Reasons and Persons

Pages: 560
Publisher: OUP Oxford (January 23, 1986)
Format: pdf, epub
Language: English

REASONS AND PERSONS At last the horizon appears free to us again, evengranted that it is not bright; at last our ships may venture out again, venture out to face any danger; all the daring of the lover of knowledge is permitted again; the sea, *our sea,* lies open again; perhaps there has never yet been such an ‘Open sea’. NIETZSCHE, p. 448 **REASONS AND PERSONS BYDEREK PARFIT CLARENDON PRESS • OXFORD**
OXFORDUNIVERSITY PRESS Great Clarendon Street, Oxford OX2 6DPOxford University Press is a department of the University of Oxford.It furthers the University’s objective of excellence in research, scholarship,and education by publishing worldwide inOxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi Kuala LumpurMadrid Melbourne Mexico City Nairobi New Delhi Shanghai Taipei TorontoWith offices inArgentina Austria Brazil Chile Czech Republic France GreeceGuatemala Hungary Italy Japan Poland Portugal SingaporeSouth Korea Switzerland Thailand Turkey Ukraine VietnamOxford is a registered trade mark of Oxford University Pressin the UK and in certain other countries
Published in the United Statesby Oxford University Press Inc., New York
First published 1984First issued in paperback (with corrections) 1986Reprinted with further corrections 1987

British Library Cataloguing in Publication DataData availableLibrary of Congress Cataloging in Publication DataParfit, Derek.Reasons and persons.Includes bibliographical references and index.
1. Ethics 2. Rationalism. 3. Self. I. Title.BJ1012.P39 1984 170 83-15139ISBN 978-0-19-824908-5 (pbk)
Printed in Great Britainon acid-free paper byCPI Antony Rowe, Chippenham, Wiltshire ***To my parentsDrs. Jessie and Norman Parfitand my sistersTheodora and Joanna***
ACKNOWLEDGEMENTS SIXTEEN years ago, I travelled to Madrid with Gareth Evans. I hoped to become a philosopher, and as we drove through France I put to him my ideas. His criticisms made me despair. But before we reached Spain, I saw that he was almost as critical of his own ideas. Like many others, I owe much to the intensity of his love of truth, and his extraordinary vitality. I record this debt first because he died when he was 34. I owe a great deal to my first teachers: Sir Peter Strawson, Sir Alfred Ayer, David Pears, and Richard Hare. I have since learnt from many people. In discussion I have learnt most from Thomas Nagel, Ronald Dworkin, Tim Scanlon, Amartya Sen, Jonathan Glover, James Griffin, Ann Davis, Jefferson McMahan, and Donald Regan. I have learnt much more from reading the writings of these and many other people. Some of my debts I

acknowledge in the endnotes to this book. But I am certain that, because of my weak memory and failure to make proper notes, this book presents, as if they were my ideas, many claims or arguments that I ought to attribute to some source. These forgotten sources, if they read this book, will be rightly aggrieved. Though they should be mentioned in the endnotes, I hope that most are at least mentioned in the Bibliography. Several people helped me to write this book. Before he died two years ago, John Mackie wrote extremely helpful comments on my earlier work. In the last few months I have received many comments on a draft of this book—so many that I have not had time to make all the needed revisions. Here is a randomly ordered list of those who have helped me in this way: Jonathan Glover, Sir Peter Strawson, John McDowell, Susan Hurley, Paul Seabright, John Vickers, Hywel Lewis, Judith Thomson, Samuel Scheffler, Martin Hollis, Thomas Nagel, Robert Nozick, Richard Lindley, Gilbert Harman, Christopher Peacocke, Peter Railton, Annette Baier, Kurt Baier, Richard Swinburne, Michael Tooley, Mark Sainsbury, Wayne Sumner, Jim Stone, Dale Jamieson, Eric Rakowski, James Griffin, Gregory Kavka, Thomas Hurka, Geoffrey Madell, Ralph Walker, Bradford Hooker, Douglas Maclean, Graeme Forbes, Bimal Matilal, Nicholas Dent, Robert Goodin, Andrew Brennan, John Kenyon, James Fishkin, Robert Elliott, Arnold Levison, Simon Blackburn, Ronald Dworkin, Amartya Sen, Peter Unger, Peter Singer, Jennifer Whiting, Michael Smith, David Lyons, Milton Wachsberg, William Ewald, Galen Strawson, Gordon Cornwall, Richard Sikora, Partha Dasgupta, Dr. Jessie Parfit, and Dr. Charles Whitty. I learnt something from everyone just named, and from some I learnt a great deal. From a few people I learnt so much that I want to thank them separately. Jonathan Bennett sent me very helpful comments on half of my draft. Bernard Williams sent me extremely helpful comments on a draft of Part Three. Six other people sent me very helpful comments on drafts of the whole book. Four of these were John Leslie, Michael Woodford, Larry Temkin, and Donald Regan. From two other people I learnt even more. John Broome was a Visiting Fellow at my College throughout the academic year at the end of which I write these words. Both in written comments, and in very many discussions, he solved very many of my problems, and suggested many great improvements. If every passage due to John Broome was mentioned in the endnotes, there would be at least thirty of these notes. As the different academic disciplines drift away from their neighbours, it is heartening to find that an economist should be, in his spare time, so good a philosopher. The person from whom I have learnt the most is Shelly Kagan. Kagan’s extraordinarily acute and penetrating comments were half as long as my draft, and many of his suggestions are printed, with little change, in this book. If his co-authorship was mentioned in the endnotes, there would be at least sixty of these notes. I write these words the day before this text goes to the printer. Because I received so many good objections or comments, I could not have revised and produced my text on time without help of other kinds. I have been helped by Patricia Morison, and greatly helped by Susan Hurley and William Ewald. Jefferson and Sally McMahan saved me many days of work in sorting papers, checking references, and compiling the Bibliography. This book is printed from camera-ready copy. Given my slowness in making the needed revisions, the four people who produced this copy have often, uncomplainingly, worked overtime, and late into the night. These generous people are Angela Blackburn, Jane Nunns, Paul Salotti, and, most generous of all, Catherine Griffin. I am grateful for the help of everyone mentioned above. To those mentioned in the last two paragraphs I record here my extreme gratitude. This book has one author, but is really the joint product of all of these people. Finally, I record my great gratitude to an entity that is not a person: All Souls College. If I had not had the extraordinary privilege of being a Prize Fellow and then a Research Fellow of this College for the past sixteen years, this book would certainly not exist. *All Souls College, Oxford12 September 1983. D. A. P.*

INTRODUCTION Like my cat, I often simply do what I want to do. I am then not using an ability that only persons have. We know that there are reasons for acting, and that some reasons are better or stronger than others. One of the main subjects of this book is a set of questions about what we have reason to do. I shall discuss several theories. Some of these are moral theories, others are theories about rationality. We are particular people. I have my life to live, you have yours. What do these facts involve? What makes me the same person throughout my life, and a different person from you? And what is the importance of these facts? What is the importance of the unity of each life, and of the distinction between

different lives, and different persons? These questions are the other main subject of this book. My two subjects, reasons and persons, have close connections. I believe that most of us have false beliefs about our own nature, and our identity over time, and that, when we see the truth, we ought to change some of our beliefs about what we have reason to do. We ought to revise our moral theories, and our beliefs about rationality. In the first two parts of the book I give other arguments for similar conclusions. I shall not describe, in advance, these arguments and conclusions. The List of Contents provides a summary. The book is long, and sometimes complicated. I have therefore separated my arguments into 154 parts, and given each part a descriptive title. I hope that this makes the arguments easier to follow, and shows what the book contains more clearly than an Index of Subjects could. If I had not rearranged the arguments into these separate parts, such an Index would have been too thick with references to be of much use. Many introductions to books of this kind try to explain the central concepts that are used. Since it would take at least a book to give a helpful explanation, I shall waste no time in doing less than this. My central concepts are few. We have *reasons for acting.* We *ought* to act in certain ways, and some ways of acting are *morally wrong.* Some outcomes are *good* or *bad,* in a sense that has moral relevance: it is bad for example if people become paralyzed, and we ought, if we can, to prevent this. Most of us understand my last three sentences well enough to understand my arguments. I shall also use the concept of what is in someone’s *self-interest,* or what would be *best for this person.* I discuss this briefly in Appendix I. My last central concept is that of a *person.* Most of us think we understand what persons are. Part Three claims that we do not. Many introductions also try to explain how, when discussing morality, we can hope to make progress. Since the best explanation would be provided by *making* progress, this is the only explanation I shall try to give. Strawson describes two kinds of philosophy, descriptive, and revisionary. Descriptive philosophy gives reasons for what we instinctively assume, and explains and justifies the unchanging central core in our beliefs about ourselves, and the world we inhabit. I have great respect for descriptive philosophy. But, by temperament, I am a revisionist. Except in my dreary Chapter 1, where I cannot avoid repeating what has been shown to be true, I try to challenge what we assume. Philosophers should not only interpret our beliefs; when they are false, they should *change* them. *Note added in 1985:* In this reprinting I have made several corrections. Thus I have withdrawn my support for the Wide Psychological Criterion of personal identity (p. 208 and elsewhere), since this conflicts with my view that we should not try to decide between the different criteria. I have replaced ‘bad’ with ‘worse’ in Claim Q (p. 360) and several later sentences. Other minor substantive corrections include those made in Notes 15 to part I and 83 to part III, and pages 66/lines 15-17, 67/19-23, 211/19-25, 312/31-2, and 374/1-2 and 12-13. For some of these corrections I am grateful to the contributors to *Ethics,* July 1986. *Note added in 1987:* In this reprinting I have widened my definition of Reductionism (p. 210), and removed an apparent circularity from the Physical and Psychological Criteria (pp. 204 and 207). Other minor substantive corrections come on pages 72/lines *2-4*, 73/5-15, 81/17-24, 82/5-11, 311/40, and Note 59 to Part I. (I have also made several stylistic corrections or revisions.) **CONTENTS PART ONE • SELF-DEFEATING THEORIES CHAPTER 1 • THEORIES THAT ARE INDIRECTLY SELF-DEFEATING**

