DEREK PARFIT

On What Matters

VOLUME ONE

On What Matters

The Berkeley Tanner Lectures

The Tanner Lectures on Human Values were established by the American scholar, industrialist, and philanthropist Obert Clark Tanner; they are presented annually at nine universities in the United States and England. The University of California, Berkeley became a permanent host of annual Tanner Lectures in the academic year 2000–2001. This two-volume work is the sixth in a series of books based on the Berkeley Tanner Lectures. The volumes include a substantially revised and expanded version of the lectures that Derek Parfit presented at Berkeley in November of 2002, together with the responses of the three invited commentators on that occasion, T. M. Scanlon, Susan Wolf, and Allen Wood; there is also a fourth set of comments, by Barbara Herman, as well as replies to the comments and additional material by Derek Parfit. The volumes are edited by Samuel Scheffler, who also contributes an introduction. The Berkeley Tanner Lecture Series was established in the belief that these distinguished lectures, together with the lively debates stimulated by their presentation in Berkeley, deserve to be made available to a wider audience. Additional volumes are in preparation.

Martin Jay
R. Jay Wallace
Series Editors

Volumes Published in the Series:

Joseph Raz, *The Practice of Value*
Edited by R. Jay Wallace
With Christine M. Korsgaard, Robert Pippin, and Bernard Williams

Frank Kermode, *Pleasure and Change: The Aesthetics of Canon*
Edited by Robert Alter
With Geoffrey Hartman, John Guillory, and Carey Perloff

Seyla Benhabib, *Another Cosmopolitanism*
Edited by Robert Post
With Jeremy Waldron, Bonnie Honig, and Will Kymlicka

Axel Honneth, *Reification: A New Look at an Old Idea*
Edited by Martin Jay
With Judith Butler, Raymond Geuss, and Jonathan Lear

Allan Gibbard, *Reconciling Our Aims*
Edited by Barry Stroud
With Michael Bratman, John Broome, and F. M. Kamm

On What Matters

VOLUME ONE

DEREK PARFIT

Edited and Introduced by
Samuel Scheffler

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford 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 in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

Portions of 'On What Matters' by Derek Parfit were delivered as a Tanner Lecture
on Human Values at the University of California, Berkeley, November 2002.
Printed with permission of the Tanner Lectures on Human Values, a Corporation,
University of Utah, Salt Lake City, Utah, USA.

First published 2011

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Parfit, Derek.
On what matters / Derek Parfit.
p. cm.
Includes bibliographical references and index.
ISBN 978–0–19–957280–9
1. Ethics. I. Title.
BJ1012.P37 2009
170—dc22 2009029662

Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Clay Ltd., St Ives plc

ISBN 978–0–19–957280–9 (Vol. 1)
978–0–19–957281–6 (Vol. 2)
978–0–19–926592–3 (set)

1 3 5 7 9 10 8 6 4 2

To Tom Nagel and Tim Scanlon

This page intentionally left blank

On What Matters

VOLUME ONE

VOLUME TWO

This page intentionally left blank

Contents

VOLUME ONE

VOLUME TWO

PART FOUR

COMMENTARIES

PART FIVE

RESPONSES

This page intentionally left blank

Introduction

Samuel Scheffler

In this densely argued and deeply original book, Derek Parfit addresses some of the most basic questions in practical philosophy. The book comprises two volumes, each containing three parts. Parfit's central chapters, which make up Parts Two and Three, deal with issues of substantive morality. These chapters descend from a series of three Tanner Lectures that Parfit delivered at the University of California at Berkeley in November of 2002. In Parts One and Six, Parfit addresses issues that were not covered in the Berkeley lectures. Part One is an extended discussion of reasons and rationality, which provides the background for his claims about morality in Parts Two and Three. Part Six takes up the meta-normative questions raised by our use of normative language in making claims both about reasons and about morality.

The three commentators who responded to Parfit's Berkeley Tanner Lectures — Thomas Scanlon, Susan Wolf, and Allen Wood — offer revised versions of their comments in Part Four. In addition, Barbara Herman, who was not a participant in the Berkeley events, contributes a set of comments written specially for inclusion in this book. Parfit replies to all of these comments in Part Five. The exchanges between him and the commentators focus primarily on the chapters deriving from the Berkeley lectures.

In his chapters on morality, Parfit aims to rechart the territory of moral philosophy. Students who take courses in the subject are usually taught that there is a fundamental disagreement between consequentialists, who believe that the rightness of an act is a function solely of its overall consequences, and Kantians, who argue — often with reference to one or another version of "the categorical imperative" — that we have certain duties that we must fulfill whether or not doing so

will produce optimal results in consequentialist terms. Although both consequentialist and Kantian views are acknowledged to admit of many variations and refinements, the division between them is assumed by most philosophers, including most consequentialists and Kantians, to be deep and fundamental.

Parfit's primary aim in Parts Two and Three of this book is to undermine this assumption, and to demonstrate the existence of a startling convergence among positions that we are accustomed to viewing as rivalrous. He begins by engaging in a sustained and searching examination of Kant's own moral philosophy, including his various formulations of the categorical imperative and many of his other central moral ideas as well. Although Kant's ethical writings, especially the *Groundwork of the Metaphysics of Morals*, are among the most widely discussed texts in the history of moral philosophy, Parfit's engagement with these texts yields a wealth of fresh observations and insights.

As is evident from his Preface, Parfit's attitude toward Kant is complex and defies easy summary. He describes him as "the greatest moral philosopher since the ancient Greeks" (235), and says that "in the cascading fireworks of a mere forty pages, Kant gives us more new and fruitful ideas than all the philosophers of several centuries" (183). He quickly adds, however, that "[o]f all the qualities that enable Kant to achieve so much, one is inconsistency" (183). Whereas many commentators explicitly present themselves either as critics of Kant or as defenders of his view, Parfit's approach is different. He treats Kant's texts as a rich fund of claims, arguments, and ideas, all of which deserve to be treated with the same seriousness that one would accord the ideas of a brilliant contemporary, but many of which require clarification or revision, and some of which are simply unworkable. Parfit examines a wide range of these claims, arguments, and ideas, subjecting them to a level of scrutiny that is remarkable for its unwavering focus and analytic intensity. His primary aim is neither to defend Kant nor to criticize him, but rather to determine which of his ideas we can use to make progress in moral philosophy. At the end of the day, it is progress that is Parfit's real goal. As he says in explaining why one of Kant's formulations should be revised, "After learning from the works of great philosophers, we

should try to make some more progress. By standing on the shoulders of giants, we may be able to see further than they could" (300).

Parfit identifies several elements of Kant's thought that he regards as particularly important and that he is prepared to endorse, albeit with some significant revisions and additions. However, he frequently differs from other leading commentators in the way he interprets the content and implications of these ideas. This is perhaps most evident in his treatment of the version of the categorical imperative known as the "Formula of Universal Law." As Parfit observes, this formulation of the categorical imperative has been subject to so many serious objections that many otherwise sympathetic commentators have concluded that it is of little value as an action-guiding principle that can help us to distinguish right from wrong. Many leading Kant scholars have concluded that other formulations of the categorical imperative are richer and more illuminating.

Parfit, by contrast, sees great potential in the Formula of Universal Law. Swimming against the prevailing tide of interpretive opinion, he insists that the FUL "*can* be made to work," and he argues that when "revised in some wholly Kantian ways, this formula is . . . remarkably successful" (294). Indeed, he goes so far as to say that a suitably revised version of this formula "might be what Kant said that he was trying to find: the supreme principle of morality" (342).

The revised version of the Formula of Universal Law that Parfit favors states that "Everyone ought to follow the principles whose universal acceptance everyone could rationally will." With its appeal to a kind of universal choice or agreement, this formulation qualifies as a form of "contractualism," and Parfit refers to it as the "Kantian Contractualist Formula." So interpreted, the Kantian position invites comparison with contemporary versions of contractualism, especially those versions that are themselves of broadly Kantian inspiration. John Rawls's appeal to principles that would be chosen behind a veil of ignorance is one example, though Rawls applied this device almost exclusively to the choice of principles of justice for the basic structure of society. He never followed up on the idea, which he had briefly entertained in *A Theory of Justice*, that the same device might be applied to the choice of moral principles more generally. Parfit nevertheless subjects this idea to severe

criticism, and concludes that it is much less promising as a general account of morality than the version of contractualism developed by Thomas Scanlon.

As Parfit states it, "Scanlon's Formula" holds that "Everyone ought to follow the principles that no one could reasonably reject." Parfit argues that, on some interpretations at least, Scanlonian Contractualism coincides with Kantian Contractualism since, on these interpretations, the principles whose universal acceptance everyone could rationally will turn out to be just the same as the principles that no one could reasonably reject. The possibility of convergence between these two forms of contractualism may not seem terribly surprising, although Parfit and Scanlon disagree about the precise extent of the convergence. What is more surprising is Parfit's assessment of the relations between contractualism and consequentialism.

As I have noted, the opposition between the Kantian and consequentialist positions is usually taken to be deep and fundamental, and the contemporary contractualisms of both Rawls and Scanlon are motivated to a significant degree by the desire to articulate a compelling alternative to consequentialism. Yet Parfit argues that Kantian contractualism actually implies a version of "Rule Consequentialism," which holds that "everyone ought to follow the principles whose universal acceptance would make things go best." The principles whose universal acceptance everyone could rationally will, he maintains, just are these "optimific" rule-consequentialist principles. Accordingly, Kantian Contractualism and Rule Consequentialism can be combined to form a view that he calls Kantian Rule Consequentialism: "Everyone ought to follow the optimific principles, because these are the only principles that everyone could rationally will to be universal laws" (411). Although this position is consequentialist in the content of its claims about the principles that people ought to follow, it is more Kantian than consequentialist in its account of why we should follow these principles. We should follow them because their universal acceptance is something that everyone could rationally will, and not because, as consequentialists would have it, all that ultimately matters is that things should go for the best.

Since Kantian Contractualism implies Rule Consequentialism, and since some versions of Kantian Contractualism coincide with some

versions of Scanlonian Contractualism, versions of all three positions can also be combined. The resulting "Triple Theory" holds that an "act is wrong just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable" (413). The upshot of these various possibilities of convergence, Parfit believes, is that it is a mistake to think that there are deep disagreements among Kantians, contractualists, and consequentialists. Instead, "[t]hese people are climbing the same mountain on different sides" (419).

In developing this central line of argument, Parfit relies heavily on substantive claims about reasons and rationality. The theories he is considering all make claims about the kinds of reasons that people have for wanting and doing various things, and about the conditions under which individuals' actions are reasonable or rational. Accordingly, Parfit's assessment of these theories consists largely in assessing the force of different claims of this sort. But claims about reasons and rationality are scarcely less controversial than claims about right and wrong. Recognizing this, Parfit prefaces his chapters on morality with a detailed exposition and defense of his own views on these topics.

Many philosophers believe that our reasons for action are all provided by our desires. We have most reason to do whatever will best fulfill either our actual desires or the desires that we would have under ideal conditions. Although such desire-based views, which Parfit classifies as "subjective theories," have been profoundly influential, both within and outside of philosophy, Parfit believes that they are deeply misguided, and his criticism of them is withering. Not only do they have wildly implausible implications, he argues, but they are ultimately "built on sand." They imply that our reasons derive their normative force from desires that we have no reason to have; but such desires, he argues, cannot themselves be said to give us reasons. In the end, then, the real implication of desire-based views is that we have no reasons for action at all and, more fundamentally, that nothing really matters, in the sense that we have no reason to care about any of the things we do care about.

Rejecting these "bleak" views, Parfit argues that we should instead accept an objective, value-based theory, according to which reasons for action are provided by the values that those acts would realize

or fulfill (or, as he puts it, by the facts that make certain things worth doing for their own sake or make certain outcomes good or bad). Understood in this way, judgments about reasons are more fundamental than judgments about rationality, for we are rational, in Parfit's view, when we respond to reasons or apparent reasons, and our acts are rational when, if our beliefs were true, we would be doing what we had good reasons to do. This contrasts with a number of popular accounts of practical rationality, such as those that identify it with the maximization of expected utility, for example, or those that interpret practical *ir*rationality as a form of inconsistency.

As Thomas Scanlon observes in his contribution, the idea that reasons have priority over rationality also conflicts with Kant's views. For Kant, both the authority and the content of the categorical imperative are to be understood with reference to the requirements of rational agency rather than to some independent conception of the reasons that people have. As Scanlon describes the Kantian view, which he calls "Kantian constructivism about reasons": "Claims about reasons (more exactly, about what a person must see as reasons) must be grounded in claims about rational agency, claims about what attitudes a person can take, consistent with seeing herself as a rational agent. Justification never runs in the other direction, from claims about reasons to claims about what rationality requires" (Volume Two, *118*).*

Parfit, like Scanlon, rejects Kantian constructivism about reasons and, as Scanlon points out, all of the moral theories whose convergence Parfit seeks to demonstrate are framed in such a way as to "appeal to an idea of 'what one can rationally will' that presupposes an independently understandable notion of the reasons that a person has and their relative strength" (*118*). This distinguishes these theories from Kant's own views and also from the views of some prominent contemporary Kantians, such as Christine Korsgaard. As Parfit acknowledges, his reliance on a primitive and "indefinable" notion of "reasons," and his concomitant commitment to the existence of irreducibly normative truths, both about reasons and about morality, makes his view a version of what Korsgaard has called "dogmatic rationalism." As such, it would be resisted not only by Kantian constructivists like Korsgaard but also

* Page numbers in italics refer to Volume Two.

by proponents of some very different meta-ethical outlooks, such as various forms of naturalism and non-cognitivism.

In Part Six, therefore, Parfit undertakes to explain and defend his conception of normativity. He endorses a view that he refers to as "Non-Metaphysical Non-Naturalist Cognitivism," which appeals to certain intuitive beliefs we are said to have about irreducibly normative truths. This view is not Platonistic in the sense of making claims about some supposed non-spatio-temporal portion of reality. Nor is its reliance on intuitions meant to suggest that normative facts are apprehended via a mental faculty that is analogous to sense perception. We do not detect the presence of normative properties like rightness or rationality as a result of being causally affected by them. Instead, we understand normative truths in something like the way we understand mathematical or logical truths. Indeed, Parfit argues, mathematical and logical reasoning themselves involve recognizing and responding to normative truths about what we have reason to believe. For example, we recognize that the truth of *p* and *if p then q* gives us conclusive reason to believe that *q*. Just as there are truths about what we have reason to believe, Parfit insists, so too there are truths about what we have reason to do.

Parfit realizes, of course, that many philosophers do not accept the existence of irreducibly normative truths in his sense. Nihilists and error theorists hold that all normative claims are false. Naturalists hold that normative facts can be reduced to natural facts. Non-cognitivists hold that normative claims, despite their importance in human life, do not function as statements of fact at all. Parfit discusses and criticizes many influential versions of such positions, including the views of Simon Blackburn, Richard Brandt, Allan Gibbard, Richard Hare, John Mackie, and Bernard Williams. None of these views, he argues, can adequately account for the normative dimension of our thought; on all such views, normativity proves to be illusory. It simply disappears. In effect, Parfit appears to believe that all such views tend toward nihilism, and that nihilism is the only genuine alternative to the recognition of irreducibly normative truths. Nor is he persuaded by Korsgaard's Kantian objections to "realism" about normativity. Contrary to what she maintains, he asserts, normativity does not have its source in the

will, but instead consists in the existence of irreducibly normative truths about what we have reason to believe, to want, and to do.

