed with regard to keeping such agreements, though such a choice is, after the fact, transparent to the other person, then we rationally ought to become agreement-keepers, would become agreement-keepers and would thus cooperate. But would it be rational for us to cooperate? Or would it just be another instance of rational irrationality?

§3 Rational Agreement-Keeping Actions!

Even if Gauthier's argument, introduced in the previous chapter, for the rationality of constrained maximization had established it might be rational to be disposed to cooperating, some objected this would not imply it was rational actually to cooperate. In the remainder of the chapter I will argue, firstly, that – contrary to the Self-Interest Theory – we are indeed rationally permitted to cooperate in the Third Counter-example, and, secondly and more generally, that given certain conditions (to be specified below), the actions resulting from rational 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).
The first half of the task for this section consists in demonstrating that in the Third Counterexample, we are rationally permitted to cooperate. I shall need to reply to the Foole, who

hath sayd in his heart, there is no such thing as Justice; and sometimes also with his tongue; seriously alleging, that every mans conservation and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also make, and not make; keep, or not keep Covenants, was not against Reason, when it conduced to ones benefit. He does not therein deny, that there be such Covenants; and that they are sometimes broken, sometimes kept; and that such breach of them may be called Injustice, and the observance of them Justice: but he questioneth, whether Injustice, ... may not sometimes stand with that Reason, which dictateth to every man his own good;16

Since, whatever the other does with their right to use force, each of us does worse for ourselves to cast off our own, the Foole will say, whether in his heart or in print,17 that it is not rational for us to do so, that we rationally ought not to do so.

My answer to the Foole takes the form of the following argument. [1.1] In the Third Counterexample, if the Self-Interest Theory (S) and Theory (B) of rational belief are both true, then so are all the following:

(1) Each rationally ought (at t1) to become an Agreement-Keeper;
(2) Each rationally ought (at t2) to believe we have agreed to cooperate; and
(3) Each rationally ought (at t3) to not cooperate.

Yet, since one rationally ought to S1, rationally ought to S2, and rationally ought to S3 only if it is logically possible that one S1, S2, and

S3 – that is, since (OP3) is true – and since it is not logically possible to become an Agreement-Keeper, believe we have agreed to cooperate, but not cooperate, then one of these claims must be false. [1.2] We have no good reason against thinking – (2) – that each rationally ought (at t2) to believe we have agreed. [1.3] And we have no good reason against thinking – (1) – each rationally ought (at t1) to become an Agreement-Keeper. Therefore, the final claim – (3) – must be false, and it must indeed be rationally permitted for each to cooperate. The Foole is wrong.

[1.1] The first premise of the argument is that, in the Third Counterexample, if the Self-Interest Theory (S) and Theory (B) of rational belief are both true, then so are (1), (2) and (3).

(1) I spent all of section §2[2] arguing that each rationally ought (at t1) to become an Agreement-Keeper. I shall not repeat the argument here.

(2) Each, then, rationally ought to become Agreement-Keeper. Since each is completely rational at the time of choosing dispositions, it follows each becomes an Agreement-Keeper. Each is transparent, and so each will have conclusive evidence to believe that the other is an Agreement-Keeper, and so (according to (B)) each rationally ought to believe so. Since each is completely rational at the time of promising, it follows each believes the other is an Agreement Keeper. In this case it has the best outcome to promise and so (according to (S)) each rationally ought to promise. Since each is completely rational at the time of promising, it follows each promises to cooperate. Each can tell whether the other has made a promise or not, and so each will have conclusive evidence to believe we have made an agreement, and so (according to (B)) each rationally ought to believe we have agreed.

(3) However, it still has the best outcome not to cooperate. If theories (S) and (B) are true, then each rationally ought not to cooperate.

Yet, since one rationally ought to S1, rationally ought to S2, and rationally ought to S3 only if it is logically possible that one S1, S2, and S3 – that is, since (OP3) is true – and since it is not logically possible to become an Agreement-Keeper, believe we have agreed, and yet not cooperate, then one of the claims (1)-(3) must be false. But which one?
The second premise of my argument is that we have no good reason against thinking - (2) - that each rationally ought (at t2) to believe we have agreed. We have no reason for rejecting the second statement.

The Poole, in order to save the Self-Interest Theory, (S), of rational action, may propose to reject theory (B), of rational belief. A response such as this, it might be suggested, is indeed the one the defender of the Self-Interest Theory would be inclined to make, since this commonly held view about the nature of rational belief is no part of the Self-Interest theory and the Poole can offer his own, alternative, account of rational belief. Roughly, this is that a person rationally ought to believe that p if and only if the expected-value to them of the outcome of believing that p exceeds that of not believing. Such a theory of rational belief is particularly a propos to the situations we are considering, for again, no matter what you do, it has the best outcome for me not to believe we have agreed, since then I will not cooperate.