1 The Self-interest Theory
2 How S Can Be Indirectly Self-defeating
3 Does S Tell Us to Be Never Self-denying?
4 Why S Does Not Fail in Its Own Terms
5 Could It Be Rational to Cause Oneself to Act Irrationally?
6 How S Implies that We Cannot Avoid Acting Irrationally
7 An Argument for Rejecting S When It Conflicts with Morality
8 Why This Argument Fails
9 How S Might Be Self-Effacing
10 How Consequentialism Is Indirectly Self-defeating
11 Why C Does Not Fail in Its Own Terms
12 The Ethics of Fantasy

13 Collective Consequentialism
14 Blameless Wrongdoing
15 Could It Be Impossible to Avoid Acting Wrongly?
16 Could It Be Right to Cause Oneself to Act Wrongly?
17 How C Might Be Self-Effacing
18 The Objection that Assumes Inflexibility
19 Can Being Rational or Moral Be a Mere Means?
20 Conclusions **CHAPTER 2 • PRACTICAL DILEMMAS**
21 Why C Cannot Be Directly Self-defeating
22 How Theories Can Be Directly Self-defeating
23 Prisoner’s Dilemmas and Public Goods
24 The Practical Problem and its Solutions **CHAPTER 3 • FIVE MISTAKES IN MORAL MATHEMATICS**
25 The Share-of-the-Total View
26 Ignoring the Effects of Sets of Acts
27 Ignoring Small Chances
28 Ignoring Small or Imperceptible Effects
29 Can There Be Imperceptible Harms and Benefits?
30 Overdetermination
31 Rational Altruism **CHAPTER 4 • THEORIES THAT ARE DIRECTLY SELF-DEFEATING**
32 In Prisoner’s Dilemmas, Does S Fail in Its Own Terms?
33 Another Weak Defence of Morality
34 Intertemporal Dilemmas
35 A Weak Defence of S
36 How Common-Sense Morality Is Directly Self-Defeating
37 The Five Parts of a Moral Theory
38 How We Can Revise Common-Sense Morality so that It Would Not Be Self-Defeating
39 Why We Ought to Revise Common-Sense Morality
40 A Simpler Revision **CHAPTER 5 • CONCLUSIONS**
41 Reducing the Distance between M and C
42 Towards a Unified Theory
43 Work to be Done
44 Another Possibility **PART TWO • RATIONALITY AND TIME** **CHAPTER 6 • THE BEST OBJECTION TO THE SELF-INTEREST THEORY**
45 The Present-aim Theory
46 Can Desires Be Intrinsically Irrational, or Rationally Required?
47 Three Competing Theories
48 Psychological Egoism
49 The Self-interest Theory and Morality
50 My First Argument
51 The S-Theorist’s First Reply
52 Why Temporal Neutrality Is Not the Issue Between S and P **CHAPTER 7 • THE APPEAL TO FULL RELATIVITY**
53 The S-Theorist’s Second Reply
54 Sidgwick’s Suggestions
55 How S Is Incompletely Relative
56 How Sidgwick Went Astray
57 The Appeal Applied at a Formal Level
58 The Appeal Applied to Other Claims **CHAPTER 8 • DIFFERENT ATTITUDES TO TIME**
59 Is It Irrational to Give No Weight to One’s Past Desires?
60 Desires that Depend on Value Judgements or Ideals
61 Mere Past Desires
62 Is It Irrational To Care Less About One’s Further Future?

63 A Suicidal Argument
64 Past or Future Suffering
65 The Direction of Causation
66 Temporal Neutrality
67 Why We Should Not Be Biased towards the Future
68 Time’s Passage
69 An Asymmetry
70 Conclusions **CHAPTER 9 • WHY WE SHOULD REJECT S**
71 The Appeal to Later Regrets
72 Why a Defeat for Proximus is Not a Victory for S
73 The Appeal to Inconsistency
74 Conclusions **PART THREE • PERSONAL IDENTITY** **CHAPTER 10 • WHAT WE BELIEVE OURSELVES TO BE**
75 Simple Teletransportation and the Branch-Line Case
76 Qualitative and Numerical Identity
77 The Physical Criterion of Personal Identity
78 The Psychological Criterion
79 The Other Views **CHAPTER 11 • HOW WE ARE NOT WHAT WE BELIEVE**
80 Does Psychological Continuity Presuppose Personal Identity?
81 The Subject of Experiences
82 How a Non-Reductionist View Might Have Been True
83 Williams’s Argument against the Psychological Criterion
84 The Psychological Spectrum
85 The Physical Spectrum
86 The Combined Spectrum **CHAPTER 12 • WHY OUR IDENTITY IS NOT WHAT MATTERS**
87 Divided Minds
88 What Explains the Unity of Consciousness?
89 What Happens When I Divide?
90 What Matters When I Divide?
91 Why There Is No Criterion of Identity that Can Meet Two Plausible Requirements
92 Wittgenstein and Buddha
93 Am I Essentially My Brain?
94 Is the True View Believable? **CHAPTER 13 • WHAT DOES MATTER**
95 Liberation From the Self
96 The Continuity of the Body
97 The Branch-Line Case
98 Series-Persons
99 Am I a Token or a Type?
100 Partial Survival
101 Successive Selves **CHAPTER 14 • PERSONAL IDENTITY AND RATIONALITY**
102 The Extreme Claim
103 A Better Argument against the Self-interest Theory
104 The S-Theorist’s Counter-Argument
105 The Defeat of the Classical Self-interest Theory
106 The Immorality of Imprudence **CHAPTER 15 • PERSONAL IDENTITY AND MORALITY**
107 Autonomy and Paternalism
108 The Two Ends of Lives
109 Desert
110 Commitments
111 The Separateness of Persons and Distributive Justice
112 Three Explanations of the Utilitarian View
113 Changing a Principle’s Scope
114 Changing a Principle’s Weight

115 Can It Be Right to Burden Someone Merely to Benefit Someone Else?
116 An Argument for Giving Less Weight to Equality
117 A More Extreme Argument
118 Conclusions **PART FOUR • FUTURE GENERATIONS** **CHAPTER 16 • THE NON-IDENTITY PROBLEM**
119 How Our Identity in Fact Depends on When We Were Conceived
120 The Three Kinds of Choice
121 What Weight Should We Give to the Interests of Future People?
122 A Young Girl’s Child
123 How Lowering the Quality of Life Might Be Worse for No One
124 Why an Appeal to Rights Cannot Solve the Problem
125 Does the Fact of Non-Identity Make a Moral Difference?
126 Causing Predictable Catastrophes in the Further Future
127 Conclusions **CHAPTER 17 • THE REPUGNANT CONCLUSION**
128 Is It Better If More People Live?
129 The Effects of Population Growth on Existing People
130 Overpopulation
131 The Repugnant Conclusion **CHAPTER 18 • THE ABSURD CONCLUSION**
132 An Alleged Asymmetry
133 Why the Ideal Contractual Method Provides No Solution
134 The Narrow Person-Affecting Principle
135 Why We Cannot Appeal to this Principle
136 The Two Wide Person-Affecting Principles
137 Possible Theories
138 The Sum of Suffering
139 The Appeal to the Valueless Level
140 The Lexical View
141 Conclusions **CHAPTER 19 • THE MERE ADDITION PARADOX**
142 Mere Addition
143 Why We Should Reject the Average Principle
144 Why We Should Reject the Appeal to Inequality
145 The First Version of the Paradox
146 Why We Are Not Yet Forced to Accept the Repugnant Conclusion
147 The Appeal to the Bad Level
148 The Second Version of the Paradox
149 The Third Version **CONCLUDING CHAPTER**
150 Impersonality
151 Different Kinds of Argument
152 Should We Welcome or Regret My Conclusions?
153 Moral Scepticism
154 How both Human History, and the History of Ethics, May Be Just Beginning **APPENDICES**
A A World Without Deception
B How My Weaker Conclusion Would in Practice Defeat S
C Rationality and the Different Theories about Self-interest
D Nagel’s Brain
E The Closest Continuer Schema
F The Social Discount Rate
G Whether Causing Someone to Exist can Benefit this Person
H Rawlsian Principles
I What Makes Someone’s Life Go Best
J Buddha’s View
Notes
Bibliography

Index of Names **PART ONESELF-DEFEATING THEORIES 1THEORIES THAT ARE INDIRECTLY SELF-DEFEATING** WHAT do we have most reason to do? Several theories answer this question. Some of these are moral theories; others are theories about rationality. When applied to some of our decisions, different theories give us different answers. We must then try to decide which is the best theory. Arguments about these theories are of many kinds. One argument is that a theory is *self-defeating.* This argument, uniquely, needs no assumptions. It claims that a theory fails even in its own terms, and thus condemns itself. The first part of this book discusses what this argument achieves. As I shall explain, all of the best known theories are in certain ways self-defeating. What does this show? In some cases, nothing. In other cases, what is shown is that a theory must be developed further, or extended. And in other cases what is shown is that a theory must be either rejected or revised. This is what is shown about the moral theories that most of us accept. I start with the best-known case. **1. THE SELF-INTEREST THEORY** We can describe all theories by saying what they tell us to try to achieve. According to all moral theories, we ought to try to act morally. According to all theories about rationality, we ought to try to act rationally. Call these our *formal* aims. Different moral theories, and different theories about rationality, give us different *substantive* aims. By ‘aim’, I shall mean ‘substantive aim’. This use of aim is broad. It can describe moral theories that are concerned, not with moral goals, but with rights, or duties. Suppose that, on some theory, five kinds of act are totally forbidden. This theory gives to each of us the aim that he never acts in these five ways. I shall first discuss the *Self-interest Theory,* or *S.* This is a theory about rationality. S gives to each person this aim: the outcomes that would be best for himself, and that would make his life go, for him, as well as possible. To apply S, we must ask what would best achieve this aim. Answers to this question I call *theories about self-interest.* As Appendix I explains, there are three plausible theories. On the *Hedonistic Theory,* what would be best for someone is what would give him most happiness. Different versions of this theory make different claims about what happiness involves, and how it should be measured. On the *Desire-Fulfilment Theory,* what would be best for someone is what would best fulfil his desires throughout his life. Here again, there are different versions of this theory. Thus the *Success Theory* appeals only to a person’s desires about his own life. On the *Objective List Theory,* certain things are good or bad for us, even if we would not want to have the good things or avoid the bad things. Here again, there are different versions. The good things might include the development of one’s abilities, knowledge, and the awareness of true beauty. The bad things might include sadistic pleasure, being deceived, and losing liberty, or dignity. These three theories partly overlap. On all these theories, happiness and pleasure are at least part of what makes our lives go better for us, and misery and pain are at least part of what makes our lives go worse. These claims would be made by any plausible Objective List Theory. And they are implied by all versions of the Desire-Fulfilment Theory. On all theories, the Hedonistic Theory is at least part of the truth. To save words, this will sometimes be the only part that I discuss. All these theories also claim that, in deciding what would be best for someone, we should give equal weight to all the parts of this person’s future. Later events may be less predictable; and a predictable event should count for less if it is less likely to happen. But it should not count for less merely because, if it happens, it will happen later. It would take at least a book to decide between the different theories about self-interest. This book discusses some of the differences between these theories, but does not try to decide between them. Much of this book discusses the Self-interest Theory. As I have said, this is not one of the theories about self-interest. It is a theory about rationality. We can discuss S without deciding between the different theories about self-interest. We can make claims that would be true on all of these theories. It will help to call some aims *ultimate.* Other aims are *instrumental,* mere means to the achievement of some ultimate aim. Thus, for all except misers, being rich is not an ultimate aim. I can now re-state the *central claim* of S. This is