As will be apparent, Parfit's aims in his discussions of reasons and normativity are very different from those he pursues in discussing substantive moral theories. In the moral case, his aim is to demonstrate that certain putatively opposing theories may actually converge, so that apparent disagreement among them evaporates. But in his discussion of different views about reasons and normativity, a convergence among rival theories is not on the agenda. Instead, he argues that a value-based theory of reasons should be accepted and that desire-based theories should be rejected. Similarly, his form of Cognitivism should be accepted in preference to all forms of Naturalism and Non-Cognitivism. Parfit is clearly troubled by substantive moral disagreement, for he thinks it threatens to undermine our conviction that there is such a thing as moral truth. That is why he is so strongly driven to demonstrate the possibility of convergence among rival moral theories. Although he is also troubled by meta-ethical or meta-normative disagreement, his response to it is different. Here he simply attempts to determine which of the contending positions is correct. Yet to the extent that the substantive moral theories whose convergence Parfit seeks to demonstrate all presuppose his views about reasons and normativity, the frankly contested character of those views may call into question the significance of the convergence he describes at the substantive moral level. Those who reject value-based theories of reasons, and those who accept one or another form of naturalism or non-cognitivism or constructivism, may be unmoved by a moral consensus that depends on accepting the very meta-ethical views that they reject. So one challenge for Parfit is to demonstrate that the significance of the convergence for which he argues is not undermined by its dependence on claims, such as those concerning reasons and normativity, about which there is no convergence. Although Parfit does not directly address this challenge, he does argue that those who have rejected the views about reasons and normativity that he favors have not always fully understood them. And he expresses the hope that, once the relevant misunderstandings have been cleared away, many more philosophers will eventually come to accept those views. If this is correct, then even though the competing theories of reasons and of

normativity do not themselves converge, there may be reason to hope for much greater convergence in the assessments that philosophers give of them. Of course, this suggestion is itself likely to be controversial.

There are many other questions that can and will be raised about Parfit's subtle and intricate arguments. One issue, different aspects of which are discussed by each of the four commentators, concerns the extent to which the views whose convergence Parfit seeks to demonstrate are authentic versions of more familiar moral views. To what extent is Kantian Contractualism really Kantian? We have already seen that, in its account of the relation between rationality and reasons, the view appears to be more Parfit's than Kant's. Similar questions can be raised about the other ostensibly convergent positions. To what extent does Scanlonian Contractualism reflect Scanlon's own views? And what is the relation between Parfit's version of Rule Consequentialism and other consequentialist formulations?

The issue is a tricky one. As Scanlon notes, Parfit is forthright about his willingness, in developing a "Kantian" position, to depart from Kant's actual views whenever he thinks he can improve upon them. As Parfit says, "We are asking whether Kant's formulas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise these formulas in ways that are clearly needed, we are developing a Kantian moral theory" (298). In his reply to Scanlon, he is similarly explicit about the fact that his argument for the convergence of Kantian Rule Consequentialism and Scanlonian Contractualism "does not apply to the view stated in Scanlon's book" (*244*), but rather to a version of that view that has been revised in ways that Parfit takes to strengthen it.

This unapologetic revisionism carries with it two risks for Parfit. The first, which Scanlon mentions, is that the degree to which any convergence he can demonstrate will seem surprising and significant may depend on how close the convergent theories are to the eponymous ancestors from which they descend. The more they have been revised in ways that depart from their original formulations, the less surprising and significant their convergence may seem. The second risk is that, in revising the original theories to bring them closer to one another, valuable elements of the original theories may be excluded.

Susan Wolf appears to harbor doubts of both of these kinds about Parfit's claims of convergence. Of Parfit's ambition to reconcile the Kantian, consequentialist, and contractualist traditions, she writes: "[I]nsofar as the remarks quoted above are meant to suggest that the values these different traditions emphasize can be interpreted and ordered in such a way as to eliminate the tensions among them, or that it would be in the spirit of these traditions' greatest exponents to accept revisions and qualifications to their stated views that would ultimately reconcile them with their opponents, Parfit departs from the explicit positions of any of the philosophers whose work he discusses, in a way that seems to me both interpretively implausible and normatively regrettable" (*32*). Wolf's view is that the Kantian, consequentialist, and contractualist traditions embody divergent evaluative perspectives, each of which has something important to contribute but which are in genuine tension with one another. These tensions reflect broader tensions within our moral thought itself. As such, she believes, they are ineliminable and not to be regretted. Any unified principle of the kind Parfit seeks will perforce be a matter of compromise rather than complete convergence, and any such principle will inevitably leave out something of value. Wolf presses this last point with special reference to Parfit's version of Kantianism, which, she argues, scants the importance of autonomy in Kant's own moral philosophy.

Barbara Herman too believes that Parfit's position departs from Kant's in fundamental ways. However, while Wolf expresses doubts about the very idea that morality rests on a unified principle of the kind that Parfit seeks, Herman is sympathetic to Kant's own unified account and believes that Parfit's theory is an unstable mixture of disparate elements. More specifically, she argues that Parfit employs a "hybrid" methodology that incorporates some Kantian features but nevertheless has "a strongly consequentialist cast" (*81*). Although Parfit's intention is to preserve what is most persuasive in Kant's view while avoiding some of the apparently unwelcome implications of that view, Herman believes that there is such a deep "mismatch" between the Kantian and consequentialist methodologies that the attempt to combine them inevitably distorts Kant's own account and obscures what is most appealing about it. In the first portion of her comments, she identifies

several elements of Parfit's methodology that she regards as deeply consequentialist in character, and she gives illustrations of the resulting methodological divide that she sees between Parfit and Kant. Perhaps the most basic difference is this: whereas Parfit appeals to various nonmoral goods to determine what people could rationally will and so to fix the content of morality itself, Kant, Herman says, seeks to establish a place for nonmoral goods within an independently established moral framework. In the remainder of her commentary, she attempts to demonstrate that this "unified" Kantian approach, properly developed, has the resources to accommodate some of the most important moral intuitions—such as those concerning permissible lies—that Kant has seemed to neglect. If this is correct, then much of the motivation for a hybrid moral methodology disappears. In his reply, Parfit does not directly engage with Herman's thoughtful attempt to develop the unified Kantian view in this way. However, he disputes her assessment of the "mismatch" between his methodology and Kant's. Most of the ostensibly consequentialist aspects of his method that she cites, he maintains, are also features of Kant's view. And although he does propose revisions in Kant's Formula of Universal Law, some of these revisions are fully in the spirit of the Kantian view, while others are necessary to avoid straightforward mistakes. The upshot, Parfit believes, is that the gap between his own position and Kant's is far narrower, and far shallower, than Herman asserts.

Like Herman, Allen Wood also argues that Parfit's philosophical methodology departs from Kant's in important ways, although he focuses on different aspects of Parfit's approach than Herman does. Wood believes that Parfit employs a method originated by Sidgwick, which sets itself the goal of providing a "scientific" ethics. The idea is to systematize our commonsense moral opinions, correcting them when necessary, with the aim of arriving at a precise set of principles that can be used algorithmically to yield a determinate moral verdict about how one should act in any conceivable situation. Wood believes that such otherwise diverse philosophers as Kant, Bentham, and Mill employ a very different method, which he himself regards as preferable to the one he ascribes to Sidgwick and Parfit. This alternative method begins not with commonsense intuitions but rather with a fundamental principle

that serves to articulate some basic value. General moral rules or duties are then derived non-deductively from the fundamental principle. These rules or duties represent an attempt to interpret the implications of the fundamental value in the conditions of human life. The rules or duties themselves admit of exceptions and require interpretation, and their application to particular cases calls for the exercise of judgment and cannot be codified in precise rules or principles. So, on the one hand, the Kantian method as Wood understands it gives less weight than the Sidgwickian method to commonsense moral intuitions; but, on the other hand, it regards as "hopeless" the aim of constructing a "scientific" ethics that can provide an algorithm for moral decision-making.

Wood believes—though Parfit's reply suggests that he would not accept this diagnosis—that the difference of method just described underlies some disagreements between Parfit and him concerning the proper interpretation of Kant's Formula of Humanity. He thinks it also underlies their sharply divergent attitudes toward one familiar type of philosophical argument. This type of argument uses our intuitive reactions to stylized and sometimes complex hypothetical examples to test candidate moral principles. Wood refers to all such examples as "trolley problems," whether or not they involve actual trolleys, in mock *hommage* to the famous case first introduced into the philosophical literature by Philippa Foot. Parfit makes frequent use of such examples in constructing his arguments. For instance, his argument for the convergence of Kantian Contractualism and Rule Consequentialism turns crucially on some claims about what a person could rationally agree to in situations where one course of action would impose a burden on the person himself and the only alternative would impose burdens on others. Parfit illustrates and defends these claims with reference to a series of hypothetical examples involving burdens of different sizes and types imposed in a range of different hypothetical circumstances. He seeks to marshal our intuitive responses in these cases to show (1) that each person could rationally will the universal acceptance of the consequentially optimific principles, even when those principles would impose some burden on the person himself, and (2) that there are no other principles whose universal acceptance everyone could rationally choose. Parfit evidently believes that the use of hypothetical

examples can help to clarify the issues that are at stake in complex moral choices and enable us to make progress in moral argument. Wood, by contrast, regards "trolley problems" as "worse than useless for moral philosophy" (*68*), and the majority of his essay is given over to an extended critique of the ways in which reliance on such problems leads moral philosophers astray.

To the extent that other people share Wood's reservations about appealing to hypothetical examples in moral philosophy, Parfit's extensive reliance on such examples may be a source of resistance to his arguments. Of course, even those who do not endorse Wood's radical rejection of all such appeals may find themselves disagreeing with Parfit's reactions to some of the specific examples he discusses, although Parfit anticipates many potential disagreements and exhibits great resourcefulness in attempting to defuse them. Yet Parfit himself points out that our reactions to some of these cases may depend, for example, on whether we accept a desire-based or a value-based theory of reasons. Since he hopes to use our reactions to support his claim of convergence among different moral theories, this kind of variation represents one way in which disagreements about reasons and rationality, like meta-ethical disagreements about the nature of normative judgment, threaten to destabilize the moral consensus that Parfit aims to establish. As I have already said, Parfit's response to this threat is not to look for convergence among the rival meta-ethical theories or theories of reasons and rationality themselves. Instead, he argues that there are decisive reasons for rejecting the alternatives to Non-Metaphysical Non-Naturalist Cognitivism and the value-based theory of reasons, and he pins his hopes for convergence on the possibility that philosophers will eventually come to accept the cognitivist and value-based positions that he favors. This is a different way of eliminating or at least taking the sting out of disagreement: by demonstrating that there is only one position that we can reasonably accept.

The drive to eliminate disagreement—whether by establishing theoretical convergence or through a decisive demonstration of the inadequacy of competing views—is a defining feature of Parfit's work. It is sometimes marked by a sense of urgency. One place where this emerges is in his reply to Susan Wolf. Wolf takes Parfit to be trying

to show "that there is a single true morality, crystallized in a single supreme principle which these different traditions may be seen to be groping towards, each in their own separate and imperfect ways" (*32*). She herself says, by contrast, that "it would not be a moral tragedy if it turned out" that morality did not have such a unifying principle (*33*). In response, Parfit agrees that it would not be a tragedy if there were no single supreme principle. But, he adds, "it *would* be a tragedy if there was no *single true morality*." He adds: "if we cannot resolve our disagreements, that would give us reasons to doubt that there are *any* true principles. There might be nothing that morality *turns out to be*, since morality might be an illusion." (*151*). It is, perhaps, the spectre of this "bleak" possibility, and the even bleaker possibility that, as Parfit worries, nothing at all may matter, that is responsible for the sense of urgency with which he pursues the elimination of disagreement. Whether or not one shares his assessment of the threat posed by deep disagreement, one cannot fail to be impressed by the extraordinary ingenuity and the sheer intellectual intensity with which he pursues his goal. His rich and challenging discussion, helpfully illuminated by his exchanges with Barbara Herman, Thomas Scanlon, Susan Wolf, and Allen Wood, casts familiar debates in a fresh and unfamiliar light, and opens up many fruitful new lines of inquiry for philosophers to investigate. Nobody who is interested in the theory of morality, rationality, or normativity will want to ignore this brilliant, provocative, and tenaciously argued book.

Preface

Since this book contains summaries, I shall say little about its contents here. Though the book is long, there are some shorter books within it. Nothing important in Part Three depends on Part Two, so you might read only Parts One and Three. If you are mainly interested in ethics, you might read only Chapters 6 to 17. If you are mainly interested in reasons, rationality, and meta-ethics, you might read only Parts One and Six.

While describing how he came to write his great, drab book *The Methods of Ethics*, Sidgwick remarks that he had 'two masters': Kant and Mill. My two masters are Sidgwick and Kant.

Kant is the greatest moral philosopher since the ancient Greeks. Sidgwick's *Methods* is, I believe, the best book on ethics ever written. There are some books that are greater achievements, such as Plato's *Republic* and Aristotle's *Ethics*. But Sidgwick's book contains the largest number of true and important claims. It is not surprising that, though a less great philosopher than Plato, Aristotle, Hume, and Kant, Sidgwick could write a better book. Sidgwick lived later. Unlike later poets or playwrights, who have no advantages over Homer or Shakespeare, later philosophers do have advantages, since philosophy makes progress.

Sidgwick and Kant both have weaknesses and flaws. Sidgwick is sometimes boring, for example, and Kant is sometimes maddening. I hope that by admitting these weaknesses, and saying why we should not be disappointed or deterred by them, I may persuade some people to read, or re-read, Sidgwick's *Methods* and some of Kant's books.

Kant and Sidgwick are a wonderfully contrasting pair. Discussing their own achievements, for example, Kant writes:

> . . . the critical philosophy must remain confident of its irresistible propensity to satisfy the theoretical as well as the moral, practical purposes of reason, confident that no change of opinions, no touching up or reconstruction into some other form, is in store for it; the system of the *Critique* rests on a fully secured foundation, established forever; it will prove to be indispensable too for the noblest ends of mankind in all future ages;

Sidgwick writes:

> The book solves nothing, but may clear up the ideas of one or two people, a little.

Kant is very original, makes some sublime claims, and is excitingly intense. Sidgwick knew that he lacked these qualities. 'I like criticizing myself', he writes to a friend, 'and have formulated the following on it:

> *Pro*: Always thoughtful, often subtle: generally sensible and impartial: approaches the subject from the right point of view.
>
> *Con*: Inconsequent, ill-arranged: stiff and ponderous in style, nothing really striking or original in the arguments.'

Sidgwick also refers to his 'one damning defect of longwinded & difficult dullness'.

This last phrase is too severe. Though Sidgwick's book is long, and some of its chapters can now be ignored, it is not longwinded. Sidgwick seldom repeats himself, and he makes many important points concisely, and only once. Nor is Sidgwick's book difficult. Some of his claims and arguments are complicated, but they are nearly all clearly written.

Sidgwick's dullness needs more discussion. Whitehead was so bored by Sidgwick's *Methods* that he never looked at another book on ethics.

But after reading a collection of Sidgwick's memoirs and letters, Keynes remarked, 'I have never found so dull a book so absorbing'. It is worth quoting from this book. Discussing the Church of England, Sidgwick writes:

> At Cambridge I get into the way of regarding it as something that once was alive and growing, but now exists merely because it is a pillar or buttress of uncertain value in a complicated edifice that no one wants just now to take to pieces. Here however, I feel rather as if I were contemplating a big fish out of water, propelling itself smoothly and gaily over the high road.

Here are two more passages:

> There is no doubt that men in England fall in love chiefly in abnormal periods: when on a reading party, or at the seaside, or at a foreign hotel, or at Christmas, or any other occasion when something, either external circumstances or any dominant emotion, thaws the eternal ice. The misfortune is that if these casual thaws do not last long enough, all the advantage gained is lost; two lines of life that causally intersected diverge perhaps for ever, and the frost sets in with redoubled force.