There are two reasons for dismissing this response. First, this Self-Interest theory of rational belief is not without independent problems. It is difficult to know whether or not real persons could actually follow the dictates of such a theory, and it may well be constituitive of belief that in a large proportion of cases conclusive evidence for believing results in the relevant belief. Furthermore, it may well be impossible to believe what maximises expected-value without having a significant proportion of one’s beliefs sensitive to evidence rather than expected-value. Second, and more importantly, rejecting theory (B) is more than we need to do in order to free ourselves from our dilemma. I have already argued in the first part of the thesis that the Self-Interest Theory is false, and thus should be rejected. This argument in no way implicated Theory (B) of rational belief, and so gave us no reason for rejecting this theory. So it is in this case. The conservative strategy - to reject in this case what needs to be rejected anyway - is to reject the Self-Interest Theory, (S). I shall consider no further the possibility that Theory (B) is false, and the possibility that (2), above, is false.

This leaves only statements (1) and (3), one of which must be rejected. But which one?

---

[1.3] The third premise of my argument is that we have no good reason against thinking – (1) – that each rationally ought (at t₁) to become an Agreement-Keeper. We have no reason to reject the first statement, and reason to reject the third.

I will show this by providing an argument directly for the claim each is rationally permitted to cooperate. In the Third Counterexample, there are eight ways you, for example, could be:

(a) \( A \) \( B \) \( C \)  
(b) \( A \) \( B \) \( \neg C \)  
(c) \( A \) \( \neg B \) \( C \)  
(d) \( A \) \( \neg B \) \( \neg C \)  
(e) \( \neg A \) \( B \) \( C \)  
(f) \( \neg A \) \( B \) \( \neg C \)  
(g) \( \neg A \) \( \neg B \) \( C \)  
(h) \( \neg A \) \( \neg B \) \( \neg C \)

where: \( A = \) becoming an Agreement-Keeper; \( B = \) Believing we have agreed to cooperate; and \( C = \) Cooperating. Maximally rational persons believe all they rationally ought to believe: if you are maximally rational, and rationally ought to believe that \( p \), then you would believe that \( p \). We have already established that you rationally ought to believe we have agreed – hence if you were maximally rational, then the only ways you might be are: (a), (b), (e), and (f). However, whether or not you are maximally rational, it is not logically possible for you to be an Agreement-Keeper, believe we have agreed, and yet not cooperate – therefore (b) is not really a way you might be. Hence, if you were maximally rational, there are only three ways you might be:

(a) \( A \) \( B \) \( C \)  
(e) \( \neg A \) \( B \) \( C \)  
(f) \( \neg A \) \( B \) \( \neg C \)  

Furthermore, maximally rational agents are those who can expect to most promote what they value. Of these three ways you might be: best is (a) – since you are an Agreement-Keeper, I will cooperate, and civil society will be possible; next is (f) – since you are not an Agreement-Keeper, I will not cooperate, and it would be foolish for you to do so; worst is (e). Hence, if you were maximally rational, then you would
become an Agreement-Keeper, believe we have agreed, and therefore cooperate. But anything a rational person would do is something ordinary mortals, like you, are rationally permitted to do: if you would A were you maximally rational, then you are rationally permitted to A. You would cooperate were you maximally rational, and so you are rationally permitted to do so.

In this way, then, we see that in the Third Counterexample you are rationally required to become an Agreement-Keeper, rationally required to believe we have agreed, and thus rationally permitted to cooperate, even though you are free not to cooperate (P3(a)) and it has the best outcome for you not to cooperate (P3(a)). This claim may be generalised.

[2] The second half of the task for this section is to demonstrate that, GIVEN that you rationally ought to adopt the enduring disposition to do what we have agreed to do, THEN if you rationally ought to believe we have agreed to do something, you are rationally permitted to perform your part of the bargain, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise (though doing so may very well be irrational absent this condition). The demonstration relies again on the deontic principles I introduced above.

The argument for this claim is as follows. The deontic principle (OP3) says that one rationally to S1, rationally ought to S2, and rationally ought to S3 only if it is logically possible that one S1, S2, and S3. Letting S1 be 'adopt the enduring disposition that if you believe we have agreed to A then you A', S2 be 'believe we have agreed to A' and S3 be 'not A', it follows from (OP3) that

(1) You rationally ought to adopt the enduring disposition that if you believe we have agreed to A then you A, you rationally ought to believe we have agreed to A, and you rationally ought not to A only if it is logically possible for you to adopt the disposition, have the belief, and yet not A.

Yet, as we have seen, this is not possible. On the basis of some propositional logic (in particular, the principle that D & B & A → P and ¬P entail D → (B → ¬A)), it follows from (1) that GIVEN that you rationally ought to adopt the enduring disposition to do what you believe we have agreed to do, THEN if you rationally ought to believe
we have agreed to perform some action, then you are rationally permitted to perform that action. The above argument depends on no particular assumptions about the action in question — A. Hence, our conclusion obtains Even if one is free not to A, and it has the best outcome not to A (though, I should point out, it may not be rationally permitted to A if the above conditions are not given).