(S1) For each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible. As we shall see, S makes several other claims. There are several objections to S. Some of these I discuss in Parts Two and Three. In what follows I discuss the objection that, like certain other theories, S is self-defeating. **2. HOW S CAN BE INDIRECTLY SELF-DEFEATING** If we call some theory *T*, call the aims that it gives us *our T-given aims.* Call T

indirectly individually self-defeating when it is true that, if someone tries to achieve his T-given aims, these aims will be, on the whole, worse achieved. On this definition, we do not simply ask whether a theory is self-defeating. We ask whether it is self-defeating when applied to certain people, during certain periods. My S-given aim is that my life go, for me, as well as possible. It can be true that, if I try to do whatever will be best for me, this will be worse for me. There are two kinds of case:

(a) My attempts may often fail. I may often do what will be worse for me than something else that I could have done.

(b) Even if I never do what, of the acts that are possible for me, will be worse for me, it may be worse for me if I am purely self-interested. It might be better for me if I had some other disposition. In cases of kind *(a),* the bad effects come from what I do. Suppose that I steal whenever I believe that I will not be caught. I may be often caught, and punished. Even in self-interested terms, honesty may therefore be the best policy for me. These cases are not worth discussing. If this is the way in which S is self-defeating, this is no objection to S. S is self-defeating here only because of my incompetence in attempting to follow S. This is a fault, not in S, but in me. We might object to some theory that it is too difficult to follow. But this is not true of S. The cases worth discussing are of kind *(b).* In these cases it will be worse for me if I am purely self-interested, even if I succeed in never doing what will be worse for me. The bad effects come, not from what I do, but from my disposition, or the fact that I am purely self-interested. What does this fact involve? I could be purely self-interested without being purely selfish. Suppose that I love my family and friends. On all of the theories about self-interest, my love for these people affects what is in my interests. Much of my happiness comes from knowing about, and helping to cause, the happiness of those I love. On the Desire-Fulfilment Theory, it will be better for me if, as I want, things go well for those I love. What will be best for me may, in these and other ways, largely overlap with what will be best for those I love. But, in some cases, what will be better for me will be worse for those I love. I am self-interested if, in all these cases, I do what will be better for me. It may be thought that, if I am self-interested, I shall *always* be trying to do whatever will be best for me. But I often act in one of two ways, believing that neither would be better for me. In these cases I am not trying to do what will be best for me; I am acting on a more particular desire. And this may be true even when I am doing what I know will be best for me. Suppose that I know that, if I help you, this will be best for me. I may help you because I love you, not because I want to do what will be best for me. In describing what it would be for me to be self-interested, it is enough to claim that, while I often act on other desires, I *never do what I believe will be worse for me.* If this is true, it will be clearer to call me, not *self-interested,* but *never self denying.* I shall now redescribe the interesting way in which, for any individual, S may be indirectly self-defeating. This is true when, if someone is never self-denying, this will be worse for him than it would be if he had some other disposition. Even if someone succeeds in never doing what would be worse for him, it may be worse for him that he is never self-denying. It might be better for him if he had some other disposition. He might then sometimes do what would be worse for him. But the costs to him of acting in this way might be less than the benefits of having this other disposition. These claims can be true on all of the different theories about self-interest. Hedonists have long known that happiness, when aimed at, is harder to achieve. If my strongest desire is that I be happy, I may be less happy than I would be if I had other desires that were stronger. Thus I might be happier if my strongest desire was that someone else be happy. Here is another example. *Kate* is a writer. Her strongest desire is that her books be as good as possible. Because she cares so much about the quality of her books, she finds her work rewarding. If her desire to write good books was much weaker, she would find her work boring. She knows this, and she accepts the Hedonistic Theory about self-interest. She therefore believes that it is better for her that her strongest desire is that her books be as good as possible. But, because of the strength of this desire, she often works so hard, and for so long, that she collapses with exhaustion, and is, for a period, very depressed. Suppose that Kate believes truly that, if she worked less hard, her books would be slightly worse, but she would be happier. She would find her work just as rewarding, and she would avoid these severe depressions. Kate therefore believes that, when she works so hard, she is doing what is

worse for her. But how could it become true that she never acts in this way? It may be a fact that she would never act in this way only if she had a much weaker desire that her books be as good as possible. And this would be even worse for her, since she would then find her work boring. On the Hedonistic Theory, it would be worse for Kate if she was never self-denying.1 Suppose that we accept not the Hedonistic but the Desire-Fulfilment Theory about self-interest. We may then deny that, in this example, Kate is doing what is worse for her. Her strongest desire is that her books be as good as possible. By working so hard, though she makes herself exhausted and depressed, she makes her books slightly better. She is thereby causing her strongest desire to be better fulfilled. On our theory about self-interest, this may be better for her. If we are not Hedonists, we need a different example. Suppose that I am driving at midnight through some desert. My car breaks down. You are a stranger, and the only other driver near. I manage to stop you, and I offer you a great reward if you rescue me. I cannot reward you now, but I promise to do so when we reach my home. Suppose next that I am *transparent,* unable to deceive others. I cannot lie convincingly. Either a blush, or my tone of voice, always gives me away. Suppose, finally, that I know myself to be never self-denying. If you drive me to my home, it would be worse for me if I gave you the promised reward. Since I know that I never do what will be worse for me, I know that I shall break my promise. Given my inability to lie convincingly, you know this too. You do not believe my promise, and therefore leave me stranded in the desert. This happens to me because I am never self-denying. It would have been better for me if I had been *trust-worthy,* disposed to keep my promises even when doing so would be worse for me. You would then have rescued me. (It may be objected that, even if I am never self-denying, I could decide to keep my promise, since making this decision would be better for me. If I decided to keep my promise, you would trust me, and would rescue me. This objection can be answered. I know that, after you have driven me home, it would be worse for me if I gave you the promised reward. If I know that I am never self-denying, I know that I shall not keep my promise. And, if I know this, I cannot decide to keep my promise. I cannot decide to do what I know that I shall not do. If I *can* decide to keep my promise, this must be because I believe that I shall not be never self-denying. We can add the assumption that I would not believe this unless it was true. It would then be true that it would be worse for me if I was, and would remain, never self-denying. It would be better for me if I was trustworthy.) I have described two ways in which it would be worse for someone if he was never self-denying. There are many other ways in which this can be true. It is probably true of most people, during most of their lives. When the Self-interest Theory is applied to these people, it is what I call indirectly individually self-defeating. Does this make S fail in its own terms? Does S condemn itself? This depends on whether S tells these people to be never self-denying. **3. DOES S TELL US TO BE NEVER SELF-DENYING?** It may seem obvious that S tells everyone to be never self-denying. But, as described so far, S claims only that, for each person, there is one supremely rational ultimate aim: that his life go, for him, as well as possible. When applied to acts, S claims both

(S2) What each of us has2 most reason to do is whatever would be best for himself, and

(S3) It is irrational for anyone to do what he believes will be worse for himself. S must also make claims about what we should do when we cannot predict the effects of our acts. We can ignore cases of *uncertainty,* where we have no beliefs about the probabilities of different effects. In *risky* cases, where we do have such beliefs, S claims

(S4) What it would be rational for anyone to do is what will bring him the greatest *expected* benefit. To calculate the expected benefit from some act, we add together the possible benefits, and subtract the possible costs, with each benefit or cost multiplied by the chance that the act will produce it. Thus, if some act has a chance of nine in ten of bringing me some benefit B, and a chance of one in ten of causing me to lose some benefit that would be twice as great as B, the expected benefit is B × 9/10—2B × 1/10, or seven-tenths of B. What should S claim about the rationality of desires and dispositions? Since S claims that, for each person, there is one supremely rational ultimate aim, S should clearly claim that the supremely rational desire is the desire that this aim be achieved. S should claim

(S5) The supremely rational desire is that one’s life go as well as possible for oneself.