> I am bearing the burden of humanity in the lap of luxury, and in consequence not bearing it well. After all, Pascal was practically right: if one is to embrace infinite doubt, if it is to come into our bowels like water, and like oil into our bones, it ought to be upon sackcloth and ashes and in a bare cell, and not amid '47 port and the silvery talk of W. G. Clark. When I go to my rooms I feel strange, ghastly, that is why I write to you. But there again — if one allows this consciousness 'the time is short' to grow and get too strong, it seems to fold up all life into a feverish moment.
>
> The world shall feel my impulse or I die.

> Think of all the second-rate men who have said this and died—and—Who cares?
>
> Butterflies may dread extinction.
>
> This is a strange mood for me. But at Trumpington today I brushed away a spider's life and said 'This is sentience.' What am I more than elaborate sentience?

Sidgwick could be amusing, and his conversation was described as 'like the sparkling of a brook whose ripples seem to give out sunshine'. But the first edition of the *Methods* contains only a few jokes, some of which Sidgwick later removed. Much of the book, however, is well-written. For example:

> to suppose . . . that the ideal of 'obeying oneself alone' can be even approximately realized by Representative Democracy is even more patently absurd. For a representative assembly is normally chosen only by a part of the nation, and each law is approved by only a part of the assembly: and it would be ridiculous to say that a man has assented to a law passed by a mere majority of an assembly *against* one member of which he has voted.

More soberly:

> . . . the Cosmos of Duty is thus really reduced to a Chaos, and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure.

This magnificently sombre claim has some of the intensity of Kant, as does another passage that is about Kant:

> I cannot fall back on the resource of thinking myself under a moral necessity to regard all my duties *as if they were* commandments of God, although not entitled to hold speculatively that any such Supreme Being really exists. I am so far from feeling bound to believe for purposes of practice what I see no ground for holding as a speculative truth, that I

> cannot even conceive the state of mind which these words seem to describe, except as a momentary half-witted irrationality, committed in a violent access of philosophic despair.

Many fine passages are too long to quote in full. One such passage ends:

> . . . the selfish man misses the sense of elevation and enlargement given by wide interests; he misses the more secure and serene satisfaction that attends continually on activities directed towards ends more stable in prospect than an individual's happiness can be: he misses the peculiar rich sweetness, depending upon a sort of complex reverberation of sympathy, which is always found in services rendered to those whom we love and who are grateful. He is made to feel in a thousand various ways . . . the discord between the rhythms of his own life and of that larger life of which his own is but an insignificant fraction.

Another passage ends:

> . . . even a man who said 'Evil be thou my good' and acted accordingly might have only an obscured consciousness of the awful irrationality of his action—obscured by a fallacious imagination that his only chance of being in any way admirable, at the point of which he has now reached in his downward course, must lie in candid and consistent wickedness.

Sidgwick warned his friends that, because his book attempts to achieve 'precision of thought', it 'cannot fail to be somewhat dry and repellent'. But this precision is often finely expressed. Discussing friendship, for example, Sidgwick describes

> the sympathy that is not quite admiration with which Common Sense regards all close and strong affections; and the regret that is not quite disapproval with which it contemplates their decay.

Many sentences, though dry, have an ironical edge or twist. For example:

> It may be said that a child owes gratitude to the authors of its existence. But life alone, apart from any provision for making life happy, seems a boon of doubtful value, and one that scarcely excites gratitude when it was not conferred from any regard for the recipient.
>
> . . . there seems to be no justice in making A happier than B, merely because circumstances beyond his control have first made him better.
>
> Thus the Utilitarian conclusion, carefully stated, would seem to be this: that the opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric.
>
> . . . really penetrating criticism, especially in ethics, requires a patient effort of sympathy which Mr Bradley has never learned to make, and a tranquillity of temper which he seems incapable of maintaining.
>
> [The book] seems smashing, but he loses by being over-controversial. There should be at least an affectation of fairness in a damaging attack of this kind.

Sidgwick's irony can make him seem stuffy, when in fact he is being subversive. Bernard Williams had been misled, for example, when he wrote that Sidgwick's discussions of sexual morality, though sometimes mildly adventurous, 'make fairly uncritical use of a notion of purity'. Sidgwick does ask 'What, then, is the conduct that Purity forbids?' But if we read him carefully, we find that his answer is: Nothing. In a book published in England in 1874, it was more than mildly adventurous to argue, though in guarded terms, that there is no moral objection to indulging in sexual pleasure for its own sake.

When people find Sidgwick dull, they are often responding not to Sidgwick's style, but to one of his greatest philosophical merits. Sidgwick describes this merit well, writing in his journal:

> Have been reading Comte and Spencer, with all my old admiration for their intellectual force and industry and more than my old amazement at their fatuous self-confidence. It does not seem to me that either of them knows what self-criticism means. I wonder if this is a defect inseparable from their excellences. Certainly I find my own self-criticism an obstacle to energetic and spirited work: but on the other hand I feel that whatever value my work has is due to it.

Sidgwick was unusually good at seeing the force of objections to his views. After hearing Sidgwick defend a paper, William James remarked:

> Sidgwick displayed that reflective candour that can at times be so irritating. A man has no right to be so fair to his opponents.

Discussing an opponent's book, for example, Sidgwick writes:

> I shall praise it as much as I can . . . it is by an author of fine qualities . . . But yet—he seems to me altogether out of it: I can scarcely treat his theory with proper respect. No doubt I seem so to him: and are we not both right? The book makes me rather depressed about ethics.

These virtues can make Sidgwick hard to read. One problem is that, as C. D. Broad explains, Sidgwick

> incessantly refines, qualifies, raises objections, answers them, and then finds further objections to the answer. Each of these objections, rebuttals, rejoinders, and surrejoinders is in itself admirable, and does infinite credit to the acuteness and candour of the author. But the reader is apt to become impatient; to lose the thread of the argument; and to rise from his desk finding that he has read a great deal with constant admiration and now remembers little or nothing.

Our first reading of the *Methods* is, in a way, the worst, since there is little that is striking or inspiring. But every time we re-read this book, we notice some new good points that we had earlier overlooked. That is what I, at least, have found.

Criticizing himself again, Sidgwick writes:

> I am not an original man: and I think less of my own thoughts every day.

This remark is also too severe. Sidgwick is in several ways original. But that is not what makes him great. Other philosophers, like Kant and Hume, are more original, and more brilliant. These philosophers are like Newton and Einstein: geniuses of the clearest kind. Sidgwick is more like Darwin. He had what has been called 'good sense intensified almost to the point of genius'. In the *Methods*, as Broad claims, 'almost all the main problems of ethics are discussed with extreme acuteness'. And Sidgwick gets very many things right. He gives the best critical accounts of three of the main subjects in ancient and modern ethics: hedonism, egoism, and consequentialism. And in the longest of his book's four parts, he also gives the best critical account of pluralistic non-consequentialist common sense morality. Though Sidgwick makes mistakes, some of which I mention in a note, he does not, I believe, make many. These facts make Sidgwick's *Methods* the book that it would be best for everyone interested in ethics to read, remember, and be able to assume that others have read.

My debts to Sidgwick are easy to describe. Of my reasons for becoming a graduate student in philosophy, one was the fact that, in wondering how to spend my life, I found it hard to decide what really matters. I knew that philosophers tried to answer this question, and to become wise. It was disappointing to find that most of the philosophers who taught me, or whom I was told to read, believed that the question 'What matters?' couldn't have a true answer, or didn't even make sense. But I bought a second-hand copy of Sidgwick's book, and I found that he at least believed that some things matter. And it was from Sidgwick that I learnt most about the other questions that moral philosophers should ask, and about some of the answers.

I turn now to my other master, Kant. When I first read Kant's *Groundwork* in the 1960s, I found this book fascinating but obscure. When I re-read this book thirty years later, and most of Kant's other books, I became unexpectedly obsessed with Kant's ethics. For the next two or three years, I thought about little else.

It seems worth confessing that, though my obsession with Kant gave me great energy, this energy was, to start with, almost entirely negative. I didn't doubt Kant's genius. But like many other people, I found myself deeply opposed both to some of Kant's main claims, and to his way of doing philosophy. By mentioning what made me so opposed to Kant, and saying how my attitude has changed, I may perhaps persuade some other people not to ignore Kant, as I nearly did.

Though Kant has some important qualities that Sidgwick lacks, Kant also lacks some important qualities that Sidgwick has. Sidgwick writes clearly, is on the whole consistent, and makes few mistakes. These things cannot be claimed of Kant.

Unlike our first reading of Sidgwick's *Methods*, our first reading of Kant's *Groundwork* is, in some ways, the best. There are some striking and inspiring claims, and we are not worried by what we can't understand. But when we re-read the *Groundwork*, many of us become discouraged, and give up. We decide that Kant, though he may be a great philosopher, is not for us.

The first problem is Kant's style. It is Kant who made really bad writing philosophically acceptable. We can no longer point to some atrocious sentence by someone else, and say 'How can it be worth reading anyone who writes like that?' The answer could always be 'What about Kant?'

There are deeper problems. When I became obsessed with Kant, I tried to restate more clearly some of Kant's main claims and arguments, and found this task very frustrating. I couldn't fit Kant's claims together in a coherent view, and many of Kant's arguments seemed to be obviously invalid or unsound. It would have helped me to know that even some of Kant's greatest admirers have similar feelings. Onora O'Neill, for example, calls the *Groundwork* 'the most exasperating' of Kant's books.

It would also have helped me to know that Kant did not have a single, coherent theory. When we ask whether Kant accepts or rejects some claim, the answer is often 'Both'. As Kemp Smith writes, 'citation of single passages is quite inconclusive'. For example, though Kant writes that 'a human being's duty at each instant is to do all the good in his power', he is not really, as this claim implies, an Act Consequentialist. Rawls remarks that, when he tried to understand Kant's texts, 'I assumed there were never plain mistakes, not ones that mattered anyway'. But there must be mistakes, since Kant makes many conflicting claims, and such claims cannot all be true. As Kemp Smith points out, Kant often 'flatly contradicts himself' and 'there is hardly a technical term which is not employed by him in a variety of different and conflicting senses. He is the least exact of the great thinkers.' (To avoid provoking Hegelians, we should perhaps say 'one of the least exact'.)

'Consistency', Kant writes, 'is a philosopher's greatest duty.' That is not true. Originality and clarity are at least as important. And Kant's greatness chiefly consists in his having many original and fruitful ideas. If Kant had always been consistent, he could not have had all these ideas.

When I first re-read Kant, what I found most irritating was not Kant's obscurities and inconsistencies, but a particular kind of overblown, false rhetoric. For example, Kant writes:

> If we look back upon all previous efforts that have ever been made to discover the principle of morality, we need not wonder why all of them had to fail. It was seen that the human being is bound to laws by his duty; but it never occurred to them that he is subject only to laws given by himself but still universal and that he is obligated only to act in conformity with his own will . . .

I didn't mind the exaggeration in the first sentence here. We can switch the volume down, turning 'all of them had to' into 'some of them did'. But since I knew that Kant believed in a Categorical Imperative, I was surprised by Kant's second sentence. I asked a Kantian, 'Does this mean that, if I don't give myself Kant's Imperative as a law, I am not subject

to it?' 'No,' I was told, 'you have to give yourself a law, and there's only one law.' This reply was maddening, like the propaganda of the so-called 'People's Democracies' of the old Soviet bloc, in which voting was compulsory and there was only one candidate. And when I said 'But I haven't given myself Kant's Imperative as a law', I was told 'Yes you have'. This reply was even worse. My irritation at such claims may have left some traces in this book.

As I have said, however, that irritation has gone. Now that I have read Kant's other works, I am aware of the passions that led Kant to make his most outrageous claims. When he is calmer, he makes other, better claims. For example, Kant is reported to have said:

> Suicide is the most abominable of the crimes that inspire horror and hatred . . . he who so utterly fails to respect his life . . . can in no way be restrained from the most appalling vices . . .

But he also said:

> In the Stoic's principle concerning suicide there lay much sublimity of soul: that we may depart from life as we leave a smoky room.

Some of Kant's impassioned arguments, moreover, have great charm. When condemning suicide, Kant said:

> If freedom is the condition of life, it cannot be employed to abolish life . . . Life is supposedly being used to bring about lifelessness, but that is a self-contradiction.

It is the word 'supposedly' that is so endearing here. Suicide involves a contradiction, one commentator suggests, because it is we, on Kant's view, who confer value on our ends. If we kill ourselves to avoid suffering, we

> cut off the source of the goodness of this end—it is no longer really an end at all, and it is no longer rational to pursue it.

This conclusion arrives too late.

For another example, consider Kant's claim that, if we tell some lie 'even to achieve some really good end', we 'violate the dignity of humanity in our own person' and make ourselves a 'mere deceptive appearance of a human being', who has 'even less worth than if he were a mere thing'. We should ignore such outbursts. On the very next page Kant suggests that, if we are asked by an author whether we like his work, we may be permitted to say what he expects.

Kant is sometimes thought of as a cold, dry, rationalist. But he is really an emotional extremist. As Sidgwick writes, 'Oh, how I sympathize with Kant! with his passionate yearning for synthesis and condemned by his reason to criticism . . . ' Kant seldom uses words like 'most', 'many', 'several', or 'some', preferring to write only 'all' or 'none'. Kant uses 'good', he says, to mean 'practically necessary'. And he seldom uses the concept of a reason: a fact that merely *counts in favour* of some act, since his preferred normative concepts are *required*, *permitted*, and *forbidden*. Temperamentally, I am an extremist too, who has to struggle to be more like Sidgwick.

Oxford University once had a useful marking grade: *Alpha Gamma*. As everyone should agree, Kant's books are pure Alpha Gamma, containing nothing that is *Beta*, or mediocre. Our disagreement should be only about how much of what Kant wrote is Alpha, and how much is Gamma. And if we have found what is Alpha, we can ignore what is Gamma.

Some of Kant's views are, I believe, too close to Hume's. Kant is a more dangerous Anti-Rationalist because, unlike Hume, he seems to be exalting what he calls *Pure Reason*. And Kant's influence has been, I believe, in some other ways bad. But he is very great, and his influence has been, in other and less obvious ways, good. Though Kant makes many claims that are false, and many of his arguments fail, he also gives us some profound truths. Like Sidgwick, I sometimes find him 'quite a revelation'. Kant's books are very thought-provoking. As Rawls writes, 'Part of the wonderful character of the works we study is the depth and variety of ways they can speak to us.'

In this book I try to say something about most of Kant's formulations of his supreme principle of morality. That is why I wrote much of Part Two, though the book's main arguments are in Parts One, Three, and Six. But except in a few sections, which are mostly in Part Two or Appendices F to I, I do not discuss the details of Kant's views.

I turn now to the other people from whom I have learned most. When I was young, most philosophers believed that there could not be any normative truths. So did most economists, other social scientists, and much of the wider Western world. Well-educated non-religious people took for granted the distinction between facts, which are objective, and mere values. Little has changed. When some economist recently claimed that his proposals involved no value judgments, someone else said 'Yes they do. You assume that we ought to do what would be better for some people and worse for no one.' 'That's not a *value judgment*,' this economist replied, 'Everyone accepts it'.

As well as finding, in the long-dead Sidgwick, someone who had greater hopes for practical and moral philosophy, I was encouraged to find some living philosophers who had such hopes. I was encouraged most by Thomas Nagel, and in particular by Nagel's claims about reasons, and about irreducibly normative truths. I have also learnt a great deal from Tim Scanlon. I often cannot remember whether some thought was mine or his. I dedicate this book to these two people.