Conclusion

The Foole, then, is wrong, and under some (but not all) conditions, rational agreements make for rational actions. The Self-Interest Theory is false. In the state of nature, it unconditionally counsels agents to retain the right to use force against others, even though the mutual retention of this right leads to the war of all against all, and short and miserable lives for each.

I started this thesis with the worry morality might be irrational. But if we can defend the contractarian analysis of morality — that one morally ought to perform some action when it is what one would agree to do in certain circumstances — and if (as I believe) Gauthier's type of argument applies to more realistic types of situations — involving more than four dispositions from which to choose, and dispositions not transparent — then we may perhaps be able to show it is rational to be moral.

It seems, after all, that there can be rational morals by agreement, and that progress might perhaps be made on what Sidgwick called 'the profoundest problem of ethics.'
I have argued that the Poole is wrong, and that given the rationality of a certain disposition, rational agreements do indeed make for rational actions. My argument for these claims, and particularly the first, depends (amongst other things) on a central assumption:

(P3) Third Counterexample: (x) Each of us, before we promise to cooperate or not, has the option of adopting one of the strategies (CIC), (CID), (DIC) or (DID); the choice each makes is causally independent of that of the other; and each of us, after the other has adopted one of their possible strategies, will have conclusive evidence to believe they have in fact done so; (a) each faces an independent choice between cooperating or defecting; defection dominates cooperation for each of us; even though each of us values joint cooperation to joint defection, (b) each of us is completely rational, (c) each of us, before we cooperate or defect, can make an agreement with the other, and (d) there is common knowledge between us that all this is so.

The Poole, however, might claim that when I add clause (x) to the Promise Puzzle, (P2), to get (P3), I change the situation in a crucial way. As we will see, this is to claim, in effect, that assumption (P3) is inconsistent.¹ There are two broad types of objections to these assumptions, and in this chapter I shall argue that neither of them succeed.

§1 Is it really still possible for you not to cooperate?

I constructed the Third Counterexample by adding to the Promise Puzzle, (P2), the stipulation that each of us, before we promise, is free to choose how we are disposed, and that this choice is, afterwards,

transparent to the other person. In effect, I added the stipulation that it maximises expected-value for each to adopt the enduring disposition of always doing what we agree to do, to get (P3). I claimed that, in these modified situations, you are rationally permitted to keep your agreement, and cooperate, even though you are free not to do so, and it maximises expected-value for you not to do so. The Foole, however, may object that this stipulative addition means each is no longer free not to cooperate. As I shall argue in this section, though, this objection is mistaken, and in the Third Counterexample each may indeed still be free not to cooperate.

[1] There are some² who suggest that constrained maximisers – who will cooperate if they believe others are likely to cooperate – are the sort of persons who cannot but cooperate if they believe others are likely to cooperate:

If CM is a permanent mechanism inducing co-operation in spite of defection being [value]-maximizing, and therefore, presumably being the more preferable action at the time of actual choice of action, it would seem that there is a crucial sense in which, at that time, one is not acting voluntarily – that is, from an immediate preference so to act – but merely at the behest of the mechanism. ...

On the mechanism interpretation, Gauthier seems wrong to think of co-operation as voluntary, free, and rational action.³

We need to know more about the disposition to cooperate. We do know that it is not a disposition to consciously decide on economic grounds to comply with agreements. If the CM were in the position to deliberate, then utility maximization would sometimes dictate noncompliance, and it would not be economically rational to adopt the disposition. Internalizing the disposition to cooperate entails binding oneself to comply strongly enough to overcome the inclination to straightforwardly maximize. In the absence of rope, mast, and sturdy sailors at one's command, the bonds must originate in one's own psyche.⁴

---

² Apart from Macintosh and Nelson, below, others making this suggestion include R. Campbell, 'Moral Justification and Freedom,' *J Phil* 85 (1988): 192-213.


MacIntosh claims a person disposed reliably to cooperate, even if it is not value-maximising to do, has a 'mechanical' disposition; Nelson compares the actions of such a person to those of Odysseus 'bound' to the mast. We see here the suggestion that, come the time for cooperating, a constrained maximiser — and so too, presumably, an Agreement-Keeper — cannot but cooperate.

[2] What might justify claims such as these? I will deal with the forms such a justification would take were it to address Agreement-Keeping, rather than constrained maximisation.

A first suggestion is that you, for example, are reliably disposed to keep agreements only if you could not but cooperate were you to believe we had agreed. On this view of the matter, all reliable dispositions are dispositions resulting in unfree acts. Yet this is surely not the case (and, I suspect, a view neither MacIntosh nor Nelson had in mind). My butcher is so disposed that were he to believe I asked him for a pound of ground beef, then he would give me a pound of ground beef. I have asked him for as much and, seeing me, he believes I have asked him for a pound of ground beef. He gives me a pound of ground beef, but this is not to say that he was compelled to do so. Not at all. He was free not to give me a pound of ground beef, but, thankfully, he is disposed not to exercise this freedom. It makes perfect sense to suppose that one has a certain capacity one is disposed not to exercise, and so makes sense to suppose that you have a capacity not to cooperate which you are disposed not to exercise.