Similarly, S should claim

(S6) The supremely rational disposition is that of someone who is never self-denying. If someone is never self-denying, though he sometimes acts on other desires, he never acts against the supremely rational desire. He never does what he believes will be worse for him. To save words, call both desires and dispositions *motives.* There are ways in which, over time, we can cause our motives to change. We can develop habits. If we act in ways that we do not now enjoy, we may come to enjoy them. If we change our work, or where we live, or read certain books, or have children, this may cause predictable changes in our motives. And there are many other ways in which we can cause such changes. According to (S2), what each person has most reason to do is to cause himself to have, or to allow himself to keep, any of the *best possible sets of motives, in self-interested terms.* These are the sets of motives of which the following is true. There is no other possible set of motives of which it is true that, if this person had these motives, this would be better for him. By ‘possible’ I mean ‘causally possible, given the general facts about human nature, and the particular facts about this person’s nature’. It would often be hard to know whether some set of motives would be causally possible for someone, or would be one of the best sets for this person in self-interested terms. But we can ignore these difficulties. There are many cases in which someone knows that it would be better for him if there was some change in his motives. And in many of these cases such a person knows that, in one of the ways described above, he could cause this change. (S3) implies that it would be irrational for this person not to cause this change. Similar claims apply to our emotions, beliefs, abilities, the colour of our hair, where we live, and anything else that we could change. What each of us has most reason to do is to make any change that would be best for himself. If someone believes that he could make such a change, it would be irrational for him not to do so. We can now return to my earlier question. We are discussing the people of whom it is true that, if they were never self-denying, this would be worse for them than if they had some other disposition. Suppose that these people know that this is true. Does S tell them to be never self-denying? S claims the following. If such a person was never self-denying, he would have the disposition that is supremely rational. But it would be irrational for this person to cause himself to have, or to keep, this disposition. It would be rational for him to cause himself to have, or to keep, the other disposition, since this would be better for him. These claims may seem to give conflicting answers to my question. They may seem to tell this person both to be, and not to be, never self-denying. This misinterprets S. When S claims that one disposition is supremely rational, it does not tell us to *have* this disposition. Remember the distinction between formal and substantive aims. Like all theories about rationality, S gives to everyone this formal aim: to be rational, and to act rationally. What distinguishes different theories is that they give us different substantive aims. In its central claim, (S1), S gives to each person one substantive aim: that his life go, for him, as well as possible. Does S give to each person *another* substantive aim: to be rational, and to act rationally? It does not. According to S, our formal aim is not a substantive aim. It may be thought that, in making these claims, I have not described the best version of the Self-interest Theory. But this is the version that would be accepted by most of those who believe this theory. Most of these people would accept (S2) and (S3). Suppose I know that it will be best for me if I make myself irrational. I shall soon describe a case in which this would almost certainly be true. If this is true, (S2) implies that what I have most reason to do is to make myself irrational. (S3) implies that it would be irrational for me *not* to do so. These claims do not give me, as a substantive aim, being rational. Does this imply that, for S, being rational is a mere means? This depends on what is the best theory about self-interest. On the Hedonistic Theory, S gives to each person this substantive aim: the greatest possible happiness for himself. Being rational is not an essential part of *this* aim. It is a mere means. So is acting rationally, and having rational desires or dispositions. Consider next the Objective List Theory. On some versions of this theory, being rational is one of the things that is good for each person, whatever its effects may be. If this is so, being rational is not a mere means, but part of the substantive aim that S gives to each person. The same would be true, on the Desire-Fulfilment Theory, in the case of those people who want to be rational, whatever the effects may be. It may be objected: ‘Suppose that we accept the Hedonistic Theory. S then tells us that being rational is a mere means. If this is so, why should we try to be rational? Why should we

want to know what we have most reason to do? If we accept what S claims, and believe that being rational is a mere means, we shall cease to be interested in the questions that S claims to answer. This must be an objection to S. An acceptable theory about rationality cannot claim that being rational is a mere means.’ We could answer: ‘A theory would be unacceptable if it claimed that being rational did not matter. But this is not what S claims. Suppose that I cling to some rock as a mere means of escaping death. Though my act is a mere means, it matters a great deal. The same can be true about being rational.’ This may not completely answer this objection. As we shall see, there is a similar objection to certain moral theories. To save words, I discuss these objections at the same time. This discussion is in Section 19. I can now explain a remark that I made above. According to S, the disposition that is supremely rational is that of someone who is never self-denying. I wrote that, in making this claim, S does not tell us to have this disposition. S gives to each person one substantive aim: that his life go, for him, as well as possible. On some theories about self-interest, being rational would, for some people, be part of this aim. But this would only be because, like being happy, being rational is part of what makes our lives go better. Being rational is not, *as such,* a substantive aim. Nor is having the supremely rational disposition. In the case of some people, according to S, being rational would *not* be part of what makes their lives go better. These are the people that I am discussing. It is true of these people that, if they were never self-denying, this would be worse for them than if they had some other disposition. Since this is true, being never self-denying would *not* be part of the aim that S gives to these people. S does not tell these people to have what S claims to be the supremely rational disposition: that of someone who is never self-denying. And, if they can change their disposition, S tells these people, if they can, *not* to be never self-denying. Since it would be better for these people if they had some other disposition, S tells them to cause themselves to have, or to keep, this other disposition. If they know that they could act in either of these ways, S claims that it would be irrational for them not to do so. It would be irrational for them to cause themselves to be, or to allow themselves to remain, never self-denying. **4. WHY S DOES NOT FAIL IN ITS OWN TERMS** These claims answer the other question that I asked. When S is applied to these people, it is what I call indirectly self-defeating. Does this make S fail in its own terms? Does S condemn itself? The answer is No. S is indirectly self-defeating because it would be worse for these people if they were never self-denying. But S does *not* tell these people to be never self-denying, and it tells them, if they can, *not* to be. If these people are never self-denying, this is worse for them. This is a bad effect, in S’s terms. But this bad effect is not the result either of their doing what S tells them to do, or of their having a disposition that S tells them to have. Since this is so, S is not failing in its own terms. It may be objected: ‘This bad effect may be the result of these people’s *belief in* S. If they believe S, they believe that it would be irrational for them to do what they believe will be worse for them. It may be true that, if they believe that it is irrational to act in this way, they will never do so. If they never act in this way, they are never self-denying. Suppose that, in one of the ways that you described, having this disposition is worse for them. This is a bad effect in S’s terms. If belief in S has this effect, S does fail in its own terms.’ In answering this objection, we need to know whether these people can change their disposition. Suppose, first, that they cannot. Why would this be true? If they cannot change their disposition, and they have this disposition *because* they believe S, the explanation must be that they cannot cause themselves to be disposed to do what they believe to be irrational. They could change their disposition only if they believed some other theory about rationality. S would then tell them, if they can, to make themselves believe this other theory. This possibility I discuss in Sections 6 to 8. As I shall argue, even if this is true, S would not be failing in its own terms. Suppose next that these people can change their disposition, without changing their beliefs about rationality. If these people are never self-denying, this will be worse for them than it would be if they had some other disposition. S tells these people to cause themselves to have this other disposition. The objection given above clearly fails. It may be true that these people are never self-denying because they believe S. But S claims that it is irrational for these people to allow themselves to remain never self-denying. If they do remain never self-denying, this cannot be claimed to be merely ‘the result of their belief in S’. It is the result of their failure to do what they could do, and what

S tells them to do. This result is worse for them, which is a bad effect in S’s terms. But a bad effect which results from *disobeying* S cannot provide an objection to S. If my doctor tells me to move to a healthier climate, he would be open to no criticism if, because I refuse to move, I die.
There is a third possibility. These people may be unable to change either their dispositions, or their beliefs about rationality. Their belief in S is bad for them, which is a bad effect in S’s terms. Is this an objection to S? It will be easier to answer this question after I have discussed other theories. My answer is in Section 18. **5. COULD IT BE RATIONAL TO CAUSE ONESELF TO ACT IRRATIONALLY?** I turn now to a new question. A theory may be unacceptable even though it does not fail in its own terms. It is true of many people that it would be worse for them if they were never self-denying. Does this give us independent grounds to reject S? According to S, it would be rational for each of these people to cause himself to have, or to keep, one of the best possible sets of motives, in self-interested terms. Which these sets are is, in part, a factual question. And the details of the answer would be different for different people in different circumstances. But we know the following, about each of these people. Since it would be worse for him if he was never self-denying, it would be better for him if he was sometimes self-denying. It would be better for him if he was sometimes disposed to do what he believes will be worse for him. S claims that acting in this way is irrational. If such a person believes S, it tells him to cause himself to be disposed to act in a way that S claims to be irrational. Is this a damaging implication? Does it give us any reason to reject S? Consider
Schelling’s Answer to Armed Robbery. A man breaks into my house. He hears me calling the police. But, since the nearest town is far away, the police cannot arrive in less then fifteen minutes. The man orders me to open the safe in which I hoard my gold. He threatens that, unless he gets the gold in the next five minutes, he will start shooting my children, one by one.
What is it rational for me to do? I need the answer fast. I realize that it would not be rational to give this man the gold. The man knows that, if he simply takes the gold, either I or my children could tell the police the make and number of the car in which he drives away. So there is a great risk that, if he gets the gold, he will kill me and my children before he drives away.
Since it would be irrational to give this man the gold, should I ignore his threat? This would also be irrational. There is a great risk that he will kill one of my children, to make me believe his threat that, unless he gets the gold, he will kill my other children.
What should I do? It is very likely that, whether or not I give this man the gold, he will kill us all. I am in a desperate position. Fortunately, I remember reading Schelling’s *The Strategy of Conflict.*[3] I also have a special drug, conveniently at hand. This drug causes one to be, for a brief period, very irrational. Before the man can stop me, I reach for the bottle and drink. Within a few seconds, it becomes apparent that I am crazy. Reeling about the room, I say to the man: ‘Go ahead. I love my children. So please kill them.’ The man tries to get the gold by torturing me. I cry out: ‘This is agony. So please go on.’
Given the state that I am in, the man is now powerless. He can do nothing that will induce me to open the safe. Threats and torture cannot force concessions from someone who is so irrational. The man can only flee, hoping to escape the police. And, since I am in this state, he is less likely to believe that I would record the number of his car. He therefore has less reason to kill me.
While I am in this state, I shall act in irrational ways. There is a risk that, before the police arrive, I may harm myself or my children. But, since I have no gun, this risk is small. And making myself irrational is the best way to reduce the great risk that this man will kill us all. On any plausible theory about rationality, it would be rational for me, in this case, to cause myself to become for a period irrational.4a This answers the question that I asked above. S might tell us to cause ourselves to be disposed to act in ways that S claims to be irrational. This is no objection to S. As the case just given shows, an acceptable theory about rationality *can* tell us to cause ourselves to do what, in its own terms, is irrational. Consider next a general claim that is sometimes made:
(GI) If there is some motive that it would be both (a) rational for someone to cause himself to have, and (b) irrational for him to cause himself to lose, then (c) it cannot be irrational for this person to act upon this motive. In the case just described, while this man is still in my house, it would be irrational for me to cause myself to cease to be irrational. During this period, I have a set of