I am grateful to Christine Korsgaard, whose impressive books led me to reread Kant, and whose critique of what she calls 'dogmatic rationalism' helped to rouse me from my undogmatic slumbers. I have also learnt much (even if not enough) from the remarkable recent series of other books and articles on or inspired by Kant, by such writers as Henry Allison, Marcia Baron, David Cummiskey, Richard Dean, Jeffrey Edwards, Stephen Engstrom, Paul Guyer, Barbara Herman, Thomas Hill, Samuel Kerstein, Patricia Kitcher, Onora O'Neill, Thomas Pogge, Andrews Reath, Jerome Schneewind, David Sussman, Roger Sullivan, and Allen Wood.

I have been greatly helped by many other people, who gave me comments on early drafts. Since it would be impossible to describe in a few pages the many ways in which I have been helped, I can only express my great gratitude to these people.

Of those who gave me comments on all parts of this book, I owe most to Robert Audi, Selim Berker, Talbot Brewer, John Broome, Ruth Chang (to whom I dedicate Chapter 16), Eugene Chislenko, Jerry Cohen, Garrett Cullity, Jonathan Dancy, David Enoch, William Fitzpatrick, Shelly Kagan, Guy Kahane, Niko Kolodny, Michael Otsuka, Ingmar Persson, Jacob Ross, Kieran Setiya, and Larry Temkin. Some parts of this book were jointly written with these people.

I was also greatly helped by Marcello Antosh, Richard Arneson, Rüdiger Bittner, Mary Coleman, Roger Crisp, Stephen Darwall, Harry Gensler, Reto Givel, Elizabeth Harman, Brad Hooker, Frances Kamm, Joseph Mendola, Jefferson McMahan, Liam Murphy, Leonard Katz, Robert Myers, Martin O'Neill, Douglas Portmore, Stuart Rachels, Peter Railton, Karl Schafer, Samuel Scheffler, Michael Slote, Saul Smilansky, Jussi Suikkanen, and Stephen White.

Of those who gave me comments only on Part One, I was helped most by Melissa Barry, David Copp, Joshua Gert, Pamela Hieronymi, Julia Markovits, Sven Nyholm, Connie Rosati, Jeffrey Sebo, David Sobel, Sigrun Svavarsdottir, David Velleman, and Michael Zimmerman.

Of those who gave me comments on my claims about Kant, I was helped most by Marcia Baron, David Cummiskey, Richard Dean, Jeffrey Edwards, Paul Guyer, Thomas Hill, Samuel Kerstein, Patricia Kitcher, Thomas Pogge, and Allen Wood. I have failed to respond adequately to the comments of Edwards, Kitcher, and Pogge on my interpretations of Kant, and to the comments by Samuel Freeman and Leif Wenar on my claims about Rawls.

Of those who gave me comments only on Part Six, I was helped most by Robert Adams, Paul Boghossian, Laurence Bonjour, Nicholas Bostrom, Philip Bricker, Justin Clarke-Doane, Terence Cuneo, Cian Dorr, Kit Fine, Stephen Finlay, Alvin Goldman, Bob Hale, Michael

Jubien, Thomas Kelly, Brian Leiter, William Lycan, Tim Maudlin, Brian McLaughlin, Charles Parsons, Simon Rippon, Stephen Schiffer, Mark Schroeder, Russ Shafer-Landau, Peter Singer, Knut Skarsaune, Robert Stalnaker, Brian Weatherson, Ralph Wedgwood, and Timothy Williamson.

I have also been helped by Larry Alexander, Henry Allison, Gustaf Arrhenius, Elizabeth Ashford, Bruce Aune, Annette Baier, Matthew Bedke, Akeel Bilgrami, Daniel Boisvert, Matthew Boyle, Sarah Buss, Krister Bykvist, Thomas Carson, Timothy Chappell, Daniel Cohen, Joshua Cohen, Robert Curtis, Gordon Davis, Paul Dinkin, Thomas Donaldson, Dale Dorsey, Jamie Dreier, Julia Driver, Jerry Dworkin, Andrew Egan, Nir Eyal, Geoffrey Ferrari, Claire Finkelstein, Katrin Flikschuh, Johann Frick, Jerry Gaus, Berys Gaut, Tamar Gendler, Pablo Gilabert, Margaret Gilbert, George Giovanni, Joshua Glasgow, James Grant, Liron Greenstein, Alex Gregory, Ish Haji, Jason Hanna, Robert Hanna, Joshua Harlan, Daniel Hausman, Allan Hazen, Christopher Heathwood, Dieter Henrich, David Heyd, Alison Hills, Nathan Holcomb, Mike Huemer, Thomas Hurka, Paul Hurley, Susan Hurley, Frank Jackson, Dale Jamieson, Justin Jeffrey, Leonard Kahn, Robert Kane, Stephen Kearns, Paul Klumpe, Richard Kraut, Rahul Kumar, Joel Kupperman, Arto Laitinen, Robin Lawlor, Mark LeBar, James Lenman, John Leslie, Hallvard Lillehammer, Don Loeb, David Lyons, Tienmu Ma, Jacqueline Marina, David McCarthy, Kris McDaniel, Dennis McKerlie, Chris McMahon, David McNaughton, Elijah Millgram, Adrian Moore, Sophia Moreau, Adam Morton, Istvan Musza, Jan Narveson, Stephen Nathanson, William Nelson, Michael Neumann, Kenneth O'Day, Avner Offer, Onora O'Neill, Serena Olsaretti, Jonas Olson, Toby Ord, Leah Orent, Francesco Orsi, David Owens, Stephen Palmquist, Herlinde Pauer-Studer, David Phillips, Christian Piller, Richard Price, Bogdan Rabanca, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, Joseph Raz, Andrews Reath, Bernard Reginster, Michael Ridge, Arthur Ripstein, Michael Rohlf, Gideon Rosen, Mike Rosen, Carol Rovane, Angelica Rudenstine, Julian Savulescu, Jerome Schneewind, Dieter Schoenecker, Frederick Schueler, Bart Schultz, Sally Sedgwick, Jeffrey Seidman, Matthew Seligman, Julius Sensat, Andrew Sepielli, Robert Shaver, Walter Sinnott-Armstrong, John Skorupski, Holly Smith, Michael Smith, Tom

Sorell, Carlos Soto, Amia Srinivasan, Cynthia Stark, Philip Stratton-Lake, Galen Strawson, Bart Streumer, Roger Sullivan, Adam Swenson, Folke Tersman, Jens Timmerman, Torbjörn Tännsjö, Pekka Vayrynen, Edna Ullmann-Margalit, David Velleman, Benjamin Vilhauer, Gerard Vong, Alex Voorhoeve, R. Jay Wallace, James Walmsley, Paul Weirich, Kenneth Westphal, Evan Williams, Chris Woodard, Helena Wright, and Masahiro Yamada.

I thank All Souls College for the immense privilege of a Research Fellowship during the many years in which I have written this book. I thank the Tanner Foundation for supporting the lectures which I expanded into Parts Two and Three. I am grateful to the Commentators on these lectures who wrote Part Four, and to Samuel Scheffler for his work as Editor. And I am very grateful to Peter Momtchiloff of the Oxford University Press for giving me, over many years, much wise advice.

SUMMARY

PART ONE REASONS

CHAPTER 1 NORMATIVE CONCEPTS

1 *Normative Reasons*

We are the animals that can both understand and respond to reasons. Facts give us reasons when they count in favour of our having some belief or desire, or acting in some way. When our reasons to do something are stronger than our reasons to do anything else, this act is what we have *most reason* to do, and may be what we *should*, *ought to*, or *must* do. Though it is facts that give us reasons, what we can *rationally* want or do depends instead on our beliefs.

2 *Reason-Involving Goodness*

Things can be good or bad by having features that might give us reasons to respond to these things in certain ways. Events can be good or bad *for* particular people, or *impersonally* good or bad, in reason-implying senses. On some widely accepted views about reasons, nothing could be in these ways good or bad.

CHAPTER 2 OBJECTIVE THEORIES

3 *Two Kinds of Theory*

According to *subjective* theories, we have most reason to do whatever would best fulfil or achieve our present desires or aims. Some Subjectivists appeal to our actual present desires or aims; others appeal to the

desires or aims that we would now have, or to the choices that we would now make, if we had carefully considered the relevant facts. Since these are all facts about *us*, we can call such reasons *subject-given*. According to *objective* theories, we have reasons to act in some way only when, and because, what we are doing or trying to achieve is in some way good, or worth achieving. Since these are facts about the *objects* of these desires or aims, we can call such reasons *object-given*. They are also *value-based*. Theories of these two kinds often deeply disagree. We ought, I shall argue, to accept some value-based objective theory.

4 *Responding to Reasons*

When we are aware of facts that give us strong reasons to have particular desires, our response to these reasons is seldom voluntary. Nor can we choose how we respond to most of our reasons to have particular beliefs. Our rationality consists in part in our non-voluntary responses to these reasons.

5 *State-Given Reasons*

When it would be good if we had certain beliefs or desires, that may seem to give us reasons to have these beliefs or desires. But such reasons would have no importance.

6 *Hedonic Reasons*

The same facts give us object-given reasons both to have and to try to fulfil certain desires. What we want is always some possible event, in the wide sense that covers acts and states of affairs. We have *telic* reasons to want some events as ends, or for their own sake, and *instrumental* reasons to want some events as a means to some good end. We have most reason to do whatever would best achieve the ends that we have most reason to want, because the intrinsic features of these ends make them relevantly best.

When we are in pain, what is bad is not our sensation but our conscious state of having a sensation that we dislike. It is similarly good to have sensations that we like. Such *hedonic likings* or *disliking* cannot be rational or irrational, since we have no reasons to like or dislike these

sensations. We also have *meta-hedonic desires* about our own and other people's pleasures and pains. Such desires or preferences *can* be rational or irrational, since we can have strong reasons to have them. It is our hedonic likings and dislikings, not our meta-hedonic desires, that make these conscious states good or bad; so the examples of pleasure and pain do not support the view that our desires can give us reasons, and can make their objects good.

7 *Irrational Preferences*

If we want some event as an end, but this event's intrinsic features give us strongly decisive reasons to want this event *not* to occur, our wanting this event is contrary to reason, and irrational. It would be irrational, for example, to prefer to have one hour of agony tomorrow rather than one minute of slight pain later today. These claims may seem too obvious to be worth making. But such claims are denied by some great philosophers, and they cannot be made by those who accept subjective theories about reasons.

CHAPTER 3 SUBJECTIVE THEORIES

8 *Subjectivism about Reasons*

Subjectivism takes several forms. Subjective theories may appeal to all of our present telic desires, or only to desires that rest on true beliefs, or only to fully informed desires. Some Subjectivists appeal to the choices that we would now make after informed and rational deliberation. Some Objectivists appeal to the choices that we would make, after such deliberation, if we were rational. Though these claims seem similar, they are very different. These Subjectivists claim only that we should deliberate in ways that are *procedurally* rational. Objectivists make claims about what we would choose if we were *substantively* rational. According to Objectivists, what we *ought rationally* to choose depends on our reasons. According to these Subjectivists, our reasons depend on what, after such deliberation, we *would in fact* choose.

9 *Why People Accept Subjective Theories*

Since so many people believe that *all* practical reasons are desire-based, aim-based, or choice-based, how could it be true that, as objective

theories claim, there are *no* such reasons? How could all these people be so mistaken? There are several possible explanations, since there are several ways in which our desires or aims may seem to give us reasons.

10 *Analytical Subjectivism*

Some claims seem to be *substantive*, but are merely *concealed tautologies*, which everyone could accept whatever else they believe. Several Subjectivists use the words 'reason', 'should', and 'ought' in *subjectivist* senses. These people's theories do not make substantive claims.

11 *The Agony Argument*

Substantive subjective theories can have implausible implications. These theories imply, for example, that we often have no reason to want to avoid some future period of agony. Some Subjectivists would respond to this objection by appealing to claims about procedural rationality. Such replies fail.

CHAPTER 4 FURTHER ARGUMENTS

12 *The All or None Argument*

Subjective theories could also imply that we have decisive reasons to cause ourselves to be in agony for its own sake, to waste our lives, and to try to achieve other bad or worthless aims. In response to this objection, Subjectivists might claim that, for some desire or aim to give us a reason, we must have some reason to have this desire or aim. But these people cannot defensibly make this claim. On subjective theories, all that matters is *whether* some act would fulfil our present fully informed desires or aims. It is irrelevant *what* we want, or are trying to achieve. Either *all* of these desires give us reasons for acting, or *none* of them do. Since it is clear that some of these desires could not give us reasons, we should conclude that none of them do.

Some of our desires can be claimed to give us reasons to have other desires, but any such chain of desire-based reasons must begin with some desire that we have no reason to have. Since such desires cannot be defensibly claimed to give us reasons, Subjectivists cannot defensibly

claim that we have desire-based reasons to have any desire or aim, or to act in any way.

13 *The Incoherence Argument*

Many Subjectivists claim that we have most reason to fulfil, not our actual present desires or aims, but the desires or aims that we would now have if we knew the relevant facts. These people also claim that, when we are making important decisions, we ought to try to learn more about the different possible outcomes of our acts, so that we shall come to have better informed desires. Since Subjectivists deny that the intrinsic features of these outcomes give us reasons, they cannot coherently make these claims.

14 *Reasons, Motives, and Well-Being*

If we are Subjectivists, we must deny that events can be good or bad for particular people, or impersonally good or bad, in the reason-implying senses. When some writers claim that some life would be best for someone, they mean that this is the life that, after fully informed and procedurally rational deliberation, this person would in fact choose. On this account, the best life for someone might be a life of unrelieved suffering. That is not a helpful claim. Some other accounts fail in other ways.

15 *Arguments for Subjectivism*

On subjective theories, *nothing matters*. We should reject the arguments for this bleak view.

CHAPTER 5 RATIONALITY

16 *Practical and Epistemic Rationality*

We are rational insofar as we respond well to reasons or apparent reasons. We have some *apparent* reason when we have beliefs about the relevant facts whose truth would give us some reason. Our desires and acts are rational when, if our beliefs were true, we would have sufficient reasons to have these desires, and to act in these ways. Some people add

that, for our desires or acts to be rational, they must depend on rational beliefs. This claim is misleading, and not worth making.

On one view, what is distinctive of epistemic rationality is the aim of reaching true beliefs. There is another, better view. As well as drawing a deeper distinction between epistemic and practical rationality, we should draw this distinction in a different way, and in a different place.

17 *Beliefs about Reasons*

According to some writers, to be fully rational, we don't need to respond to reasons, or apparent reasons. It is enough to avoid certain kinds of inconsistency, such as failing to respond to what we ourselves believe to be reasons. Such views are too narrow.

18 *Other Views about Rationality*

The rationality of our desires does not depend, as many people claim, on whether these desires are consistent, or on how we came to have them, or on whether our having them has good effects. Our desires are rational when they depend on beliefs whose truth would make the objects of these desires, or what we want, in some way good or worth achieving.

CHAPTER 6 MORALITY

19 *Sidgwick's Dualism*

We can assess the strength of our reasons, Sidgwick seems to argue, from two points of view. When assessed from our personal point of view, self-interested reasons are supreme. When assessed from an impartial point of view, impartial reasons are supreme. To compare the strength of these two kinds of reason, we would need some third, neutral point of view. Since there is no such point of view, self-interested and impartial reasons are *wholly incomparable*. When reasons of these two kinds conflict, neither could be stronger. We would always have sufficient or undefeated reasons to do either what would be impartially best or what would be best for ourselves.

We should reject Sidgwick's argument. We ought to assess the strength of all our reasons from our actual, personal point of view, and we do not need a neutral point of view. We should also revise Sidgwick's conclusion. We have personal and partial reasons to be specially concerned, not only about our own well-being, but also about the well-being of certain other people, such as our close relatives and those we love. These are the people to whom we have *close ties*. We also have impartial reasons to care about anyone's well-being, whatever that person's relation to us. Though there are truths about the relative strengths of these two kinds of reason, Sidgwick's view is partly right, since these comparisons are, even in principle, very imprecise. As *wide value-based objective* theories claim, when one of two possible acts would be impartially better, but the other act would be better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either way.