A better suggestion (and the one I think MacIntosh and Nelson actually have in mind) is that you are reliably disposed to keep agreements, even if it is not expected-value maximising to do so, only if you could not but cooperate were you to believe we had agreed. On this view of the matter, all reliable dispositions to act in a non-expected-value maximising manner are dispositions resulting in unfree acts. Yet this is still not quite correct. From the fact that I am reading the day's newspapers in the library, rather than doing what I believe it is best for me to do — looking up some references — it does not follow that my action of reading the papers is an unfree one; it only follows (if at all) that it is irrational. Thus, from the fact that I do something non-expected-value maximising it does not follow that my action is unfree. Similarly from the fact I am disposed to doing
something not expected-value maximising - like cooperating - it does not follow that I am disposed to performing unfree actions. It makes perfect sense to suppose one has a certain capacity which one is (if at all) irrationally disposed not to exercise, and so makes sense to suppose that I have a capacity not to cooperate which I am (if at all) irrationally disposed not to exercise.

The Foole might claim that the stipulative addition I made to the Promise Puzzle, (P2), to the Third Counterexample means that neither of us is free to renege on our agreement, but neither of the two possible interpretations of this claim is convincing. More convincing would it be to claim that you are reliably disposed to keep agreements only if you are reliably disposed to value cooperation over non-cooperation whenever you believe we have agreed. On this view of the matter, all reliable dispositions cannot but be dispositions resulting in (free) expected-value maximising actions. But to make this suggestion is just to object that in the situation described by (P3) - where you are actually disposed to cooperating, since you ought to be disposed to cooperating, and are rational - it must no longer have the best outcome for you not to cooperate. And this is the second objection with which I want to deal.

§2 Does it really still have the best outcome for you not to cooperate?

I constructed the Third Counterexample by adding to the Promise Puzzle, (P2), the stipulation that each of us, before we promise, is free to choose how we are disposed, and that this choice is, afterwards, transparent to the other person, to get (P3). I claimed that, in these modified situations, you are rationally permitted to keep your agreement, and cooperate, even though you are free not to do so, and it maximises expected-value for you not to do so. The Foole, however, may now object this stipulative addition means it no longer has the best outcome for you not to cooperate. As I shall argue in this section, however, this second objection is also mistaken, and that in the Third Counterexample it may indeed still have the best outcome not to cooperate.

[1] The problem, again, is that it seems the action of your cooperating is unintelligible, unless we suppose that, somehow, it now has the best
outcome for you to cooperate. There are some philosophers who suggest that a constrained maximiser—who will cooperate if they believe others are likely to cooperate—must be the sort of person now valuing the outcome of cooperation to that of non-cooperation:

Gauthier assumes that two people totally unconcerned with each other's interests can have equally strong commitments to keep an agreement, even though either of them could increase his utility payoff by breaking the agreement. I find this assumption simply incoherent—unless these two people do develop some concern from each other's interest.

The problem is set in the following passage: "[a] constrained maximiser is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and does co-operate in such ways when she may actually expect to benefit" (177). The problem is that Gauthier does not say exactly why she does co-operate (at least not in that context). Certainly she would benefit more by defecting. It seems to me then that the disposition merely determines co-operation, for it simply cannot rationalise it.

Harsanyi's claim implies a coherent notion of a constrained maximiser will be one which, for example, says that the constrained maximiser develops some concern for the other agent. Macintosh insists that if one would benefit more from defection, then there is nothing to rationalise cooperation were it to occur.

[2] What might justify claims such as these? I will deal with the forms such a justification would take were it to address Agreement-Keeping, rather than constrained maximisation. At base, the issue is whether or

---


not the following claim, a similar one to which we met in Chapter Three, is true:

(G3) In the Promise Puzzle, you cooperate intentionally only if the expected-value of doing so is no less than that of not cooperating.

If statement (G3) were true, then my opponent would have established that my counterexample is, indeed, incoherent. It is this statement which lies at the heart of the dispute.8

Statement (G3), though, is not true. Suppose, for example, you are aware of all the relevant outcomes: if you were to cooperate, then the outcome would be that you either forgo the opportunity to exploit the other or relinquish the ability to defend yourself; if not, not. Suppose further you value the outcome of not cooperating to that of cooperating. Nevertheless, seeing the benefits to be obtained by agreement, you agree to cooperate, and so do I. You come to believe, correctly, that we have made an agreement. Suppose finally, you believe (rightly or wrongly, but surely as many believe) that the fact we have an agreement to cooperate is a sufficient reason, in and of itself, to cooperate. Come the time for cooperating, you do so. In this case, (G3) is false. On the one hand, the antecedent is true: you cooperate, and do so intentionally, since (roughly) cooperation is caused, presumably in the right way, by there being a consideration (namely, that we agreed to do so) you took to be the case, and took to be (rightly or wrongly) a sufficient reason for you to cooperate. From your perspective, you had all the reason you needed to cooperate. On the other hand, the consequent is false: the expected-value of cooperating is less than that of not cooperating, since you are completely aware of the fact that the only outcome of cooperating will be the loss of the ability to defend yourself or to exploit the other, and you would rather not suffer this outcome. From your own perspective, however, this reason against cooperating is not of greater strength than the reason you take yourself to have for cooperating – namely, that we agreed. Intentional action, then, is not necessarily action maximising (agent-

relative) expected-value, but is, if anything at all, action which maximises (agent-relative) expected-strength-of-reasons.