motives of which both (a) and (b) are true. But (c) is false. During this period, my acts are irrational. We should therefore reject (G1). We can claim instead that, since it was rational for me to cause myself to be like this, this is a case of *rational* irrationality. **6. HOW S IMPLIES THAT WE CANNOT AVOID ACTING IRRATIONALLY** Remember Kate, who accepts the Hedonistic Theory about self-interest. We may accept some other theory. But on these other theories there could be cases that, in the relevant respects, are like Kate’s. And the claims that follow could be restated to cover these cases. It is best for Kate that her strongest desire is that her books be as good as possible. But, because this is true, she often works very hard, making herself, for a period, exhausted and depressed. Because Kate is a Hedonist, she believes that, when she acts in this way, she is doing what is worse for her. Because she also accepts S, Kate believes that, in these cases, she is acting irrationally. Moreover, these irrational acts are quite voluntary. She acts as she does because, though she cares about her own interests, this is not her strongest desire. She has an even stronger desire that her books be as good as possible. It would be worse for her if this desire became weaker. She is acting on a set of motives that, according to S, it would be irrational for her to cause herself to lose. It might be claimed that, because Kate is acting on such motives, she cannot be acting irrationally. But this claim assumes (Gl), the claim that was shown to be false by the case I called Schelling’s Answer to Armed Robbery. If we share Kate’s belief that she is acting irrationally, in a quite voluntary way, we might claim that *she* is irrational. But Kate can deny this. Since she believes S, she can claim: ‘When I do what I believe will be worse for me, my *act* is irrational. But, because I am acting on a set of motives that it would be irrational for me to cause myself to lose, I am *not* irrational. More precisely, I am *rationally irrational.’* She can add: ‘In acting on my desire to make my books better, I am doing what will be worse for me. This is a bad effect, in self-interested terms. But it is part of a set of effects that is one of the best possible sets. Though I sometimes suffer, because this is my strongest desire, I also benefit. And the benefits are greater than the losses. That I sometimes act irrationally, doing what I know will be worse for me, is the price I have to pay if I am to get these greater benefits. This is a price worth paying.’ It may be objected: ‘You do not *have* to pay this price. You *could* work less hard. You could do what would be better for you. You are not compelled to do what you believe to be irrational.’ She could answer: ‘This is true. I *could* work less hard. But I only *would* do this if my desire to make my books better was much weaker. And this would be, on the whole, worse for me. It would make my work boring. How could I bring it about that I shall not in the future freely choose, in such cases, to do what I believe to be irrational? I could bring this about only by changing my desires in a way that would be worse for me. This is the sense in which I cannot have the greater benefits without paying the lesser price. I cannot have the desires that are best for me without sometimes freely choosing to act in ways that will be worse for me. This is why, when I act irrationally in these ways, I need not regard *myself* as irrational.’ This reply assumes one view about voluntary acts: *Psychological Determinism.* On this view, our acts are always caused by our desires, beliefs, and other dispositions. Given our actual desires and dispositions, it is not causally possible that we act differently. It may be objected: ‘If it is not causally possible that Kate act differently, she should not believe that, to act rationally, she *ought* to act differently. We only *ought* to do what we *can* do.’ A similar objection will arise later when I discuss what we ought morally to do. It will save words if Kate answers both objections. She can say: ‘In the doctrine that *ought* implies *can,* the sense of ‘can’ is compatible with Psychological Determinism. When my act is irrational or wrong, I ought to have acted in some other way. On the doctrine, I ought to have acted in this other way only if I could have done so. If I could *not* have acted in this other way, it cannot be claimed that this is what I ought to have done. The claim (1) that I could not have acted in this other way is not the claim (2) that acting in this way would have been impossible, given my actual desires and dispositions. The claim is rather (3) that acting in this way would have been impossible, even if my desires and dispositions had been different. Acting in this way would have been impossible, whatever my desires and dispositions might have been. If claim (1) was claim (2), Determinists would have to conclude that it is not possible for anyone ever to act wrongly or irrationally. They can justifiably reject this conclusion. They can insist that claim (1) is claim (3).’ Kate could add: ‘I

am not claiming that *Free Will* is compatible with Determinism. The sense of ‘can’ required for Free Will may be different from the sense of ‘can’ in the doctrine that ought implies can. These senses are held to be different by most of those Determinists who believe that Free Will is *not* compatible with Determinism. This is why, though these Determinists do not believe that anyone deserves punishment, they continue to believe that it is possible to act wrongly or irrationally.’ Kate may be wrong to assume Psychological Determinism. I claimed earlier that our beliefs about rationality may affect our acts, because we may want to act rationally. It may objected:

This misdescribes how these beliefs affect our acts. We do not *explain* why someone has acted rationally by citing his desire to do so. Whenever someone acts rationally, it may be trivially true that he wanted to do so. But he acted as he did because he had a belief, not a belief *and* a desire. He acted as he did simply because he believed that he had a reason to do so. And it is often causally possible for him to act rationally whatever his desires and dispositions are.[4] Note that this objector cannot claim that it is *always* possible for someone to act rationally, whatever his desires and dispositions are. Even if he denies Determinism, this objector cannot claim that there is *no* connection between our acts and our dispositions. This objector must also admit that our desires and dispositions may make it *harder* for us to do what we believe to be rational. Suppose that I am suffering from intense thirst, and am given a glass of iced water. And suppose I believe that I have a reason to drink this water slowly, since this would increase my enjoyment. I also have a reason not to spill this water. It is much easier to act upon this second reason than it is, given my intense thirst, to drink this water slowly. If the objector’s claims are true, Kate’s reply must be revised. She might say: ‘It would be worse for me if my strongest desire was to avoid doing what I believe to be irrational. It is better for me that my strongest desire is that my books be as good as possible. Since this is my strongest desire, I sometimes do what I believe to be irrational. I act in this way because my desire to make my books better is much stronger than my desire not to act irrationally. You claim that I could often avoid acting in this way. By an act of will, I could often avoid doing what I most want to do. If I could avoid acting in this way, I cannot claim that I am in no sense irrational. But, given the strength of my desire to make my books better, it would be *very hard* for me to avoid acting in this way. And it would be irrational for me to change my desires so that it would be easier for me to avoid acting in this way. Given these facts, I am irrational only in a very weak sense.’ Kate might add: ‘It is not possible *both* that I have one of the best possible sets of motives, in self-interested terms, *and* that I never do what I believe to be irrational. This is not possible in the relevant sense: it is not possible *whatever* my desires and dispositions are. If I was never self-denying, my ordinary acts would never be irrational. But I would have acted irrationally in causing myself to become, or allowing myself to remain, never self-denying. If instead I cause myself to have one of the best possible sets of motives, I shall sometimes do what I believe to be irrational. If I do not have the *disposition* of someone who is never self-denying, it is not possible that I *always act* like someone with this disposition. Since this is not possible, and it would be irrational for me to cause myself to be never self-denying, I cannot be criticised for sometimes doing what I believe to be irrational.’ It may now be said that, as described by Kate, S lacks one of the essential features of any theory. It may be objected: ‘No theory can demand what is impossible. Since Kate cannot always avoid doing what S claims to be irrational, she cannot always do what S claims that she ought to do. We should therefore reject S. As before, *ought* implies *can.’* Even if we deny Determinism, this objection still applies. As I have claimed, we must admit that, since Kate does not have the disposition of someone who is never self-denying, she cannot *always* act like such a person. Is it a good objection to S that Kate cannot always avoid doing what S claims to be irrational? Remember Schelling’s Answer to Armed Robbery. In this case, on any plausible theory about rationality, it would be irrational for me not to make myself very irrational. But, if I do make myself very irrational, I cannot avoid acting irrationally. On both alternatives, at least one of my acts would be irrational. It is therefore true that, in this case, I cannot avoid acting irrationally. Since there can be such cases, an acceptable theory can imply that we cannot avoid acting irrationally. It is no objection to S that it has this implication. We may believe that these claims do

not fully answer this objection. A similar objection will be raised later against certain moral theories. To save words, I discuss these objections together, in Section 15. I shall now summarize my other conclusions. In the case of many and perhaps most people, the Self-interest Theory is indirectly self-defeating. It is true, of each of these people, that it would be worse for him if he was never self-denying—disposed never to do what he believes would be worse for him. It would be better for him if he had some other set of motives. I have claimed that such cases do not provide an objection to S. Since S does not tell these people to be never self-denying, and tells them, if they can, not to be, S is not failing in its own terms. Nor do these cases provide an independent objection to S. Though they do not refute S, for those who accept S these cases are of great importance. In these cases S must cover, not just ordinary acts, but also the acts that bring about changes in our motives. According to S, it would be rational to cause ourselves to have, or to keep, one of the best possible sets of motives, in self-interested terms. If we believe that we could act in either of these ways, it would be irrational not to do so. In the case of most people, any of the best possible sets would cause these people sometimes to do, in a quite voluntary way, what they know will be worse for them. If these people believe S, they will believe that these acts are irrational. But they need not believe *themselves* to be irrational. This is because, according to S, it would be irrational for them to change their motives so that they would cease to act irrationally in this way. They will in part regret the *consequences* of these irrational acts. But the *irrationality* of these acts they can regard with complacency. This is *rational* irrationality. It may be objected, to these claims, that they falsely assume Psychological Determinism. It may sometimes be possible for these people to do what they believe to be rational, whatever their desires and dispositions are. If this objection is correct, these claims need to be revised. When these people do what they believe to be irrational, they cannot claim that they are in no sense irrational. But they can claim that, given their actual motives, it would be very hard for them to avoid acting in this way. And it would be irrational for them, on their theory, to change their motives so that it would be easier to avoid acting in this way. They can therefore claim that they are irrational only in a very weak sense. Having explained once how these claims could be revised, I shall not mention this objection whenever, in what follows, it would be relevant. It would be easy to make the needed revisions to any similar claims. **7. AN ARGUMENT FOR REJECTING S WHEN IT CONFLICTS WITH MORALITY** It has been argued that the Self-interest Theory might tell us to believe, not itself, but some other theory. This is clearly possible. According to S, it would be rational for each of us to cause himself to believe some other theory, if this would be better for him. I have already mentioned one way in which this might be true. It might not be possible for us to do what we believe to be irrational. S would then tell us, in the cases I have been discussing, to try to believe a different theory. There are also other ways in which this might be true. Let us return, for an example, to the keeping of our promises. One kind of mutual agreement has great practical importance. In these agreements, each person in some group makes a conditional promise. Each person promises to act in a certain way, provided that all the others promise to act in certain ways. It can be true both (1) that it will be better for each of these people if all rather than none of them keep their promises, and (2) that, whatever the others do, it will be worse for each person if he himself keeps his promise. What each person loses if he keeps his promise is less than what he gains if all the others keep their promises. This is how (1) and (2) are both true. Such agreements are *mutually advantageous, though requiring self-denial.* If I am known to be never self-denying, I shall be excluded from such agreements. Others will know that I cannot be trusted to keep my promise. It has been claimed that, since this is true, it would be better for me if I ceased to be never self-denying and became trustworthy.[5] This claim overlooks one possibility. It may be best for me if I *appear* to be trustworthy but remain really never self-denying. Since I appear to be trustworthy, others will admit me to these mutually advantageous agreements. Because I am really never self-denying, I shall get the benefits of breaking my promises whenever this would be better for me. Since it is better for me to appear trustworthy, it will often be better for me to keep my promise so as to preserve this appearance. But there will be some promises that I can break secretly. And my gain from breaking some promises may outweigh my loss in ceasing to appear trustworthy. Suppose, however, that I am transparent, unable to lie convincingly. This is true of many people. And it might become more