20 *The Profoundest Problem*

As well as asking 'What do I have most reason to do?', we can ask 'What ought I morally to do?' If these questions often had conflicting answers, because we often had most reason to act wrongly, morality would be undermined. Like other normative requirements, moral requirements matter only when they give us reasons.

Though reasons are more fundamental, much of what follows is about morality. But I shall also be discussing reasons. Several moral principles and theories appeal to claims about what, in actual or imagined situations, we would have most reason or sufficient reason to consent to, or agree to, or to want, or choose, or do.

CHAPTER 7 MORAL CONCEPTS

21 *Acting in Ignorance or with False Beliefs*

By distinguishing several senses of 'ought morally' and 'wrong', we can recognize some important truths and avoid some unnecessary disagreements. Acts can be wrong in *fact-relative*, *evidence-relative*,

belief-relative, and *moral-belief-relative* senses. Facts about these kinds of wrongness provide answers to different questions. When what we ought to do depends on the goodness of our act's effects, we ought to try to do, not what would in fact be best, but what would be *expectably-best*.

22 *Other Kinds of Wrongness*

There are several other senses of 'wrong', which may refer to different kinds of wrongness. Most of these senses are worth using.

It is a difficult question whether, as I believe, there are some irreducibly normative truths, some of which are moral truths. These questions will be easier to answer when we have made more progress in our thinking about practical and epistemic reasons, and about morality. Rather than proposing a new moral theory, I shall try to develop existing theories of three kinds: Kantian, Contractualist, and Consequentialist.

PART TWO PRINCIPLES

CHAPTER 8 POSSIBLE CONSENT

23 *Coercion and Deception*

We act wrongly, Kant claims, when we treat people in any way to which they cannot possibly consent. This claim may seem to imply that we ought never to coerce or deceive people, since these may seem to be acts whose nature makes consent impossible. But that is not relevantly true.

24 *The Consent Principle*

Kant's claims about consent can be interpreted in two ways. On the *Choice-Giving Principle*, it is wrong to treat people in any way to which these people *cannot actually* give or refuse consent, because we have failed to give these people the power to choose how we treat them. This principle is clearly false. On the *Consent Principle*, it is wrong to treat people in any way to which they *could not rationally* consent, if we gave them the power to choose how we treat them. This principle is more likely to be what Kant means, and might be true.

Kant's claims gives us an inspiring ideal of how, as rational beings, we ought to be related to each other. We might be able to treat everyone only in ways to which they could rationally consent; and this might be how everyone ought always to act.

25 *Reasons to Give Consent*

Whether we could achieve Kant's ideal depends on which are the acts to which people could rationally give informed consent, because they would have sufficient reasons to consent. If the best theory about reasons were either some subjective theory, or Rational Egoism, the Consent Principle would fail, since there would be countless permissible or morally required acts to which some people could not rationally consent. But if the best theory is some wide value-based objective theory, as I believe, the Consent Principle may succeed. As some examples suggest, there may always be at least one possible act to which everyone could rationally consent. And we have reasons to believe that, in all such cases, it would be wrong to act in any way to which anyone could not rationally consent.

26 *A Superfluous Principle?*

According to some writers, even if the Consent Principle is true, this principle adds nothing to our moral thinking. What is morally important is not the fact that people could not rationally consent to certain acts, but the various facts that give these people decisive reasons to refuse consent. When applied to acts that affect only one person, this objection has some force. But when our acts would affect many people, if there is only one possible act to which everyone could rationally consent, this fact would give us a strong reason to act in this way, and would help to explain why the other possible acts would be wrong. It is also worth asking whether we could achieve Kant's ideal.

27 *Actual Consent*

It is wrong to treat people in certain ways if these people either do not, or would not, actually consent to these acts. Such acts are wrong even if these people could have rationally given their consent. That is no

objection to the Consent Principle, which claims to describe only one of the facts that can make acts wrong.

On one view, it is wrong to treat people in any way to which they actually refuse consent. That is clearly false. It may seem that no one could rationally consent to being treated in any way to which they actually refuse consent. If that were true, the Consent Principle would also be clearly false. But this objection can be answered.

According to *the Rights Principle*, everyone has rights not to be treated in certain ways without their actual consent. In stating and applying this principle, we would need to answer some difficult questions.

28 *Deontic Beliefs*

To explain why the Consent Principle does not mistakenly require certain wrong acts, we must appeal to the fact that these acts are wrong in other ways, or for other reasons. On some plausible assumptions, the Consent Principle could never require us to act wrongly, because any act's wrongness would give everyone sufficient reason to consent to our failing to act in this way.

29 *Extreme Demands*

The Consent Principle can require us to bear great burdens, when that would save some other people from much greater burdens. If this requirement is too demanding, we would have to revise this principle. But we might still be able to achieve Kant's ideal.

CHAPTER 9 MERELY AS A MEANS

30 *The Mere Means Principle*

It is wrong, Kant claims, to treat any rational being merely as a means. We treat people in this way when we both use these people and regard them as mere tools, whom we would treat in whatever way would best achieve our aims. On a better version of Kant's principle, it is wrong to treat people merely as a means, or to *come close* to doing that.

We do not treat someone merely as a means, nor are we close to doing that, if either (1) our treatment of this person is governed in sufficiently important ways by some relevant moral belief or concern, or (2) we do or would relevantly choose to bear some great burden for this person's sake.

Suppose that some Egoist benefits himself by keeping some promise to someone whose help he needs, and saving some drowning child for the sake of getting some reward. Since this man treats these other people merely as a means, Kant's principle mistakenly condemns these acts. We could qualify this principle, so that it condemns treating someone merely as a means only if our act is also likely to harm this person.

Suppose next that some driverless runaway train is headed for a tunnel in which it would kill five people. These people's lives cannot be saved except by your causing me, without my consent, to fall onto the track, thereby killing me but stopping the train. It may seem that, if you acted in this way, you would be treating me merely as a means. But in some versions of this case that would not be true. And I could rationally consent to being treated in this way. Though such acts may be wrong, that wrongness is not implied by either the Mere Means Principle or the Consent Principle.

31 As *a Means and* Merely *as a Means*

It is widely believed that if we harm people, without their consent, as a means of achieving some aim, we thereby treat these people merely as a means, in a way that makes our act wrong. This view involves three mistakes. When we *harm* people as a means, we may not be treating *these people* as a means. Even if we *are* treating these people as a means, we may not be treating them *merely* as a means. And even if we *are* treating them merely as a means, we may not be acting wrongly.

Some people give other accounts of what is involved in treating people merely as a means. These accounts seem to be either mistaken, or unhelpful.

32 *Harming as a Means*

If it would be wrong to impose certain harms on people as a means of achieving certain aims, these acts would be wrong even if we were *not* treating these people *merely* as a means. And if it would *not* be wrong to impose certain other harms on people as a means of achieving certain aims, these acts would not be wrong even if we *were* treating these people merely as a means. Though it is wrong to *regard* anyone merely as a means, the wrongness of our *acts* never or hardly ever depends on whether we are treating people merely as a means.

CHAPTER 10 RESPECT AND VALUE

33 *Respect for Persons*

We ought to respect everyone, but that does not tell us how we ought to act. It is wrong, some writers claim, to treat people in ways that are incompatible with respect for them. This claim does not help us to decide, in difficult cases, whether some act would be wrong.

34 *Two Kinds of Value*

Some things have a kind of value that is to be *promoted*. Possible acts and other events are in this way good when there are facts about them that give us reasons to make them actual. People have a kind of value that is to be *respected*. Such value is not a kind of goodness.

35 *Kantian Dignity*

Kant uses 'dignity' to mean supreme value or worth. It is sometimes claimed that, on Kant's view, such supreme value is had only by rational beings, or persons, and is the kind of value that should be respected rather than promoted. But that is not Kant's view. There are several ends or outcomes that Kant claims to have supreme value, and to be ends that everyone ought to try to promote.

Some of Kant's remarks suggest that non-moral rationality has supreme value. But Kant's main claims do not commit him to this implausible view. Kant fails to distinguish between being supremely good and having

a kind of moral status that is compatible with being very bad. But we can add this distinction to Kant's view.

36 *The Right and the Good*

Some ancient Greeks, Kant claims, mistakenly tried to derive the moral law from their beliefs about the Greatest Good. But Kant describes an ideal world, which he calls the *Highest* or *Greatest Good*, and he claims that everyone ought always to strive to produce this world. Kant may seem here to be making what he calls the 'fundamental error' of these ancient Greeks. But that is not so.

37 *Promoting the Good*

In Kant's ideal world, everyone would be virtuous and would have all the happiness that their virtue would make them deserve. We can do most to produce this world, Kant claims, by strictly following his other principles. It is often thought that, when Kant claims that lying is always wrong, he is thereby rejecting Act Consequentialism. That is not so. But when Kant, Hume, and others make such claims, they fail to draw some distinctions that we need to draw.

CHAPTER 11 FREE WILL AND DESERT

38 *The Freedom that Morality Requires*

If our acts were merely events in time, Kant argues, these acts would be causally determined, so we could never have acted differently, and morality would be an illusion. Since morality is not an illusion, our acts are not merely events in time. This argument fails. Though we *ought* to have acted differently only if we *could* have done so, the relevant sense of 'could' is compatible with determinism.

39 *Why We Cannot Deserve to Suffer*

According to another of Kant's arguments, if our acts were merely events in time, we could never be responsible for these acts in some way that could make us deserve to suffer. Since we *can* be responsible for our acts in this desert-involving way, our acts are not merely such events.

Though this argument is valid, it is not sound. We ought to accept Kant's claim that, if our acts were merely such events, we could not deserve to suffer. But since we ought to reject this argument's conclusion, we ought to reject Kant's other premise. Our acts *are* merely events in time. So we cannot deserve to suffer.

PART THREE THEORIES

CHAPTER 12 UNIVERSAL LAWS

40 *The Impossibility Formula*

By our *maxims* Kant means, roughly, our policies and underlying aims. According to Kant's *stated* version of what we can call his *Impossibility Formula*, it is wrong to act on any maxim that could not be a universal law. There is no useful sense in which this could be claimed to be true.

According to Kant's *actual* version of his Impossibility Formula, it is wrong to act on any maxim of which it is true that, if everyone accepted and acted on this maxim, or everyone believed that they were morally permitted to act upon it, that would make it impossible for anyone successfully to act upon it. This formula spectacularly fails, since it does not condemn acts of self-interested killing, injuring, coercing, lying, and stealing. Kant's formula rightly condemns the making of lying promises. But this formula condemns such acts for a bad reason, and it mistakenly condemns some good or morally required acts.

41 *The Law of Nature and Moral Belief Formulas*

Kant proposes another, better formula. To apply this formula, we suppose that we have the power to *will*, or choose, that certain things be true. We act wrongly, Kant claims, if we act on some maxim that we could not rationally will to be a universal law. There are three versions of this *Formula of Universal Law*. According to

> *the Law of Nature Formula*, it is wrong to act on some maxim unless we could rationally will it to be true that everyone accepts this maxim, and acts upon it when they can.

According to

> *the Permissibility Formula*, it is wrong to act on some maxim unless we could rationally will it to be true that everyone is morally permitted to act upon it.

According to

> *the Moral Belief Formula*, it is wrong to act on some maxim unless we could rationally will it to be true that everyone believes that such acts are morally permitted.

It will be enough to consider Kant's Law of Nature and Moral Belief Formulas. These formulas develop the ideas that are expressed in two familiar questions: 'What if everyone did that?' and 'What if everyone thought like you?'

When we apply these formulas, we must appeal to some view about rationality and reasons. Since we are asking what Kant's formulas can achieve, we should appeal to what we believe to be the best view. But we should not appeal to our beliefs about which acts are wrong, or to the *deontic* reasons that such wrongness might provide, since Kant's formulas would then achieve nothing.

42 *The Agent's Maxim*

Whether some act is wrong, Kant's formulas assume, depends on the agent's maxim. Most of the maxims that Kant discusses are, or include, *policies*. Suppose that some Egoist has only one maxim or policy: 'Do whatever would be best for me'. This man could not rationally will it to be true either that everyone acts on this maxim, or that everyone believes such acts to be permitted. Most Egoists could not rationally choose to live in a world of Egoists, since that would be much worse for them than worlds in which people act on various moral maxims. Whenever our imagined Egoist acts on his maxim, Kant's formulas imply that this man's acts are wrong. This man acts wrongly even when, for self-interested reasons, he pays his debts, puts on warmer clothing, and saves some drowning child in the hope of getting some reward. These implications are clearly

false. When this Egoist acts in these ways, his acts do not have what Kant calls *moral worth*, but they are not wrong.

Consider next Kant's maxim 'Never lie'. Kant could not have rationally willed it to be true that no one ever tells a lie, not even to a would-be murderer who asks where his intended victim is. Kant's formula therefore implies that, if Kant acted on this maxim by telling anyone the truth, he acted wrongly. That is clearly false. As these and other cases show, whether some act is wrong cannot depend on the agent's maxim, in the sense that can refer to policies. There are many policies on which it is sometimes but not always wrong to act. Nor does an act's moral worth depend on the agent's maxim.

Kant's appeal to the agent's maxim raises other problems. Such problems have led some people to believe that Kant's Formula of Universal Law cannot help us to decide which acts are wrong. When used as such a criterion, these people claim, Kant's Formula is unacceptable, worthless, and cannot be made to work.

Kant's Formula *can* be made to work. When revised in certain ways, I shall argue, this formula is remarkably successful.

Some writers suggest that, rather than appealing to the agent's actual maxim, Kant's Formula should appeal to the possible maxims on which the agent might have been acting. This suggestion fails.

In revising our two versions of Kant's Formula, we should drop the concept of a maxim, and use instead the morally relevant description of the acts that we are considering. The Law of Nature Formula could become:

> We act wrongly unless we are doing something that we could rationally will everyone to do, in similar circumstances, if they can.

The Moral Belief Formula could become:

> We act wrongly unless we could rationally will it to be true that everyone believes such acts to be permitted.

These formulas will need some further revisions.

It may be objected that, if we revise Kant's formulas by dropping the concept of a maxim, we are no longer discussing Kant's view. That is true, but no objection. We are developing a Kantian moral theory, in a way that may make progress.

CHAPTER 13 WHAT IF EVERYONE DID THAT?

43 *Each-We Dilemmas*

It will be simpler to go on discussing Kant's formulas, returning to our revised versions when that is needed.

On Kant's Law of Nature Formula, it is wrong to act on some maxim unless we could rationally will it to be true that *everyone* rather than *no one* acts upon it. We are often members of some group of whom it is true that, if *each* rather than *none* of us did what would be *better* for ourselves, *we together* would be doing what would be *worse* for all of us. In many such cases, each of us could either benefit ourselves or give some greater benefit to others. We can face similar *each-we dilemmas* when we have certain other morally permitted or required aims, such as the aim of promoting our children's well-being. It may be true that, if each rather than none of us did what would be better for our own children, *we* would be doing what would be worse for everyone's children. We could not rationally will it to be true that everyone rather than no one acts in these ways. So if everyone followed Kant's Law of Nature Formula, no one would act in these ways, and that would be better for everyone. These are the cases in which we can best think and say 'What if everyone did that?'

Kant's formula is especially valuable when the bad effects of any single act are spread over so many people that the effects on each person are trivial or imperceptible. One example are the acts with which we are selfishly over-heating the Earth's atmosphere. By requiring us to do only what we could rationally will everyone to do, Kant's formula helps us to see how much harm we are doing, and strongly supports the view that such acts are wrong. In some of these cases, we can add, common sense morality is *directly collectively self-defeating*, and should therefore be revised.