Again, my opponent may not be impressed. I claimed you cooperate intentionally, even though this is not expected-value maximising, if the fact we agreed gives you (at least by your own lights) an extra, non-consequentialist, reason to cooperate. My opponent may claim, though, that while this intention does really give you a reason, it is in fact a consequentialist reason.

[3] The suggestion is that if you cooperate, your action manifests the value you place on keeping promises. My opponent's response is as follows: if you are aware of the details of the situation, then you will be aware that the outcome of cooperating is that you fulfil your promise, and the outcome of not cooperating that you do not. However, if you are aware that these are the relevant outcomes, and you cooperate, then you must (implicitly or explicitly) value keeping your promises no less than the sacrifice that may be involved in doing so. Furthermore, in a Promise Puzzle you are aware of the details of the situation. Thus, in a Promise Puzzle, if you cooperate then (a) you must be aware that the outcome of cooperating is the fulfilment of your promise and the outcome of not doing so is the non-fulfilment of promise, and (b) you must (implicitly or explicitly) value fulfilling your promises no less than the sacrifice involved in doing so. However, if (a) and (b) are true, then, by definition, the expected-value of cooperating is in fact no less than that of not cooperating. It follows that, in the Promise Puzzle case, if you cooperate, then the expected-value of doing so must be no less than that of not doing so. Statement (G3) is true.

This objection to my claim that agreements provide extra reasons for action is very similar to the objection to my claim that intentions provide extra reasons for action, but it seems the responses will have to be different. The (not so short) response in the case of intentions was that the relevant objection did not do justice to the causal role intention plays in guiding action. Intention enters into the guidance of

---

action not because its presence becomes just one more fact to be considered in deliberation, but rather because it typically pre-empts any further consideration of actions which are inconsistent with the fulfilment of that intention, even though such actions may be ones the agent is perfectly free to do. But it seems this response is not available in the case of agreements, since the belief that we have made an agreement can, and it seems will always, guide action in virtue of the fact that its presence becomes just one more fact to be considered in deliberation, rather than by pre-empting the consideration of options inconsistent with keeping the agreement. It may be that intention guides action via pre-emption, but unlikely that belief about agreements does.

Can the belief that we have agreed play an intention-like role in the guidance of action? I think the answer to the question is ‘yes’, and to see this, consider the following account – that of Michael Bratman’s – of the functional role of intention. The account requires an internal state to satisfy three functional conditions if it is to count as an intention to A. First, the state in question must have a characteristic stability, in that, unexpected conditions to one side, it will continue to exist even when not causally interacting with other internal states. Second, if the agent is in the state, then they will view the issue of whether or not to A as a closed one, and will be disposed not to engage in further thought about whether to A. Finally, the state plays a causal role in deliberation about how to go about doing A. For example, a person in this state, and believing the only way they can A is if they to B, will typically arrive at an intention to B.

It is clear that your belief that we have agreed can play each of these functional roles. First, it has the required stability, since we have in fact agreed, and you may be expected not to forget this. Second, in some persons, but not necessarily all, the belief they have agreed to do something will render closed the question of whether or not to do it. To be sure, there will be some (untrustworthy) people for whom this will not be true, but these people do not tell against my suggestion that your belief can play this second causal role. Finally, in some persons,

---

though again not necessarily all, a belief that they have agreed to do something will lead to deliberations about how to keep the agreement. For example, your belief that we have agreed to relinquish the use of our arms, and your belief (suppose) that the only reliable way to do this is to turn your swords into ploughshares, leads to an intention to do precisely this. If Bratman is right, then it is a conceptual truth that an intention to A satisfies these three requirements. It may not be a conceptual truth, but rather just an empirical truth about you, that the belief we have agreed also satisfies these three requirements. Such a belief state, then, is playing an intention-like role in guiding your cooperation, and, if it is, then we may suppose, just as with intention, that it guides your cooperation by pre-empting the consideration of options inconsistent with keeping the agreement, rather than by just becoming one more fact to be considered in deliberation about whether to keep the agreement or not.

The Poole might claim that the stipulative addition I made to the Promise Puzzle, (P2), to get the Third Counterexample means that it no longer maximises expected-value for each not to cooperate, but, as we have now seen, such a claim is false. In agreeing to cooperate, each commits themselves to cooperating, but there is no reason to suppose that a commitment to such an action occurs only when each is incapable of doing otherwise, or only when it has the best outcome to perform the action. (Indeed, if only these could provide us with the motive to perform some action, it would be tempting to say we lacked commitment to that action.) For all the Poole has said, in the Third Counterexample it may indeed still have the best outcome not to cooperate.

§3 What reason is there for you to cooperate?