widely true if we develop cheap and accurate lie-detector tests. Let us assume that this has happened, so that we are all transparent—unable to deceive others. Since we are to some degree transparent, my conclusions may apply to our actual situation. But it will simplify the argument to assume that all direct deception has become impossible. It is worth seeing what such an argument might show. We should therefore help the argument, by granting this assumption. If we were all transparent, it would be better for each of us if he became trustworthy: reliably disposed to keep his promises, even when he believes that doing so would be worse for him. It would therefore be rational, according to S, for each of us to make himself trustworthy. Assume next that, to become trustworthy, we would have to change our beliefs about rationality. We would have to make ourselves believe that it is rational for each of us to keep his promises, even when he knows that this would be worse for him. I shall later describe two ways in which this assumption might be true. It is hard to change our beliefs when our reason for doing so is merely that this change will be in our interests. We would have to use some form of self-deception. Suppose, for example, that I learn that I am fatally ill. Since I want to believe that I am healthy, I pay a hypnotist to give me this belief. I could not keep this belief if I remembered how I had acquired it. If I remembered this, I would know that the belief was false. The same would be true of our beliefs about rationality. If we pay hypnotists to change these beliefs, because this will be better for us, the hypnotists must make us forget why we have our new beliefs. On the assumptions made above, S would tell us to change our beliefs. S would tell us to believe, not itself, but a revised form of S. On this revised theory, it is irrational for each of us to do what he believes will be worse for himself, *except when he is keeping a promise.* If S told us to believe this revised theory, would this be an objection to S? Would it show that it *is* rational to keep such promises? We must focus clearly on this question. We may be right to believe that it is rational to keep our promises, even when we know that this will be worse for us. I am asking, ‘Would this belief be supported if S itself told us to cause ourselves to have this belief?’ Some people answer Yes. They argue that, if S tells us to make ourselves have this belief, this shows that this belief is justified. And they apply this argument to many other kinds of act which, like keeping promises, they believe to be morally required. If this argument succeeded, it would have great importance. It would show that, in many kinds of case, it is rational to act morally, even when we believe that this will be worse for us. Moral reasons would be shown to be stronger than the reasons provided by self-interest. Many writers have tried, unsuccessfully, to justify this conclusion. If this conclusion could be justified in the way just mentioned, this would solve what Sidgwick called ‘the profoundest problem of Ethics’.6 **8. WHY THIS ARGUMENT FAILS** There is a simple objection to this argument. The argument appeals to the fact that S would tell us to make ourselves believe that it is rational to keep our promises, even when we know that this will be worse for us. Call this belief *B.* B is incompatible with S, since S claims that it is irrational to keep such promises. Either S is the true theory about rationality, or it is not. If S is true, B must be false, since it is incompatible with S. If S is not true, B might be true, but S cannot support B, since a theory that is not true cannot support any conclusion. In brief: if S is true, B must be false, and if S is not true, it cannot support B. B is either false, or not supported. So, even if S tells us to try to believe B, this fact cannot support B. We may think that a theory about rationality cannot be true, but can at most be the best, or the best justified theory. The objection just given could be restated in these terms. There are two possibilities. If S is the best theory, we should reject B, since it is incompatible with S. If S is not the best theory, we should reject S. B cannot be supported by a theory that we should reject. Neither of these possibilities gives any support to B.7 This objection seems to me strong. But I know some people whom it does not convince. I shall therefore give two more objections. These will also support some wider conclusions. I shall first distinguish threats from warnings. When I say that I shall do *X* unless you do *Y,* call this a *warning* if my doing X would be worse for you but not for me, and a *threat* if my doing X would be worse for both of us. Call me a *threat-fulfiller* if I would always fulfil my threats. Suppose that, apart from being a threat-fulfiller, someone is never self-denying. Such a person would fulfil his threats even though he knows that this would be worse for him. But he would not *make* threats if he believed that doing so would be worse for him. This is because, apart from being a threat-fulfiller, this person is never self-denying. He never does what he

believes will be worse for him, *except when he is fulfilling some threat.* This exception does not cover *making* threats. Suppose that we are all both transparent and never self-denying. If this was true, it would be better for me if I made myself a threat-fulfiller, and then announced to everyone else this change in my dispositions. Since I am transparent, everyone would believe my threats. And believed threats have many uses. Some of my threats could be defensive, intended to protect me from aggression by others. I might confine myself to defensive threats. But it would be tempting to use my known disposition in other ways. Suppose that the benefits of some co-operation are shared between us. And suppose that, without my co-operation, there would be no further benefits. I might say that, unless I get the largest share, I shall not co-operate. If others know me to be a threat-fulfiller, and they are never self-denying, they will give me the largest share. Failure to do so would be worse for them. Other threat-fulfillers might act in worse ways. They could reduce us to slavery. They could threaten that, unless we become their slaves, they will bring about our mutual destruction. We would know that these people would fulfil their threats. We would therefore know that we can avoid destruction only by becoming their slaves. The answer to threat-fulfillers, if we are all transparent, is to become a *threat-ignorer.* Such a person always ignores threats, even when he knows that doing so will be worse for him. A threat-fulfiller would not threaten a transparent threat-ignorer. He would know that, if he did, his threat would be ignored, and he would fulfil this threat, which would be worse for him. If we were all both transparent and never self-denying, what changes in our dispositions would be better for each of us? I answer this question in Appendix A, since parts of the answer are not relevant to the question I am now discussing. What is relevant is this. If we were all transparent, it would probably be better for each of us if he became a trustworthy threat-ignorer. These two changes would involve certain risks; but these would be heavily outweighed by the probable benefits. What would be the benefits from becoming trustworthy? That we would not be excluded from those mutually advantageous agreements that require self-denial. What would be the benefits from becoming threat-ignorers? That we would avoid becoming the slaves of threat-fulfillers. We can next assume that we could not become trustworthy threat-ignorers unless we changed our beliefs about rationality. Those who are trustworthy keep their promises even when they know that this will be worse for them. We can assume that we could not become disposed to act in this way unless we believed that it *is* rational to keep such promises. And we can assume that, unless we were known to have this belief, others would not trust us to keep such promises. On these assumptions, S tells us to make ourselves have this belief. Similar remarks apply to becoming threat-ignorers. We can assume that we could not become threat-ignorers unless we believed that it is always rational to ignore threats. And we can assume that, unless we have this belief, others would not be convinced that we are threat-ignorers. On these assumptions, S tells us to make ourselves have this belief. These conclusions can be combined. S tells us to make ourselves believe that it is always irrational to do what we believe will be worse for us, *except when we are keeping promises or ignoring threats.* Does this fact support these beliefs? According to S, it would be rational for each of us to make himself believe that it is rational to ignore threats, even when he knows that this will be worse for him. Does this show this belief to be correct? Does it show that it *is* rational ignore such threats? It will help to have an example. Consider

My Slavery. You and I share a desert island. We are both transparent, and never self-denying. You now bring about one change in your dispositions, becoming a threat-fulfiller. And you have a bomb that could blow the island up. By regularly threatening to explode this bomb, you force me to toil on your behalf. The only limit on your power is that you must leave my life worth living. If my life became worse than that, it would cease to be better for me to give in to your threats. How can I end my slavery? It would be no good killing you, since your bomb will automatically explode unless you regularly dial some secret number. But suppose that I could make myself transparently a threat-ignorer. Foolishly, you have not threatened that you would ignore this change in my dispositions. So this change would end my slavery. Would it be rational for me to make this change? There is the risk that you might make some new threat. But since doing so would be clearly worse for you, this risk would be small. And, by taking this small risk, I would almost certainly gain a very great benefit. I would almost certainly end my slavery. Given the

wretchedness of my slavery, it would be rational for me, according to S, to cause myself to become a threat-ignorer. And, given our other assumptions, it would be rational for me to cause myself to believe that it is always rational to ignore threats. Though I cannot be wholly certain that this will be better for me, the great and nearly certain benefit would outweigh the small risk. (In the same way, it would never be wholly certain that it would be better for someone if he became trustworthy. Here too, all that could be true is that the probable benefits outweigh the risks.) Assume that I have now made these changes. I have become transparently a threat-ignorer, and have made myself believe that it is always rational to ignore threats. According to S, it was rational for me to cause myself to have this belief. Does this show this belief to be correct? Let us continue the story.

How I End My Slavery. We both have bad luck. For a moment, you forget that I have become a threat-ignorer. To gain some trivial end—such as the coconut that I have just picked—you repeat your standard threat. You say, that, unless I give you the coconut, you will blow us both to pieces. I know that, if I refuse, this will certainly be worse for me. I know that you are reliably a threat-fulfiller, who will carry out your threats even when you know that this will be worse for you. But, like you, I do not now believe in the pure Self-interest Theory. I now believe that it is rational to ignore threats, even when I know that this will be worse for me. I act on my belief. As I foresaw, you blow us both up. Is my act rational? It is not. As before, we might concede that, since I am acting on a belief that it was rational for me to acquire, I am not irrational. More precisely, I am *rationally* irrational. But what I am doing is not rational. It is irrational to ignore some threat when I know that, if I do, this will be disastrous for me and better for no one. S told me here that it was rational to make myself believe that it is rational to ignore threats, even when I know that this will be worse for me. But this does not show this belief to be correct. It does not show that, in such a case, it *is* rational to ignore threats. We can draw a wider conclusion. This case shows that we should reject

(G2) If it is rational for someone to make himself believe that it is rational for him to act in some way, it *is* rational for him to act in this way. Return now to B, the belief that it is rational to keep our promises even when we know that this will be worse for us. On the assumptions made above, S implies that it is rational for us to make ourselves believe B. Some people claim that this fact supports B, showing that it *is* rational to keep such promises. But this claim seems to assume (G2), which we have just rejected. There is another objection to what these people claim. Even though S tells us to try to believe B, S implies that B is false. So, if B is true, S must be false. Since these people believe B, they should believe that S is false. Their claim would then assume