44 *The Threshold Objection*

Whether it is wrong to act on some maxim sometimes depends on how many people act upon it. There are some maxims on which it is permissible or good for some people to act, though it would be very bad if everyone acted on them. Two examples are the maxims 'Consume food without producing any,' and 'Have no children, so as to devote my life to philosophy'. Most of us could not rationally will it to be true that everyone acts on these maxims, so Kant's Law of Nature Formula condemns such acts even when they are not wrong. This objection is partly answered by the fact that most people's maxims implicitly take into account what other people are doing. For a complete answer, we must revise Kant's formula.

45 *The Ideal World Objections*

Kant's Law of Nature Formula, it is often claimed, requires us to act as if we were living in an ideal world, even when in the real world such acts would have predictably disastrous effects and be clearly wrong. We are required, for example, never to use violence even in self-defence, and required to act in various ways that mistakenly ignore what other people will in fact do. This *Ideal World Objection* can be answered. Kant's formula does not require such acts.

There is a different problem. Once a few people have failed to do what we could rationally will everyone to do, Kant's formula permits the rest of us to do whatever we like. Similar objections apply to some *Rule Consequentialist* moral theories. To answer this *New Ideal World Objection*, we should revise Kant's formula in another way. It is wrong to act on some maxim, this formula could claim, unless we could rationally will it to be true that this maxim be acted on, not only by everyone rather than by no one, but also by *any other number* of people rather than by no one. Rule Consequentialists could make similar claims.

Of the two versions of Kant's Formula of Universal Law, the Moral Belief Formula is better. When people object 'What if everyone did that?', it is often enough to reply 'Most people won't'. But when people object 'What if everyone thought like you?', it is *not* enough merely to reply 'Most people won't'.

CHAPTER 14 IMPARTIALITY

46 *The Golden Rule*

Kant's contempt for the Golden Rule is not justified.

47 *The Rarity and High Stakes Objections*

When people act wrongly, they may either be doing something that cannot often be done, or be giving themselves benefits that are unusually great. In some of these cases, these people could rationally will it to be true both that everyone acts like them, and that everyone believes such acts to be permitted. So Kant's formulas mistakenly permit these people's wrong acts.

48 *The Non-Reversibility Objection*

Many wrong acts benefit the agent but impose much greater burdens on others. The Golden Rule condemns such acts, since we would not be willing to have other people do such things to us. But when we apply Kant's formulas, we don't ask whether we could rationally will it to be true that *other* people do these things to *us*. We ask whether we could rationally will it to be true that *everyone* does these things to *others*. And we may know that, even if everyone did these things to others, *no one* would do these things to *us*. In such cases, many wrong-doers could rationally will it to be true both that everyone acts like them, and that everyone believes such acts to be morally permitted. So Kant's formulas mistakenly permit these people's acts.

This objection applies to many actual cases. Some examples are the acts with which many men benefit themselves by treating women as inferior, denying women certain rights and privileges, and giving less weight to women's well-being. To argue that Kant's formulas condemn these men's acts, we would have to claim that these men could not rationally will it to be true either that they and other men continue to benefit themselves in these ways, or that everyone, including all women, believes these acts to be justified. Since we cannot appeal to our belief that these acts are wrong, we cannot plausibly defend this claim. So Kant's formulas mistakenly permit such acts. Similar claims apply to

some of the acts with which many people who are powerful or rich exploit and oppress some other people who are weak or poor.

49 *A Kantian Solution*

To avoid this and some of our other objections, we should again revise Kant's Formula of Universal Law. It will be enough to revise Kant's Moral Belief Formula, which could become:

> It is wrong to act in some way unless *everyone* could rationally will it to be true that everyone believes such acts to be morally permitted.

When everyone believes some act to be permitted, everyone accepts some principle that permits such acts. If some moral theory appeals to the principles that everyone could rationally choose to be universally accepted, this theory is *Contractualist*. So we can restate this formula, and give it another name. According to

> *the Kantian Contractualist Formula*: Everyone ought to follow the principles whose universal acceptance everyone could rationally will.

This formula might be what Kant was trying to find: the supreme principle of morality.

CHAPTER 15 CONTRACTUALISM

50 *The Rational Agreement Formula*

Many Contractualists ask us to imagine that we and others are trying to reach agreement on which moral principles everyone will accept. According to

> *the Rational Agreement Formula*: Everyone ought to follow the principles to whose universal acceptance it would be rational in self-interested terms for everyone to agree.

This version of Contractualism either has no clear implications, or gives unfair advantages to those who would have greater bargaining power.

51 *Rawlsian Contractualism*

Rawls claims that, to avoid these objections, we should add a *veil of ignorance*. According to

> *Rawls's Formula*: Everyone ought to follow the principles to whose universal acceptance it would be rational in self-interested terms for everyone to agree, if everyone had to reach this agreement without knowing any particular facts about themselves or their circumstances.

This version of Contractualism, Rawls claims, provides an argument against all forms of Utilitarianism. That is not true. Nor does Rawlsian Contractualism support acceptable non-Utilitarian principles.

52 *Kantian Contractualism*

To reach a better version of Contractualism, we should return to the Kantian Formula. We should ask which principles each person could rationally choose, if this person knew all of the relevant facts, and had the power to choose which principles everyone would accept. According to the Kantian Formula, everyone ought to follow the principles that, in these imagined cases, everyone could rationally choose.

53 *Scanlonian Contractualism*

According to Scanlon's partly similar formula, everyone ought to follow the principles that no one could *reasonably reject*. Since Scanlon appeals to claims about what is reasonable in a partly moral sense, it may seem that, if we accept Scanlon's Formula, that would make no difference to our moral thinking. But that is not so.

Scanlon once claimed that his formula gives an account of wrongness itself, or of *what it is* for acts to be wrong. Contractualist formulas are better claimed to describe one of the facts that can *make* acts wrong. Scanlon's view now takes this form.

54 *The Deontic Beliefs Restriction*

When we apply any Contractualist formula, Contractualists must claim, we cannot appeal to our intuitive beliefs about which acts are wrong.

If we appealed to such *deontic* beliefs, these formulas would achieve nothing. Some Contractualists claim that we should never appeal to such intuitive deontic beliefs, which involve mere prejudice, or cultural conditioning. We should reject such claims. When we are trying to decide which acts are wrong, we must appeal to these intuitive beliefs. Contractualists should claim instead that, though we cannot appeal to such beliefs *while* we are working out what their formula implies, we *can* appeal to these beliefs when we later try to decide whether, given these implications, we ought to accept this formula.

CHAPTER 16 CONSEQUENTIALISM

55 *Consequentialist Theories*

Whatever moral view we hold, we can use 'best' in the impartial-reason-implying sense. Some outcome is in this sense best when it is the outcome that, from an impartial point of view, everyone would have most reason to want. These outcomes should be taken to include acts, and their goodness may in part depend on facts about the past. *Consequentialist* moral theories appeal only to claims about how it would be best for things to go. *Direct* Consequentialists apply this criterion to everything. When these people apply this criterion to acts, they are *Act Consequentialists. Indirect* Consequentialists apply this criterion directly to some things, but indirectly to others. According to some *Motive Consequentialists*, for example, though the best motives are the motives whose being had by everyone would make things go best, the best or right acts are not the acts that would make things go best, but the acts that would be done by people with the best motives. Indirect Consequentialism can take many other forms.

56 *Consequentialist Maxims*

According to *Maxim Consequentialists*, everyone ought to act on the maxims whose being acted on by everyone would make things go best. On every plausible or widely accepted view about rationality, Kant's original Law of Nature Formula permits some people to be Maxim Consequentialists.

57 to 62 *The Kantian Argument*

According to one version of

> *Rule Consequentialism*: Everyone ought to follow the principles whose universal acceptance would make things go best.

Such principles we can call *optimific*.

Kantians could argue:

> Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.
>
> Everyone could rationally choose whatever they would have sufficient reasons to choose.
>
> There are some optimific principles.
>
> These are the principles that everyone would have the strongest impartial reasons to choose.
>
> No one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons.

Therefore

> Everyone would have sufficient reasons to choose these optimific principles.
>
> There are no other significantly non-optimific principles that everyone would have sufficient reasons to choose.

Therefore

> It is only these optimific principles that everyone would have sufficient reasons to choose.

Therefore

> Everyone ought to follow these principles.

This argument's first premise is the Kantian Contractualist Formula. The argument is valid, and its other premises are true. So this Kantian Formula requires us to follow these Rule Consequentialist principles.

This Kantian Argument, we may suspect, must have at least one Consequentialist premise. If that were true, this argument would have no importance. But none of this argument's premises assumes the truth of Consequentialism. Here is how, without any such premise, this argument validly implies a Consequentialist conclusion:

> Consequentialists appeal to claims about what it would be rational for everyone to choose from an impartial point of view. The strongest objections to Consequentialism are provided by some of our intuitive beliefs about which acts are wrong.
>
> Contractualists appeal to claims about what it would be rational for everyone to choose, in some way that would make these choices impartial. In Contractualist moral reasoning, we cannot appeal to our intuitive beliefs about which acts are wrong.

Since both kinds of theory appeal to what it would be rational for everyone impartially to choose, and Contractualists tell us to ignore our non-Consequentialist moral intuitions, we should expect that valid arguments with some Contractualist premise could have some Consequentialist conclusion.

We can draw another conclusion. There are, I have claimed, some decisive objections to Kant's Formula of Universal Law. To avoid these objections, Kant's Formula must be revised. In its best revised form, this formula requires us to follow the principles whose universal acceptance everyone could rationally will, or choose. There are no significantly non-optimific principles that everyone could rationally choose. So this formula cannot succeed unless it is true that, as I have argued, everyone could rationally choose the optimific principles. Kant's Formula of Universal Law cannot succeed unless, in this revised form, this formula implies Rule Consequentialism.

CHAPTER 17 CONCLUSIONS

63 *Kantian Consequentialism*

According to the Act Consequentialist principle, everyone ought always to do whatever would make things go best. This is not one of the principles whose universal acceptance would make things go best. So the Kantian Formula does not require us to be Act Consequentialists.

According to another version of the Kantian Formula, everyone ought to follow the principles whose being universally *followed*, or *successfully* acted upon, everyone could rationally will, or choose. This version of the Kantian Formula implies a version of Rule Consequentialism that is significantly closer to Act Consequentialism.

Since Kantian Contractualism implies Rule Consequentialism, these theories can be combined. Principles can be universal laws by being either universally accepted or universally followed. According to

> *Kantian Rule Consequentialism*: Everyone ought to follow the principles whose being universal laws would make things go best, because these are the only principles whose being universal laws everyone could rationally will.

64 *Climbing the Mountain*

When there is only one set of principles that everyone could rationally will to be universal laws, these are the only principles, we can argue, that no one could reasonably reject. If that is true, this combined theory could also include Scanlon's Formula. According to what we can call this

> *Triple Theory*: An act is wrong just when such acts are disallowed by the principles that are optimific, uniquely universally willable, and not reasonably rejectable.

If we accept this theory, we should admit that acts can have other properties that make them wrong. The Triple Theory should claim to describe a single complex higher-level property under which all other

wrong-making properties can be subsumed. If this theory succeeds, it would describe what these other properties have in common.

This theory may succeed, since it has many plausible implications. The Kantian and Scanlonian Formulas are also in themselves plausible. Of this theory's three components, Rule Consequentialism is, in one way, the hardest to defend. Some Rule Consequentialists appeal to the claim that

(Q) all that ultimately matters is how well things go.

This claim is in itself very plausible. If we reject (Q), that is because this claim supports Act Consequentialism, which conflicts too often, or too strongly, with some of our intuitive beliefs about which acts are wrong. Rule Consequentialism conflicts much less often and less strongly with these intuitive beliefs. But if Rule Consequentialists appeal to (Q), their view faces a strong objection. On this view, it is wrong to do what is disallowed by the optimific principles even when we know that our acts would make things go best. We can plausibly object that, if all that ultimately matters is how well things go, such acts cannot be wrong.

Kantian Rule Consequentialism avoids this objection. On this view what is fundamental is not this belief about what ultimately matters, but the belief that we ought to follow the principles whose being universal laws everyone could rationally will.

Of our reasons for doubting that there are moral truths, one of the strongest is provided by some kinds of moral disagreement. If we and others hold conflicting views, and we have no reason to believe that *we* are the people who are more likely to be right, that should at least make us doubt our view. It may also give us reasons to doubt that any of us could be right.

It has been widely believed that there are such deep disagreements between Kantians, Contractualists, and Consequentialists. That, I have argued, is not true. These people are climbing the same mountain on different sides.

APPENDIX A STATE-GIVEN REASONS

When certain facts would make it better if we had a certain belief, these facts give us object-given reasons to *want* to have this belief, and to *cause* ourselves to have it, if we can. There is no point in adding that we would also have *state-given* reasons to *have* this belief. Though we cannot now respond to such alleged reasons, our psychology might change. When we believed that it would be better if we had some epistemically irrational belief, we might find ourselves coming to have this belief in a direct non-voluntary way. But this should not be regarded as a response to state-given reasons. Nor could such reasons ever conflict with our epistemic reasons. It is more plausible to claim that, when certain facts would make it better if we had some desire, these facts give us a reason to have this desire. But we also have strong reasons to reject this claim.

APPENDIX B RATIONAL IRRATIONALITY AND GAUTHIER'S THEORY

Gauthier claims that, when we have rationally caused ourselves to have some disposition, it would be rational for us to act upon it. This claim has several implausible implications. Though it might be rational to cause ourselves to believe that it would be rational to act on such dispositions, this fact could not show that this belief is true. Gauthier also claims that, if we accept a Hobbesian version of Contractualism and a minimal version of morality, his argument shows that we are rationally required never to act wrongly. Since this argument fails, it gives us no reason to accept Gauthier's minimal morality.

APPENDIX C DEONTIC REASONS

In defending the Kantian Argument for Rule Consequentialism, I suggest that

> (X) if the optimific principles require certain acts that we believe to be wrong, we would not have decisive *non*-deontic reasons to act in these ways. Any such decisive reasons would

> have to be *deontic*, in the sense of being provided by the wrongness of these acts.

When some people claim that some act is wrong, these people mean that we have decisive moral reasons not to act in this way. These people would deny that there are any deontic reasons. On this view,

> (2) when some act is wrong, this fact is the second-order fact that certain other facts give us decisive moral reasons not to act in this way, and the fact that we had these reasons would not give us a further reason not to act in this way.

If (2) were true, (X) would be partly undermined. Given what most of us mean by 'wrong', however, we can justifiably reject (2). And (2) is least plausible in the very cases to which (X) most importantly applies.

PART ONE

REASONS

This page intentionally left blank

1

Normative Concepts

1 Normative Reasons

We are the animals that can both understand and respond to reasons. These abilities have given us great knowledge, and power to control the future of life on Earth. Though there may be life elsewhere, there may be no other animals like us. We may be the only rational beings in the Universe.

We can have reasons to believe something, to do something, to have some desire or aim, and to have many other attitudes and emotions, such as fear, regret, and hope. Reasons are given by facts, such as the fact that someone's finger-prints are on some gun, or that calling an ambulance would save someone's life.

It is hard to explain the *concept* of a reason, or what the phrase 'a reason' means. Facts give us reasons, we might say, when they count in favour of our having some attitude, or our acting in some way. But 'counts in favour of' means roughly 'gives a reason for'. Like some other fundamental concepts, such as those involved in our thoughts about time, consciousness, and possibility, the concept of a reason is indefinable in the sense that it cannot be helpfully explained merely by using words. We must explain such concepts in a different way, by getting people to think thoughts that use these concepts. One example is the thought that we always have a reason to want to avoid being in agony.

We can have reasons, I shall say, of which we are unaware. Suppose that I ask my doctor, 'Since I'm allergic to apples, do I have any reason

not to eat any other kind of food?' If my doctor knows that walnuts would kill me, her answer should be Yes. This fact gives me a reason.