The conclusion of the previous chapter stands: you are rationally permitted to cooperate, even though you are free not to do so, and even though it has the best outcome for you not to do so. This, though, raises a question: if you really are rationally permitted to cooperate, then what reason could you possibly have to do so? I shall argue in this section that in the Third Counterexample, and given some conditions, the fact that we have agreed to cooperate is reason enough for you to cooperate.
[1] I claim that if you become an Agreement-Keeper in the situation described by (P3), then the fact that we have agreed to cooperate is (sufficient) reason for you to cooperate. It is important to note two things about this claim. First, I am making a claim about the reasons you have only in those situations described by assumption (P3), and am not claiming that, in all cases, the fact that one has made an agreement to do something is a sufficient reason to do that thing. Second, even in this restricted class of cases, I claim only that you have reason to cooperate if the specified conditions obtain, and am not claiming you have reason to cooperate, even if we have not agreed, or you have not become an Agreement-Keeper.

Since the Poole is wrong, and the Self-Interest Theory false, there must be a non-consequentialist reason for cooperating; my further speculation, then, concerning this non-consequentialist reason is that it is the fact we have agreed to cooperate. The idea that agreements provide one with extra, non-consequentialist, reasons for action is hardly new, but I want to add my own examples in support of it.

It is important to be clear about the form of my argument. It does not depend on an inference from the rationality, or the fact, of your believing that our agreement is a reason to cooperate to that of its actually being a reason. This would be to invoke just another bridging principle, which Parfit claims (and I agree) is false. Rather, I am arguing that (a) we may use your belief that it is a reason to defeat an objection to my position, allowing me (b) to reassert my claim that you are rationally permitted to cooperate, and thus leading to my speculation (c) that the fact we have agreed is in indeed a reason to cooperate. There are, though, those who would claim there is no reason to cooperate. In the remainder of this chapter, I will consider what one of them has to say.

[2] Gregory Kavka, in a review of Morals by Agreement, comes at one point to consider the reasons a constrained maximiser might have for doing the non-maximising thing, and cooperating. Kavka canvasses a number of possibilities:

---

12 G. Kavka, 'Review of 'Morals by Agreement', ' Mind 96 (1987): 117-121. All page references in this subsection will be to this review.
B correctly calculates that she would maximize long-run expected utility by not carrying out her part. What valid reason could she have to comply rather than maximize? (1) Compliance is morally right. This reason doesn't help Gauthier because it gets things backwards from his perspective. ... (2) Compliance will increase B’s future opportunities for cooperation. By hypothesis this factor is outweighed in B’s calculation of expected utility. (3) If B had not been disposed to comply, she would not have reached a beneficial agreement with A. True, perhaps, but the agreement has been made and A has already complied. The past cannot be changed; at present non-compliance maximizers expected-utility. ... (4) B acts contrary to her rationally adopted disposition if she does not comply. True, but the question is whether this disposition is one to perform rational or irrational acts. Asserting that the rationality of dispositions establishes the rationality of acts in accordance with those dispositions begs the question at issue. ... (5) If B always acted in this way (non-compliance), she would have few opportunities to cooperate. This is simply a combination of reasons (2) and (3) and is correspondingly outweighed, confused, or both. (6) If everyone acted like this, B (or everyone) would lose the benefits of cooperation. This has no relevance to B who is concerned only to promote her own utility, ... Being unable to think of other plausible reasons B might have to comply, I am inclined to doubt the rationality of acts of constrained maximization.13

Kavka is right, it seems to me, to claim that none of (1) to (6) provides sufficient reason for a constrained maximiser to cooperate. Kavka would be right, also, were he to claim that no statements similar to (1) to (6) provide sufficient reason for an Agreement-Keeper to cooperate. Kavka would be wrong, though, to infer from this that there is no reason to cooperate. Having been quoted in length, we can see that he has been thorough in his search for reasons, but, unfortunately, not thorough enough.

Kavka has neglected to consider the possibility that the reason for cooperating is simply that one has agreed to do so. But it may seem that he has a more general argument up his sleeve. He says

in the absence of sufficient forward-looking reasons for complying, he would not find it rational to comply. Unless, that it, the fact that he has in the past committed himself to constrained maximization itself constitutes a reason for compliance. But, as argued in the previous paragraph [when discussing reason

13 Kavka, p. 120
In this passage, Kavka suggests that the only way one could argue it is rational to cooperate is to assume – what I agree it would be question begging to assume – that actions expressing a rational disposition are themselves rational. His point may be well taken against Gauthier, but not against the argument I provided in the previous chapter. I argued that it might be rational to cooperate by assuming instead – and more reasonably – that any adequate theory of rationality will not imply that an agent rationally ought to perform each of a number of actions they cannot perform together. We faced a choice, then, whether to reject the claim that you rationally ought to adopt the cooperative disposition, or reject the claim that you rationally ought not to cooperate, and I argued we should reject the latter. This means that you are rationally permitted to cooperate, and this in turn means that there must be a reason for you to cooperate, which the consequentialist aspect of the Self-Interest Theory omits. I have suggested (entirely plausibly in my view) that this extra reason is the fact that we have agreed to cooperate.