(G3) If some false theory about rationality tells us to make ourselves have a particular belief, this shows this belief to be true. But we should obviously reject (G3). If some false theory told us to make ourselves believe that the Earth was flat, this would not show this to be so. S told us to try to believe that it is rational to ignore threats, even when we know that this will be worse for us. As my example shows, this does not support this belief. We should therefore make the same claim about keeping promises. There may be *other* grounds for believing that it is rational to keep our promises, even when we know that doing so will be worse for us. But this would not be shown to be rational by the fact that the Self-interest Theory itself told us to make ourselves believe that it was rational. It has been argued that, by appealing to such facts, we can solve an ancient problem: we can show that, when it conflicts with self-interest, morality provides the stronger reasons for acting. This argument fails. The most that it might show is something less. In a world where we are all transparent—unable to deceive each other—it might be rational to deceive ourselves about rationality.[8] **9. HOW S MIGHT BE SELF-EFFACING** If S told us to believe some other theory, this would not support this other theory. But would it be an objection to S? Once again, S would not be failing in its own terms. S is a theory about practical not theoretical rationality. S may tell us to make ourselves have false beliefs. If it would be better for us to have false beliefs, having true beliefs, even about rationality, would not be part of the ultimate aim given to us by S. The arguments given above might be strengthened and extended. This would be easier if, as I supposed, the technology of lie-detection made us all wholly transparent. If we could never deceive each other, there might be an argument that showed that, according to S, it would

be rational for everyone to cause himself not to believe S. Suppose that this was true. Suppose that S told everyone to cause himself to believe some other theory. S would then be *self-effacing.* If we all believed S, but could also change our beliefs, S would remove itself from the scene. It would become a theory that no one believed. But to be self-effacing is not to be self-defeating. It is not the aim of a theory to be believed. If we personify theories, and pretend that they have aims, the aim of a theory is not to be believed, but to be true, or to be the best theory. That a theory is self-effacing does not show that it is not the best theory. S would be self-effacing when, if we believed S, this would be worse for us. But S need not tell us to believe itself. When it would be better for us if we believed some other theory, S would tell us to try to believe this theory. If we succeeded in doing what S told us to do, this would again be better for us. Though S would remove itself from the scene, causing no one to believe itself, it would still not be failing in its own terms. It would still be true that, because each of us has followed S—done what S told him to do—each has thereby made the outcome better for himself. Though S would not be failing in its own terms, it might be claimed that an acceptable theory cannot be self-effacing. I deny this claim. It may seem plausible for what, when examined, is a bad reason. It would be natural to *want* the best theory about rationality not to be self-effacing. If the best theory was self-effacing, telling us to believe some other theory, the truth about rationality would be depressingly convoluted. It is natural to hope that the truth is simpler: that the best theory would tell us to believe itself. But can this be more than a hope? Can we assume that the truth *must* be simpler? We cannot. **10. HOW CONSEQUENTIALISM IS INDIRECTLY SELF-DEFEATING** Most of my claims could, with little change, cover one group of moral theories. These are the different versions of *Consequentialism,* or C. C’s *central claim* is

(C1) There is one ultimate moral aim: that outcomes be as good as possible. C applies to everything. Applied to acts, C claims both

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

(C3) If someone does what he believes will make the outcome worse, he is acting wrongly. I distinguished between what we have most reason to do, and what it would be rational for us to do, given what we believe, or ought to believe. We must now distinguish between what is *objectively* and *subjectively* right or wrong. This distinction has nothing to do with whether moral theories can be objectively true. The distinction is between what some theory implies, given (i) what are or would have been the effects of what some person does or could have done, and (ii) what this person believes, or ought to believe, about these effects. It may help to mention a similar distinction. The medical treatment that is objectively right is the one that would in fact be best for the patient. The treatment that is subjectively right is the one that, given the medical evidence, it would be most rational for the doctor to prescribe. As this example shows, what it would be best to know is what is objectively right. The central part of a moral theory answers this question. We need an account of subjective rightness for two reasons. We often do not know what the effects of our acts would be. And we ought to be blamed for doing what is subjectively wrong. We ought to be blamed for such acts even if they are objectively right. A doctor should be blamed for doing what was very likely to kill his patient, even if his act in fact saves this patient’s life. In most of what follows, I shall use *right, ought, good,* and *bad* in the objective sense. But *wrong* will usually mean *subjectively* wrong, or *blameworthy.* Which sense I mean will often be obvious given the context. Thus it is clear that, of the claims given above, (C2) is about what we ought objectively to do, and (C3) is about what is subjectively wrong. To cover risky cases, C claims

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

of lives saved would be greater. Consequentialism covers, not just acts and outcomes, but also desires, dispositions, beliefs, emotions, the colour of our eyes, the climate, and everything else. More exactly, C covers anything that could make outcomes better or worse. According to C, the best possible climate is the one that would make outcomes best. I shall again use ‘motives’ to cover both desires and dispositions. C claims
(C5) The best possible motives are those of which it is true that, if we have them, the outcome will be best. As before, ‘possible’ means ‘causally possible’. And there would be many different sets of motives that would be in this sense best: there would be no other possible set of motives of which it would be true that, if we had this set, the outcome would be better. I have described some of the ways in which we can change our motives. (C2) implies that we ought to try to cause ourselves to have, or to keep, any of the best possible sets of motives. More generally, we ought to change both ourselves, and anything else, in any way that would make the outcome better. If we believe that we could make such a change, (C3) implies that failing to do so would be wrong.9 To apply C, we must ask what makes outcomes better or worse. The simplest answer is given by *Utilitarianism.* This theory combines C with the following claim: the best outcome is the one that gives to people the greatest net sum of benefits minus burdens, or, on the Hedonistic version of this claim, the greatest net sum of happiness minus misery. There are many other versions of C. These can be *pluralist* theories, appealing to several different principles about what makes outcomes better or worse. Thus, one version of C appeals both to the Utilitarian claim and to the Principle of Equality. This principle claims that it is bad if, through no fault of theirs, some people are worse off than others. On this version of C, the goodness of an outcome depends both on how great the net sum of benefits would be, and on how equally the benefits and burdens would be distributed between different people. One of two outcomes might be better, though it involved a smaller sum of benefits, because these benefits would be shared more equally. A Consequentialist could appeal to many other principles. According to three such principles, it is bad if people are deceived, coerced, and betrayed. And some of these principles may essentially refer to past events. Two such principles appeal to past entitlements, and to just deserts. The Principle of Equality may claim that people should receive equal shares, not at particular times, but in the whole of their lives. If it makes this claim, this principle essentially refers to past events. If our moral theory contains such principles, we are not concerned only with *consequences* in the narrow sense: with what happens *after* we act. But we can still be, in a wider sense, Consequentialists. In this wider sense our ultimate moral aim is, not that outcomes be as good as possible, but that history go as well as possible. What I say below could be restated in these terms. With the word ‘Consequentialism’, and the letter ‘C’, I shall refer to all these different theories. As with the different theories about self-interest, it would take at least a book to decide between these different versions of C. This book does not discuss this decision. I discuss only what these different versions have in common. My arguments and conclusions would apply to all, or nearly all, the plausible theories of this kind. It is worth emphasizing that, if a Consequentialist appeals to all of the principles I have mentioned, his moral theory is *very* different from Utilitarianism. Since such theories have seldom been discussed, this is easy to forget. Some have thought that, if Consequentialism appeals to many different principles, it ceases to be a distinctive theory, since it can be made to cover all moral theories. This is a mistake. C appeals only to principles about what makes outcomes better or worse. Thus C might claim that it would be worse if there was more deception or coercion. C would then give to all of us two common aims. We should try to cause it to be true that there is less deception or coercion. Since C gives to all agents common moral aims, I shall call C *agent-neutral.* Many moral theories do not take this form. These theories are *agent-relative,* giving to different agents different aims. It can be claimed, for example, that each of us should have the aim that *he* does not coerce other people. On this view, it would be wrong for me to coerce other people, even if by doing so I could cause it to be true that there would be less coercion. Similar claims might be made about deceiving or betraying others. On these claims, each person’s aim should be, not that there be less deception or betrayal, but that he himself does not deceive or betray others. These claims are not Consequentialist. And these are the kinds of claim that most of us accept. C can appeal to principles about deception and

betrayal, but it does not appeal to these principles in their familiar form. I shall now describe a different way in which some theory *T* might be self-defeating. Call *T* *indirectly collectively self-defeating* when it is true that, if several people try to achieve their T-given aims, these aims will be worse achieved. On all or most of its different versions, this may be true of C. C implies that we should always try do whatever would make the outcome as good as possible. If we are disposed to act in this way, we are *pure do-gooders.* If we were all pure do-gooders, this might make the outcome worse. This might be true even if we always did what, of the acts that were possible for us, would make the outcome best. The bad effects might come, not from our acts, but from our disposition. There are many ways in which, if we were all pure do-gooders, this might have bad effects. One is the effect on the sum of happiness. On any plausible version of C, happiness is a large part of what makes outcomes better. Most of our happiness comes from having, and acting upon, certain strong desires. These include the desires that are involved in loving certain other people, the desire to work well, and many of the strong desires on which we act when we are not working. To become pure do-gooders, we would have to act against or even to suppress most of these desires. It is likely that this would enormously reduce the sum of happiness. This would make the outcome worse, even if we always did what, of the acts that were possible for us, made the outcome best. It might not make the outcome worse than it *actually* is, given what people are actually like. But it would make the outcome worse than it would be if we were not pure do-gooders, but had certain other causally possible desires and dispositions.[10] There are several other ways in which, if we were all pure do-gooders, this might make the outcome worse. One rests on the fact that, when we want to act in certain ways, we shall be likely to deceive ourselves about the effects of our acts. We shall be likely to believe, falsely, that these acts will produce the best outcome. Consider, for example, killing other people. If we want someone to be dead, it is easy to believe, falsely, that this would make the outcome better. It therefore makes the outcome better that we are strongly disposed not to kill, even when we believe that doing so would make the outcome better. Our disposition not to kill should give way only when we believe that, by killing, we would make the outcome *very much* better. Similar claims apply to deception, coercion, and several other kinds of act. **11. WHY C DOES NOT FAIL IN ITS OWN TERMS** I shall assume that, in these and other ways, C is indirectly collectively self-defeating. If we were all pure do-gooders, the outcome would be worse than it would be if we had certain other sets of motives. If we know this, C tells us that it would be wrong to cause ourselves to be, or to remain, pure do-gooders. Because C makes this claim, it is not failing in its own terms. C does not condemn itself. This defence of C is like my defence of S. It is worth pointing out one difference. S is indirectly individually self-defeating when it is true of some person that, if he was never self-denying, this would be worse for him than if he had some other set of desires and dispositions. This would be a bad effect in S’s terms. And this bad effect often occurs. There are many people whose lives are going worse because they are never, or very seldom, self-denying. C is indirectly collectively self-defeating when it is true that, if some or all of us were pure do-gooders, this would make the outcome worse than it would be if we had certain other motives. This would be a bad effect in C’s terms. But this bad effect may *not* occur. There are few people who are pure do-gooders. Because there are few such people, the fact that they have this disposition may not, on the whole, make the outcome worse. The bad effect in S’s terms often occurs. The bad effect in C’s terms may not occur. But this difference does not affect my defence of S and C. Both theories tell us not to have the dispositions that would have these bad effects. This is why S is not, and C would not be, failing in their own terms. It is irrelevant whether these bad effects actually occur. My defence of C assumes that we can change our dispositions. It may be objected: ‘Suppose that we were all pure do-gooders, because we believe C. And suppose that we could not change our dispositions. These dispositions would have bad effects, in C’s terms, and these bad effects would be the result of our belief in C. So C would be failing in its own terms.’ There was a similar objection to my defence of S. I discuss these objections in Section 18. **12. THE ETHICS OF FANTASY** I have assumed that C is indirectly collectively self-defeating. I have assumed that, if we were all pure do-gooders, the outcome would be worse than it would be if we had certain other sets of motives. If this claim is