Rather than saying that certain facts *give* us reasons, some people say that these facts *are* reasons for us. And some people say that, to *have* some reason, we must be aware of the fact that gives us this reason. But these people's claims do not conflict with mine, since these are merely different ways of saying the same things. My doctor might say, 'No, you don't have any reason not to eat any other kind of food, but you will have such a reason after I've told you that eating walnuts would kill you'. It is simpler to say that I already have this reason.

When we must choose between different possible acts, our reasons may conflict, and they can differ in what we can call their force, strength, or weight. If I enjoy walnuts, this fact gives me a reason to eat them; but, if they would kill me, this fact gives me a stronger or weightier conflicting reason *not* to eat them. When we have several reasons to act in some way, these reasons may *together* be stronger than, or outweigh, some single stronger conflicting reason. If I could either save you from ten hours of pain, or do something else that would both save you from nine hours of pain and save someone else from eight hours of pain, I would have a stronger set of reasons to act in this second way. As we can more briefly say, I would have *more reason* to act in this way.

If our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, these reasons are *decisive*, and acting in this way is what we have *most reason* to do. If such reasons are much stronger than any set of conflicting reasons, we can call them *strongly* decisive. Though most kinds of reason are decisive only in certain cases, there may be some kinds of reason that are always decisive. On some views, for example, we always have decisive reasons not to act wrongly.

When we are aware of facts that give us decisive reasons to act in some way, we *respond* to these reasons if our awareness of these facts leads us to do, or try to do, what we have these reasons to do. If we ignore these reasons, we are not responding to them, just as ignoring someone's cry for help is not responding to this cry.

There is often nothing that we have decisive reasons to do, or *most* reason to do, because we have *sufficient* reasons, or *enough* reason,

to act in any of two or more ways. Our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways. We might have sufficient reasons, for example, to eat either a peach or a plum or a pear, to choose either law or medicine as a career, or to give part of our income either to Oxfam or to some other similar aid agency, such as Médecins Sans Frontières. When neither of two conflicting reasons is stronger, that is seldom because these reasons are precisely equally strong. Though there are truths about the relative strength of different reasons, these truths are often very imprecise.

Reasons can be related in more complicated ways. Some facts give us reasons, for example, to ignore some other reasons. If I am judging who deserves some prize, that would give me a reason to ignore the fact that one of the contestants is my best friend. And some facts give us reasons, not in all cases, but only when combined with certain other facts. I shall mainly be discussing simpler reasons.

When we have decisive reasons, or most reason, to act in some way, this act is what we *should* or *ought* to do in what we can call the *decisive-reason-implying* senses. Even if we never use the phrases 'decisive reason' or 'most reason', most of us often use 'should' and 'ought' in these reason-implying senses. There is a similar sense of 'must'. These words imply reasons of different strengths. I might say that you *should* see some film, that you *ought* to give up smoking, and that you *mustn't* touch some live electric cable. Though the word 'should' is used more often, and the word 'must' has more force, I shall mostly use the less ambiguous word 'ought'.

As well as asking what we ought to do in the decisive-reason-implying sense, we can ask what we ought *rationally* to do. When we call some act 'rational', using this word in its ordinary, non-technical sense, we express the kind of praise or approval that we can also express with words like 'sensible', 'reasonable', 'intelligent', and 'smart'. We use the word 'irrational' to express the kind of criticism that we express with words like 'senseless', 'stupid', 'idiotic', and 'crazy'. To express weaker criticisms of this kind, we can use the phrase 'less than fully rational'.

When we must choose between several possible acts, there may be several facts that give us reasons to act in these ways. I shall call these the *relevant, reason-giving* facts. What we ought rationally to do depends in part on our beliefs about these facts. These beliefs may include assumptions of which we are not consciously aware—such as the assumption that we would not harm ourselves or others if we eat a walnut, or touch some electric cable, or push open some swinging door. If we have certain beliefs about the relevant facts, and what we believe would, if it were true, give us a reason to act in some way, I shall call these *beliefs whose truth* would give us this reason. In most cases, I believe, some possible act of ours would be

> *rational* if we have beliefs about the relevant facts whose truth would give us sufficient reasons to act in this way,
>
> what we *ought rationally* to do if these reasons would be decisive,
>
> *less than fully rational* if we have beliefs whose truth would give us clear and decisive reasons *not* to act in this way,

and

> *irrational* if these reasons would be strongly decisive.

On this view, when we know all of the relevant facts, what we ought rationally to do is the same as what we ought to do in the decisive-reason-implying sense. But when we are ignorant or have false beliefs, these *oughts* may conflict. Suppose that, while walking in some desert, you have disturbed and angered a poisonous snake. You believe that, to save your life, you must run away. In fact you must stand still, since this snake will attack only moving targets. Given your false belief, it would be irrational for you to stand still. You ought rationally to run away. But that is *not* what you ought to do in the decisive-reason-implying sense. You have no reason to run away, and a decisive reason *not* to run away. You ought to stand still, since that is your only way to save your life.

Some people would say that you do have a reason to run away, which is provided by your false belief that this act would save your life. But if

we say that false beliefs can give people reasons, we would need to add that these reasons do not have any *normative force*, in the sense that they do not count in favour of any act. And we would have to ignore such reasons when we are trying to decide what someone has most reason to do. It is better to describe such cases in a different way. When we have beliefs whose truth would give us a reason to act in some way, we have what I shall call an *apparent* reason to act in this way. If these beliefs are true, this apparent reason is also a *real* reason. If these beliefs are false, we have what *merely* appears to be a reason. In the case of the angry snake, given your false belief that running away would save your life, you have a *merely apparent* reason to run away. We have a different kind of apparent reason when we believe that we have some reason. We can now claim that all reasons have normative force. When we give people advice, we can ignore the merely apparent reasons that are provided by these people's false beliefs. But what it would be *rational* for people to do depends on their *apparent* reasons, whether or not these reasons are real, or merely apparent.

We can now turn from possible to actual acts. I believe that, in most cases, we act

> *rationally* if we act in some way because we have beliefs about the relevant facts whose truth would give us sufficient reasons to act in this way,

and

> *irrationally* if we act in some way despite having beliefs whose truth would give us clear and strongly decisive reasons *not* to act in this way.

Such an act would be most irrational if these beliefs are conscious. When these reasons would be less clear, or would be only weakly decisive, our act may be only less than fully rational. It would be irrational, for example, to start smoking, knowing that we shall be likely to become addicted and shorten our lives. It would be merely less than fully rational to buy some book that we know we won't read, or to try to ring up some phone service to report that our phone isn't working.

It is worth explaining why, though it is facts that give us reasons, the rationality of our acts depends instead on our beliefs. When we are trying to decide what we or others ought to do, what matters are the reason-giving facts. In the case of the angry snake, you ought to stand still because that is in fact your only way to save your life. When we ask whether someone has acted rationally, we have a different aim. We are asking whether this person deserves the kind of criticism that we express with words like 'foolish', 'stupid', and 'crazy'. When people are ignorant, or have false beliefs, they may do what they ought *not* to do in the decisive-reason-implying sense. But these people may not deserve any criticism, since they may have false beliefs whose truth would have given them sufficient reasons to act as they do. At least in most cases, that is enough to make their act rational. If you ran away from the snake because you believed falsely that this act would save your life, your fatal act wouldn't be foolish, stupid, or crazy. You would merely be very unlucky.

For us to be acting rationally, many people claim, it is not enough that we are acting on beliefs whose truth would give us sufficient reasons to act as we do. Our act is rational only if our beliefs are rational. This is not, I shall argue later, the best view.

To be fully rational, we may also need to meet certain other *rational requirements*, by avoiding certain kinds of inconsistency and other mismatch between our intentions, beliefs, and other mental states. We may be rationally required, for example, not to have contradictory intentions, and to intend to do what we believe that we ought to do. Though these requirements raise several interesting questions, I shall say little about them. Questions about reasons are, I believe, more fundamental. And while it often matters greatly whether we are wanting what we have reasons to want, and doing what we have reasons to do, it seldom matters, or matters much, whether we are being inconsistent and thereby failing to meet some rational requirement. Some people claim that, to be rational, we don't need to respond to reasons or apparent reasons, since it is enough to meet these rational requirements. I shall later give some arguments against this view.

There are some other, similar questions that I shall mention briefly and then set aside. When we are deciding what to do, and we don't know all of the relevant facts, we must base our decision on what we believe,

and on the available evidence. In such cases, we can ask what we *should* or *ought* to do in what we can call the *evidence-relative* senses. It may seem that, in such cases, we ought to try to do what we have most reason to do. But such attempts may be too risky, or too unlikely to succeed. We often ought to act in ways that are more likely to achieve less ambitious aims. If many people's lives are in danger, for example, we ought to do what would certainly save most of these people, rather than doing what has only a small chance of being the act that would save them all.

It is of great practical importance what we ought to do in cases that involve risk or uncertainty. These questions have been well discussed by many philosophers, decision theorists, and others. Certain other questions about reasons, though more fundamental, have been less well discussed. These are also questions about which people disagree more deeply. Since I shall be mainly discussing these questions, I shall mostly consider cases in which we know all of the relevant, reason-giving facts.

These claims have been about about *normative* reasons. When we have such a reason or apparent reason, and we act *for this reason*, this becomes our *motivating* reason. If I avoid walnuts, for example, my motivating reason might be that, as my doctor has told me, eating them would kill me. This distinction is clearest when we have only a motivating reason for acting in some way. If you ran away from the angry snake, your motivating reason would be provided by your false belief that this act would save your life. But, as I have said, you have no normative reason to run away. You merely think you do. In an example of a different kind, we might claim: 'His reason was to get revenge, but that was no reason to do what he did'. Since I shall not be discussing why people act as they do, I shall say little about motivating reasons.

As well as asking what we ought to do in the decisive-reason-implying sense, and what we ought rationally to do, we sometimes ask what we ought to do in one of several *moral* senses. Most of these senses differ in at least two ways from the decisive-reason-implying sense. First, we often have decisive reasons that are not moral reasons. If I need to catch some train, for example, I may have a decisive reason to leave some meeting now. If I hate commuting, I may have most reason to live close to where I work. These may not be things that I ought morally to do. Second, when we believe that we ought morally to act in some way, we

may not believe that we have decisive reasons to act in this way. On some views, we might have *no* reason to do what we ought morally to do. In these chapters I shall first discuss reasons, turning only later to morality.

It is easy to confuse the decisive-reason-implying sense of 'ought' either with 'ought rationally' or with 'ought morally'. So rather than discussing what we ought to do in the decisive-reason-implying sense, I shall often discuss what we have decisive reasons, or most reason, to do.

2 Reason-Involving Goodness

We can next consider some ways in which things can be *good* or *bad*. When we call something

> *good*, in what we can call the *reason-implying* sense, we mean roughly that there are certain kinds of fact about this thing's nature, or properties, that would in certain situations give us or others strong reasons to respond to this thing in some positive way, such as wanting, choosing, using, producing, or preserving this thing.

Some book may be good, for example, by being enjoyable, or inspiring, or containing useful information. Some medicine may be the best by being the safest and the most effective. These facts may give us or others reasons to read this book, or to take this medicine. There are similar senses of 'better', 'bad', 'worse', and 'worst'.

Things can be good or bad in other senses. If I claimed, for example, that some tree has good roots, that moles have bad eye-sight, or that the best metaphor is

> Ice formed on the butler's upper slopes,

and the best palindrome is not 'Madam I'm Adam' but

> A MAN A PLAN A CANAL: PANAMA,

I would not intend these uses of 'good', 'bad', and 'best' to be reason-implying. Moles could not have reasons to wear spectacles, nor do we

have reasons to be amused by the ice on the butler's upper slopes. And many uses of 'good' mean only that something meets certain standards. But the most important uses of 'good' and 'bad' are, I believe, reason-implying.

When something is in this sense good, Scanlon claims, this thing's goodness could not give us reasons. Such goodness is the property of having *other* properties that might give us certain reasons, and the second-order fact that we had these reasons would not itself give us any reason not to act in this way.

This view needs, I think, one small revision. If some medicine or book is the best, these facts could be truly claimed to give us reasons to take this medicine, or to read this book. But these would not be *further, independent* reasons. These reasons would be *derivative*, since their normative force would derive entirely from the facts that made this medicine or book the best. That is why it would be odd to claim that we had *three* reasons to take some medicine: reasons that are given by the facts that this medicine is the safest, the most effective, *and* the best. Since such derivative reasons have no independent normative force, it would be misleading to mention them in such a claim.

Of our reasons for acting, many are provided by facts about what would be

> *good for us*, in the sense of being in our interests, benefiting us, or contributing to our well-being.

When people say that something would be good for us, or in our interests, these people often mean that this thing would have good effects on our health, character, or bank balance. In my intended wider sense, something is *intrinsically* or in itself good for us if it is one of the features of our lives in which our well-being consists, because these are the features that make our lives worth living. Something is *instrumentally* good for us if it has effects that are intrinsically good for us. On *hedonistic* theories, our well-being consists, roughly, in pleasure and happiness, and avoiding pain and suffering. On theories that appeal to *substantive goods*, our well-being may also partly consist in some other states or activities, such as loving and being loved, being morally

good and acting well, and various other kinds of achievement. On *desire-based* theories, our well-being consists in the fulfilment of some of our desires, such as our informed desires about our own life. On any plausible theory, hedonism covers at least a large part of the truth, so my examples will often involve hedonic well-being.

We have *self-interested* reasons to care about our own well-being, and *altruistic* reasons to care about the well-being of other people. These are reasons to want certain things to happen for our own sake, or for the sake of these other people. 'Self-interested' does not mean 'selfish'. Even the most unselfish people have self-interested reasons, since they have reasons to care about their own future well-being.

We can have strong reasons to care about the well-being of certain other people, such as our close relatives and other people whom we love. Like self-interested reasons, these altruistic reasons are

> *person-relative* or *partial* in the sense that these are reasons to be specially concerned about the well-being of people who are in certain ways *related to us*.

We also have some reasons, I believe, to care about everyone's well-being. Such reasons are

> *impartial* in the sense that
>
> > (1) these are reasons to care about anyone's well-being whatever that person's relation to us,
>
> so that
>
> > (2) we would have these reasons even if our situation gave us an impartial point of view.

I use the phrase 'point of view' in something close to its literal sense, not the looser sense in which we talk of the reasons that we might have from a financial, aesthetic, or other such point of view. We have an impartial point of view when we are considering possible events that would affect or involve people who are all strangers to us. When our actual point of view is not impartial, we can think about possible events from an imagined impartial point of view. We can do that by imagining possible

events that are relevantly similar, except that the people involved are all strangers to us.

We have impartial reasons, I believe, to care equally about everyone's well-being. That is a substantive belief, not something that is implied by my definition of an impartial reason. On some other, widely held views, we have impartial reasons to care more about the well-being of certain kinds of people, such as those who are morally good, or those who have the greatest abilities. When our *point of view* is impartial, that does not ensure that *we* are impartial. We might care more about the well-being of certain strangers, such as those who are more similar to us, or those whose faces we like. But we would have no *reasons*, I believe, to care more about the well-being of these people.

We can next describe two ways in which events can be good or bad. When we call some possible event

> *good for someone*, in the *reason-implying* sense, we mean that there are certain facts that give this person self-interested reasons to want this event to occur, and that give other people altruistic reasons to want or hope, for this person's sake, that this event will occur.

This definition may seem to tell us little, since it refers to *self-interested* reasons. As we shall see, however, it is controversial whether we have any such reasons.

When we call one of two events

> *better* in the *impartial-reason-implying* sense, we mean that everyone would have, from an impartial point of view, stronger reasons to want this event to occur, or to hope that it will.

