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
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
Introduction

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. 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.

INTRODUCTION

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
INTRODUCTION

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


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

---

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?'