In response to Kavka, then, I do not assume that this is a reason in order to argue that it is rational for you to cooperate; rather, I argue that it is rational for you to cooperate, conclude that there must be some reason or other for you to cooperate, and speculate that the reason is that we have agreed to cooperate. My argument runs in exactly the opposite order to that which Kavka finds objectionable.

In the Third Counterexample, and given some conditions, the fact we have agreed to cooperate is reason enough for one to cooperate. My argument for this claim has been that, in these situations, the deontic principle \((OP_n)\) entails there is some non-consequentialist reason for you to drink the toxin, and I simply make the further speculation that this reason is simply the fact we have agreed. This is just as one would expect. In common parlance, the fact that one has made an agreement to do something is often taken to be a reason, in and of itself, for one to do that thing, a reason independent of any consequentialist considerations. We see that, in some circumstances at least, agreements are indeed reasons for action.

\[14\] Kavka, p. 121
Conclusion

The Poole might claim the Third Counterexample is, in effect, incoherent, but he would be mistaken. We end the second part of the thesis with the conclusions of the previous chapter standing. First, that the Poole's claim that it is rational to keep an agreement only if it maximises expected-value to do so is mistaken. Second, that GIVEN that you rationally ought to adopt the enduring disposition to do what we have agreed to do, THEN if you rationally ought to believe we have agreed to do something, you are rationally permitted to perform your part of the bargain, EVEN IF you are free to do otherwise and it has the best outcome for you to do otherwise (though doing so may very well be irrational absent this condition). It was Nietzsche who said that '[m]an himself must first of all have become calculable, regular, necessary, even in his own image of himself, if he is to be able to stand security for his own future, which is what one who promises does!' If I am right, then reason, at least, need not stand in the way of such security.

RATIONAL DILEMMAS

PART III
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 S₁, rationally ought to S₂, ..., and rationally ought to Sₙ, then it is logically possible that he S₁'s, he S₂'s, ..., and he Sₙ's. (n = 1, 2, 3, ...)

(OC_n) If an agent rationally ought to S₁, rationally ought to S₂, ..., and rationally ought to Sₙ, then he can be such that he S₁'s, he S₂'s, ..., and he Sₙ'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, 'Contrary-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.

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 Blackhall, 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 \((\text{OP}_n)\) and \((\text{OC}_n)\) should be clear:

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

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

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 (OC2) – 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 (OP3) – 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 (OPn) and (OCn) 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 (OPn) and (OCn), 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.
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 \((\text{OP}_n)\) and \((\text{OC}_n)\) 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 \((\text{OP}_n)\) and \((\text{OC}_n)\).

§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 be a *Agreement-Keeper* (an action), rationally ought to be 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 arguments 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. 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. 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 (eg, 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 \( A_0, A_1, \ldots \) 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...
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 agentrationally 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 Makropulos 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 '[t]he 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 milenia 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 ... [l]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. 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 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. 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.

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 \((OP_n)\) and \((OC_n)\) 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.
Chapter Eight

Derek Parfit on Rational Irrationality

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 defense, that 'it is not possible both that I have one of the best possible sets of
motives, in self-interested terms, and that I never do what I believe to be irrational. This is not possible in the relevant sense: it is not possible whatever my desires and dispositions are. If I 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. 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 (OCn) and (OPn) 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

---

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

5 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.
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.
agon. 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.\textsuperscript{10} 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 \textit{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 \textit{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,
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 blameworthy 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 (OP_n) and (OC_n) 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 (OP_n) and (OC_n) 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:

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:

\[
\text{I face a choice between adopting the deterrent intention (I) or not. If a}
\text{certain condition - to be introduced below - obtains, it turns out that I}
\text{rationally ought to adopt the intention to retaliate.}
\]

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

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

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 UDSSs 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.4

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 \( x \).

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.\textsuperscript{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.\textsuperscript{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.

\textsuperscript{7} Gauthier, 'Deterrence, Maximization, and Rationality,' p. 488.

\textsuperscript{8} Kavka, 'The Paradox of Deterrence Revisited,' pp. 45-6.
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:

\begin{itemize}
  \item \textbf{(OP\textsubscript{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. \((n = 1, 2, 3, \ldots)\)
  \item \textbf{(OC\textsubscript{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. \((n = 1, 2, 3, \ldots)\)
\end{itemize}

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 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 (OC₃) or (OP₃). On the one hand, it is an objection to (OC₃) 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 (OC₃). On the other hand, it is an objection to (OP₃) 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 (OP₃). So, as defined, no UDS is an objection to principles (OCₙ) or (OPₙ).