true, C tells us that we should try to have one of these other sets of motives. Whether this claim is true is in part a factual question. I believe that it is probably true. But I shall not try to show this here. It seems more worthwhile to discuss what this claim implies. I also believe that, even if we became convinced that Consequentialism was the best moral theory, most of us would not *in fact* become pure do-gooders. Because he makes a similar assumption, Mackie calls Act Utilitarianism ‘the ethics of fantasy’.11 Like several other writers, he assumes that we should reject a moral theory if it is in this sense *unrealistically demanding:* if it is true that, even if we all accepted this theory, most of us would in fact seldom do what this theory claims that we ought to do. Mackie believes that a moral theory is something that we *invent.* If this is so, it is plausible to claim that an acceptable theory cannot be unrealistically demanding. But, on several other views about the nature of morality, this claim is not plausible. We may *hope* that the best theory is not unrealistically demanding. But, on these views, this can only be a hope. We cannot assume that this must be true. Suppose that I am wrong to assume that C is indirectly collectively self-defeating. Even if this is false, we can plausibly assume that C is unrealistically demanding. Even if it would not make the outcome worse if we were all pure do-gooders, it is probably causally impossible that all or most of us become pure do-gooders. Though these are quite different assumptions, they have the *same* implication. If it is causally impossible that we become pure do-gooders, C again implies that we ought to try to have one of the best possible sets of motives, in Consequentialist terms. This implication is therefore worth discussing if (1) C is either indirectly self-defeating or unrealistically demanding, or both, and (2) neither of these facts would show that C cannot be the best theory. Though I am not yet convinced that C is the best theory, I believe both (1) and (2). **13. COLLECTIVE CONSEQUENTIALISM** It is worth distinguishing C from another form of Consequentialism. As stated so far, C is *individualistic* and concerned with *actual* effects. According to C, *each* of us should try to do what would make the outcome best, *given what others will actually do.* And each of us should try to have one of the possible sets of motives whose effects would be best, given the actual sets of motives that will be had by others. Each of us should ask: ‘Is there some other set of motives that is both possible for *me* and is such that, if I had this set, the outcome would be better?’ Our answers would depend on what we know, or can predict, about the sets of motives that will be had by others. What can I predict as I type these words, in January 1983? I know that most of us will continue to have motives much like those that we have now. Most of us will love certain other people, and will have the other strong desires on which most happiness depends. Since I know this, C may tell *me* to try to be a pure do-gooder. This may make the outcome better even though, if we were *all* pure do-gooders, this would make the outcome worse. If most people are *not* pure do-gooders, it may make the outcome better if a few people are. If most people remain as they are now, there will be much suffering, much inequality, and much of most of the other things that make outcomes bad. Much of this suffering I could fairly easily prevent, and I could in other ways do much to make the outcome better. It may therefore make the outcome better if I avoid close personal ties, and cause my other strong desires to become comparatively weaker, so that I can be a pure dogooder. If I am lucky, it may not be bad for me to become like this. My life will be stripped of most of the sources of happiness. But one source of happiness is the belief that one is doing good. This belief may give me happiness, making my austere life, not only morally good, but also a good life for me. I may be less lucky. It may be true that, though I could come close to being a pure do-gooder, this would not be a good life for me. And there may be many other possible lives that would be much better for me. This could be true on most of the plausible theories about self-interest. The demands made on me by C may then seem unfair. Why should I be the one who strips his life of most of the sources of happiness? More exactly, why should I be among the few who, according to C, ought to try to do this? Would it not be fairer if we all did more to make outcomes better? This suggests a form of Consequentialism that is both *collective* and concerned with *ideal* effects. On this theory, each of us should try to have one of the sets of desires and dispositions which is such that, if *everyone* had one of these sets, this would make the outcome better than if everyone had other sets. This statement can be interpreted in several ways, and there are well-known difficulties in removing the ambiguities. Moreover, some versions of this theory are open to strong objections. They tell us to

ignore what would in fact happen, in ways that may be disastrous. But Collective Consequentialism, or *CC,* has much appeal. I shall suggest later how a more complicated theory might keep what is appealing in CC, while avoiding the objections. CC does not differ from C only in its claims about our desires and dispositions. The two theories disagree about what we ought to do. Consider the question of how much the rich should give to the poor. For most Consequentialists, this question ignores national boundaries. Since I know that most other rich people will give very little, it would be hard for me to deny that it would be better if I gave away almost all my income. Even if I gave nine-tenths, some of my remaining tenth would do more good if spent by the very poor. Consequentialism thus tells me that I ought to give away almost all my income. Collective Consequentialism is much less demanding. It does not tell me to give the amount that would in fact make the outcome best. It tells me to give the amount which is such that if we *all* gave this amount, the outcome would be best. More exactly, it tells me to give what would be demanded by the particular International Income Tax that would make the outcome best. This tax would be progressive, requiring larger proportions from those who are richer. But the demands made on each person would be much smaller than the demands made by C, on any plausible prediction about the amounts that others will in fact give. It might be best if those as rich as me all give only half their income, or only a quarter. It might be true that, if we all gave more, this would so disrupt our own economies that in the future we would have much less to give. And it might be true that, if we all gave more, our gift would be too large to be absorbed by the economies of the poorer countries. The difference that I have been discussing arises only within what is called *partial compliance theory.* This is the part of a moral theory that covers cases where we know that some other people will not do what they ought to do. C might require that a few people give away almost all their money, and try to make themselves pure do-gooders. But this would only be because most other people are *not* doing what C claims that they ought to do. They are not giving to the poor the amounts that they ought to give. In its partial compliance theory, C has been claimed to be excessively demanding. This is not the claim that C is *unrealistically* demanding. As I have said, I believe that this would be no objection. What is claimed is that, in its partial compliance theory, C makes *unfair* or *unreasonable* demands. This objection may not apply to C’s *full compliance theory.* C would be much less demanding if we *all* had one of the possible sets of motives that, according to C, we ought to try to cause ourselves to have.[12] **14. BLAMELESS WRONGDOING** Though C is indirectly self-defeating, it is not failing in its own terms. But it may seem open to other objections. These are like those I raised when discussing S. Suppose that we all believe C, and all have sets of motives that are among the best possible sets in Consequentialist terms. I have claimed that, at least for most of us, these sets would not include being a pure dogooder. If we are not pure do-gooders, we shall sometimes do what we believe will make the outcome worse. According to C, we shall then be acting wrongly.

This book challenges, with several powerful arguments, some of our deepest beliefs about rationality, morality, and personal identity. The author claims that we have a false view of our own nature; that it is often rational to act against our own best interests; that most of us have moral views that are directly self-defeating; and that, when we consider future generations the conclusions will often be disturbing. He concludes that moral non-religious moral
philosophy is a young subject, with a promising but unpredictable future.

Derek Parfit, Reasons and Persons PART III â€“ P Chapter 13 - Challenging, with several powerful arguments, some of our deepest beliefs about rationality, morality,

and personal identity, Derek Parfit claims that we have a Derek Parfit, Reasons and Persons PART III â€“ P Chapter 13 - "Reasons, Persons, and Effective Altruism" April 21, 2015. Derek Parfit's Reasons and Persons - Reason Papers - Rahva Raamat Reasons And Persons Derek Parfit - Hof Emsauen - Reasons and Persons is arguably the most influential of the two books published in his lifetime and hailed as a classic work of ethics and "13 Reasons Why" "Riverdale" & "Mr. Robot" - Bleeding Cool - Parfit was not a prolific author; he tended to write his books over the course In the end, he wrote only two: 1984's Reasons and Persons, and Reasons and Persons PDF ePUB MOBI - Emma Bowey - Whatever programming learning you are into, don't learn from one guy, book or tutorial. Learn from 5 to 10 different persons, who are the best Royal Road Small Chests - Lebenszeichen fÃ¼r die Ewigkeit - You must not circulate this book in any other binding or cover Reasons and persons... (2) What makes a person at two different times one and the same. (ebook) Reasons and Persons - 9780191622441 - Dymocks - Open individualism is the view in the philosophy of personal identity, according to which there exists only one numerically identical subject, Buy Reasons And Persons Book at Easons - This is the wisdom behind forgiving others and the Holy Bible also provides us with reasons why we should forgive. 1. Forgive others because we are all sinners. Download e-book Reasons and Persons (Oxford Paperbacks) - Search The Phone Book from BT to find contact details of businesses and.. chance with the person you saw or met, but for whatever reason couldn't talk to. Dangerous Goods(HAZMAT) - IATA - A quote on the back cover says: â€œReasons and Persons may be the greatest movement due to Derek Parfit, and specifically due to this book.

Relevant Books

[DOWNLOAD]