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 (OCₙ) and (OPₙ) – 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 (OCₙ) and (OPₙ), 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 (OCₙ) and (OPₙ) 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 (OCn) is that situations of the following form are logically possible:

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

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 (OC3).
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:

\[(D₃b) \ (x) \ I \ am \ free \ to \ adopt \ the \ enduring \ disposition \ that \ if \ I \ intend \ to \ do \ something, \ and \ am \ aware \ that \ the \ relevant \ conditions \ obtain, \ then \ I \ do \ it, \ and \ the \ expected-value \ of \ adopting \ this \ disposition \ exceeds \ that \ of \ my \ not \ adopting \ it. \ \text{(a')} \ Whether \ or \ not \ I \ adopt \ this \ disposition, \ I \ am \ free \ then \ to \ adopt \ the \ intention \ to \ retaliate; \ the \ expected \ value \ of \ my \ adopting \ the \ intention \ is \ greater \ than \ that \ of \ not \ adopting \ it, \ since \ it \ is \ necessary \ and \ very \ likely \ sufficient \ to \ adopt \ the \ intention \ if \ you \ are \ to \ cooperate, \ (b') \ whether \ or \ not I\]
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 (OPₙ). 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 (OP₄), and so it seems that (OP₄) 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).
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: [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.

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:

\[(I_2) \text{ it is not logically possible that (i) x adopt the intention at } t_1 \text{ that if } p \text{ at } \tau \text{ then he A's at } \tau, \text{ (ii) x not reconsider this intention from } t_1 \text{ to } t_2 \text{ inclusive, (iii) x believes at } t_2 \text{ that } p \text{ obtains, and (iv) x not (try to)}^{14} \text{ A at } t_2.\]

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 – \((I_2)\) – does not falter on this point.\(^{15}\) If this principle is correct, deontic principle \((OP_4)\) entails with \((I_2)\) that

\[(2) \text{ it is not all the case that (i') x rationally ought to adopt the intention at } t_1 \text{ that if } p \text{ at } \tau \text{ then he A's at } \tau, \text{ (ii') x rationally ought not to reconsider this intention from } t_1 \text{ to } t_2 \text{ inclusive, (iii') x rationally ought to believe at } t_2 \text{ that } p \text{ obtains, and (iv') x rationally ought not (try to) A at } t_2.\]

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

\(^{14}\) Without this type of qualification, the principle would be false, since x might not be able to A at \(t_2\). I shall largely ignore this complexity in what follows, and shall assume that were I to try to retaliate, then I would succeed.

\(^{15}\) 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.\(^{16}\) (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.\(^{17}\) 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-consequentalist 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-consequentalist theory under consideration, though, denies this, and so is false.\textsuperscript{18}

In short, the act-consequentalist 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 consequentalist theory of the rationality of unreflective (non) reconsideration will be a rule-consequentalist 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.

\textsuperscript{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].
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 nth 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.\textsuperscript{19}

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).

§4 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.

\textsuperscript{19} 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.
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.22 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-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
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
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:

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

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:

(OC2) If an agent rationally ought to $S_1$, and rationally ought to $S_2$, then he can be such that he $S_1$'s and he $S_2$'s.

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 (OCn) (and (OPn)) 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 in a way that deserves to be blamed, and not that it is impossible for someone to avoid acting irrationally. 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 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 Reasons and
Persons, of the words 'can' and 'cannot'. Here is what he has to say on this topic:

\[
(P1) \text{In the doctrine that ought implies can, the sense of can is compatible with}
\]

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

\[
(P2) \text{It is not possible both that I have one of the best possible set of motives}
\]

[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):

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

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 the 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 \textit{problematic} reference to Kate's desires and dispositions.

Suppose some theory tells Kate that (a') she ought \textit{to cause herself to have} the desire that her books be as good as possible, and tells her, in addition, that (b') she ought \textit{not to have} this desire. It might be theory S, since this theory in effect tells Kate \textit{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 \textit{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:

\begin{quote}
\begin{itemize}
\item[(P2')] It is not possible \textit{both} that I cause myself to have the desire that my books be as good as possible, \textit{and} that I not have this desire. This is not possible in the relevant sense: it is not possible \textit{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.
\end{itemize}
\end{quote}

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

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
REFERENCES

Brink, D. O., 'Utilitarian Morality and the Personal Point of View,' *J Phil* 83 (1986): 417-438
Castenada, H.-N., 'A Problem for Utilitarianism,' *Analysis* 28 (1968): 231-4
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., 'Against Moral Dilemmas,' *Phil Rev* 91 (1982): 87-97
Conee, E., 'Why Moral Dilemmas are Impossible,' *Amer Phil Quart* 26 (1989): 133-141


Darwell, S., 'Rational Agent, Rational Act,' Phil Topics 14 (1986): 33-57


Dworkin, R., 'Nuclear Intentions', Ethics 95 (1985): 445-460


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


REFERENCES

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,
REFERENCES


Harman, G., ‘Reasons,’ *Critica* 7 (1975): 3-17


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
REFERENCES


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,
REFERENCES

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 "Protagorus" 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)


REFERENCES

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
van Fraassen, B., 'Values and the Heart's Command,' *J Phil* 70 (1973): 5-19
Vorobej, M., 'Gauthier on Deterrence,' *Dialogue* 25 (1986): 471-6