It would be in this sense better, I believe, if some plague or earthquake killed fewer people, or if any person or other animal ceased to be in pain. This kind of goodness is *impersonal* in the sense that, when we call some event in this sense good, we don't mean that this event would be *good for* some person or group of people. But many events are impersonally

good because they are good for one or more people. The benefits to these people are what make these events impersonally good. And since everyone has reasons to want such events to occur, such impersonal goodness involves *omnipersonal* reasons.

If some possible event would be in these senses good for someone, or impersonally good, this fact could be truly claimed to give us a reason to want this event to occur. But as before, this reason would be derivative, since this reason's force would derive from the facts that would make this event good for this person, or impersonally good. When we use 'good for' and 'good' in these senses, these are merely briefer ways of implying that there are such other, reason-giving facts. Unlike the concept of *a reason*, and the decisive-reason-implying concept *should* or *ought*, these versions of the concept *good* are not fundamental.

On some widely accepted views about reasons, no events could be in these senses either good or bad for particular people, or impersonally good or bad. If such a view were true, that would greatly affect what we had most reason to want, and to do. But we ought, I shall argue, to reject such views.

2

Objective Theories

3 Two Kinds of Theory

The word 'desire' often refers to our sensual desires or appetites, or to our being attracted to something, by finding the thought of it appealing. I shall use 'desire' in a wider sense, which refers to any state of being motivated, or of wanting something to happen and being to some degree disposed to make it happen, if we can. The word 'want' already has both these senses. If you and I were planning how we shall spend some day together, I might say without self-contradiction, 'I want us to do, not what *I* want us to do, but what *you* want us to do'. What I want, in the *wide* sense, is not what *I* want but what *you* want, in the *narrow* sense. I want us to do what *you* are attracted to, or find appealing, even if it doesn't appeal to me.

Some people think: 'Whenever people act voluntarily, they are doing what they want to do. Doing what we want is selfish. So everyone always acts selfishly.' This argument for *Psychological Egoism* fails, because it uses the word 'want' first in the wide sense and then in the narrow sense. If I voluntarily gave up my life to save the lives of several strangers, my act would not be selfish, though I would be doing what in the wide sense I wanted to do.

Our desires have *objects*, which are *what* we want. These objects are all *events* in the sense that includes acts, processes, and states of affairs. We can be correctly said to want things of other kinds. I might want an apartment in Venice, a glass of water, and a piano teacher. Some fugitive may be wanted by the police. But what we really want is to own,

live in, drink, be taught by, find, use, or have some other relation to some thing or person. Rather than saying that we want some event *to occur*, I shall say, for short, that we want this event.

Our desires are *teleological* or *telic* when we want some event as an *end*, or for its own sake. Our desires are *instrumental* when we want some event as a *means*, because this event would or might cause some other event that we want. We want some acts or other events both as an end and as a means to some other end. Two such events might be a thrilling search for some important truth, and, when we want to have a child, making love. When we decide to try to fulfil some telic desire, we thereby make this desire's fulfilment one of our *aims*.

We often have long chains of instrumental desires, but such chains all begin with, or are grounded on, some telic desire. I might want medical treatment, for example, not for its own sake, but only to restore my health, and I might want health only so that I can finish writing some great novel, and I might want to finish this novel only to achieve posthumous fame. This desire might also be purely instrumental, since I might want to achieve such fame only to refute my critics, or to increase the income of my heirs. But if I want posthumous fame for its own sake, this telic desire would begin this particular chain.

Psychological Hedonists claim that, at the beginning of all such chains of instrumental desires, there is some telic desire for pleasure, or the avoidance of pain. That is false. Of those who hold this view, some confuse it with the view that we always get pleasure in advance from the thought of our desire's fulfilment, or are pained by the thought of its non-fulfilment. That is also false. And even if it were true, that would not show that what we really want is always to get pleasure, or avoid pain. If I want posthumous fame, for example, I may get pleasure from thinking about how, after my death, people will remember me and admire my great novel. But that would not show that I want such fame for the sake of this pleasure. On the contrary, this pleasure would depend on my wanting such fame for its own sake. Another example is the fact that, to enjoy many games, it is not enough to want to enjoy them, since we shall enjoy these games only if we also want to win.

As well as wanting such other things, some people do not even want pleasure as an end. Suppose that we know some relentlessly ambitious

politician, whom we find basking in the sun, sipping champagne. When we ask this man what he is doing, he replies 'Enjoying myself'. Given our knowledge of this man's character, this reply is baffling. This man never does anything merely for enjoyment. He then explains that his doctor warned that, unless he allows himself such pleasures, his health will worsen, thereby hindering his pursuit of power. Our bafflement disappears. This man wants these pleasures, not for their own sake, but only as a means.

There are two main kinds of view about what I shall call *practical* reasons. According to one group of views, there are certain facts that give us reasons both to have certain desires and aims, and to do whatever might achieve these aims. These reasons are given by facts about the *objects* of these desires or aims, or what we might want or try to achieve. We can therefore call such reasons *object-given*. If we believe that all practical reasons are of this kind, we are *Objectivists about Reasons*, who accept or assume some *objective* theory.

Object-given reasons are provided by the facts that make certain outcomes worth producing or preventing, or make certain things worth doing for their own sake. In most cases, these reason-giving facts also make these outcomes or acts good or bad for particular people, or impersonally good or bad. So we can also call these objective reasons and theories *value-based*.

According to another group of theories, our reasons for acting are all provided by, or depend upon, certain facts about what would fulfil or achieve our present desires or aims. Some of these theories appeal to our actual present desires or aims. Others appeal to the desires or aims that we would now have, or to the choices that we would now make, if we had carefully considered all of the relevant facts. Since these are all facts about *us*, we can call these reasons *subject-given*. If we believe that all practical reasons are of this kind, we are *Subjectivists about Reasons*, who accept some *subjective* theory.

These two kinds of theory are very different. According to Objectivists, though many reasons for acting can be claimed to be given by the fact that some act would achieve one of our aims, these reasons *derive their force* from the facts that give us reasons to *have* these aims. These are the facts that make these aims relevantly good, or worth achieving.

According to Subjectivists, we have no such reasons to have our aims. Some Subjectivists even claim that it is *we* who, with our desires or choices, make things good. While defending such a view, for example, Korsgaard writes:

> most things are good because of the interest human beings have in them . . . Objectivism reverses this relation . . . Instead of saying that what we are interested in is therefore good, the objectivist says that the goodness is in the object, and we ought therefore to be interested in it.

Such goodness would give us reasons in the way the sun gives light, 'because it's out there, shining down'. If Subjectivism is true, we must make our choices in the dark.

Subjectivists and Objectivists often partly agree. According to all plausible objective theories, we have reasons to try to promote our future well-being. Since most of us want to promote our future well-being, subjective theories also imply that most of us have reasons to act in this way. And most of us have many other desires that both kinds of theory tell us to try to fulfil, since what we want is often something that is worth having or achieving.

Though theories of both kinds often agree that we have reasons to try to fulfil our present desires, these theories often disagree about which of these desires we have *stronger* reasons to try to fulfil. On many subjective theories, the strength of these reasons depends on the strength of these desires, or on our preferences. On objective theories, the strength of these reasons depends instead on how good, or worth achieving, the fulfilment of these desires would be. Many of us often have stronger desires for what would be less worth achieving. Many such cases involve an attitude to time that we can call *the bias towards the near*. We may prefer to have enjoyable experiences in the nearer future, though we know that, if we waited, our enjoyment would be greater. We may prefer to postpone some tedious chore, or unavoidable ordeal, though we know that this postponement will only make this chore more tedious or this ordeal more painful. And we may choose to spend all our money now, though we know that some of this money would later bring us greater benefits. By fulfilling such desires and preferences, many of us

make our lives go worse. In these and many other ways, subjective and objective theories often disagree about what we have *most* reason to do.

There are other, deeper disagreements. As we shall see, theories of either kind can imply that we have decisive reasons to do something, though theories of the other kind imply that we have *no* reason to do this thing, and have decisive reasons *not* to do it. And these two kinds of theory wholly disagree about our reasons to have our desires and aims.

We ought, I shall argue, to accept some value-based, objective theory. On these theories, reasons for acting all derive their force from the facts that give us reasons to have certain desires and aims. These other reasons are more fundamental.

4 Responding to Reasons

The same facts can give us reasons both to want something to happen and to try to make it happen by acting in some way. That is why I call both kinds of reason *practical*. Though these two kinds of reason are very closely related, there is a striking difference between the ways in which we can respond to them. When we are aware of facts that give us reasons to act in some way, we can often respond to these reasons by acting in this way. This response is voluntary in the sense that, if we had wanted not to act in this way, we could have chosen not to do so. But when we are aware of facts that give us strong reasons to have some desire, our response to these reasons is seldom voluntary. It is seldom true that, if we had wanted not to have such desires, we could have chosen not to have them. We could seldom choose, for example, whether we want to stay alive, or to avoid great pain. If some whimsical despot threatens to kill me unless, one minute from now, I want to be killed, I could not choose to have this desire.

Similar claims apply to our *epistemic* reasons to have particular beliefs. These reasons are provided by facts that are related to the *truth* of some belief, by being evidence for its truth, or by logically implying this belief, or in some other way. If we see dark grey clouds, for example, that gives us some reason to believe that it will soon rain. If we know that gold weighs more than lead, which weighs more than iron, these facts give us a decisive reason to believe that gold weighs more than iron. When we

are aware of facts that give us decisive reasons to have some belief, we can respond to these reasons by coming to have and continuing to have this belief. But our responses to such reasons are seldom voluntary. We could seldom choose *not* to believe what we have such decisive reasons to believe. If my imagined despot threatens to kill me unless, one minute from now, I no longer believe that $2 + 2 = 4$, I could not choose to lose this belief.

Some writers claim that, when we come to have some belief or desire in this direct non-voluntary way, this is an act, or something that we do. But I shall use 'act' and 'do' more narrowly, to refer only to *voluntary* acts. Many such acts are purely mental. If you find yourself asking, for example, whether you still have enough time to catch some train, you might voluntarily do a mental calculation to answer this question. With this complex act you would intentionally bring it about that you come to have *some* belief about this question. But if this calculation leads you to believe that you don't have enough time to catch your train, your coming to have this *particular* belief would not be an act, or be voluntary. You could not, for example, choose to believe that you would be able to catch your train by running ten miles in ten minutes.

Though we can seldom choose how we respond to our reasons to have particular beliefs and desires, our responses to these reasons are not things that merely happen to us, like an automatic knee-jerk, or our slipping on a banana skin. Our being rational consists in part in our responding to such reasons or apparent reasons in these non-voluntary ways. We can be asked *why* we believe something, or want something, and we can often give our reasons.

It is worth asking whether our responses to such reasons might take other forms, by being always or often voluntary. Suppose that, when you are aware of certain facts that give you decisive epistemic reasons to have some belief, you fail to respond to these reasons in the rational non-voluntary way, by coming to have this belief. Though you can see smoke and flames rising towards you up the stairs inside your hotel, you fail to believe that your life is in danger. Could you correct your mistake, by choosing to have this belief?

The answer is likely to be No. Suppose first that, as well as failing to believe that your life is in danger, you also fail to believe that the

smoke and flames give you any reasons to have this belief. You could not then correct your mistake, since you would not believe that you had made any mistake. You could not choose to believe, for these epistemic reasons, that your life is in danger, since you would not believe that you had these reasons.

Suppose instead that you do believe that the smoke and flames give you decisive reasons to believe that your life is in danger. It is unlikely that you could then choose to believe that your life is in danger. In most cases, in coming to believe that we have decisive epistemic reasons to have some belief, we also come to have this belief. And when we already have some belief, we cannot choose to have it.

There might be exceptions. Suppose next that, though you believe that the smoke and flames give you decisive reasons to believe that your life is in danger, you don't yet have this second belief. We can perhaps imagine that you could then choose to make yourself have this belief for these reasons. But your response to these epistemic reasons would still be only partly voluntary. When you saw that smoke and flames were rising up the stairs in your hotel, you did not choose to believe that these facts gave you decisive reasons to believe that your life is in danger.

There are other reasons why our responses to most epistemic reasons could not be voluntary. For us to have knowledge of the world around us, our beliefs must be reliably caused by our visual and other perceptual experiences, or by our awareness of other facts that give us epistemic reasons to have these beliefs. Such causation could not be reliable if we could freely choose all of our beliefs. And to have knowledge of necessary truths, such as logical or mathematical truths, we must also respond to some epistemic reasons in rational but non-voluntary ways, by recognizing or realizing what follows from what, and what must be true.

Similar claims apply to our desires and preferences. We can seldom choose what it is that we want or prefer, because we cannot choose either what we have reasons to want, or how strong these reasons are. What we can choose is only which of our desires we adopt as aims, and try to fulfil. Our responses to these reasons might become somewhat more voluntary than they are now. That would be, in some ways, better, since we could then more easily transform our desires, attitudes, and emotions, by

making ourselves become the kind of person that we have reasons to want to be. We might be able to ensure, for example, that we shall never lose our youthful ideals. But such abilities would also be dangerous, like our recently discovered mechanical ways of moving our bodies at great speed. If we changed ourselves for the worse, our new, deliberately chosen desires might lead us not to undo such mistakes.

5 State-Given Reasons

Our reasons to have some desire are provided, I have claimed, by facts about this desire's *object*, or the event that we want. Such reasons I am calling *object-given*. Many people assume that we can also have *state-given* reasons to have some desire. Such reasons would be provided by certain facts, not about some desire's object, but about our state of having this desire. We would have such reasons when our having some desire would be in some way good, either as an end or as a means.

On this view, we can have at least four kinds of reason to have some desire, which can be described as follows:

	telic and intrinsic	instrumental
object-given	The event that we want would be in itself good, or worth achieving	This event would have good effects
state-given	Our wanting this event would be in itself good	Our wanting this event would have good effects

We might have reasons of all these kinds to have the same desire. If you are in pain, for example, I might have all these reasons to want your pain to end. What I want would be in itself good, and it might also have the good effect of allowing you to enjoy life again. My wanting your pain to end might be in itself good, and this desire might also have good effects, such as your being comforted by my sympathy.

Similar claims apply to our reasons to have beliefs. Since our epistemic reasons are related to the truth of *what* we believe, these reasons can also be called *object-given*. Many people assume that we can also have *state-given* reasons to have certain beliefs. Such reasons would be provided

by facts that would make our *having* some belief in some way good. It is often claimed, for example, that we have such reasons to believe that God exists and that we shall have a life after death. These reasons would not be epistemic, or truth-related, but *goodness-related*, or *value-based*. Such alleged reasons to have beliefs are sometimes called *practical* or *pragmatic*.

If we can have such state-given reasons, these reasons would not, I believe, have any importance. When it would be better if we were in some state, we would have reasons to want to be in this state. If we could cause ourselves to be in this state, we would have reasons to act in this way. It is not worth claiming that, as well as having reasons to *want* to be and to *cause* ourselves to be in this state, we would also have reasons to *be* in this state. Suppose for example that I would be healthier and happier if I weighed less, owned a bicycle, knew how to dance, and had some friends. These facts would give me reasons to want and to try to make myself lose weight, to buy a bicycle, to learn how to dance, and to make some friends. It is not worth claiming that, as well giving me reasons to act in these ways, these facts would give me reasons to weigh less, to own a bicycle, to know how to dance, and to have some friends. Such reasons would make no difference.

Suppose next that, though it would be better if we were in a certain state, we could not possibly cause ourselves to be in this state. We would then have reasons to wish that we were in this better state. I might have reasons, for example, to wish that I were ten inches taller, twenty years younger, and could run faster than a cheetah. We needn't claim that I would also have reasons to *be* ten inches taller, to *be* twenty years younger, and to *be able* to run faster than a cheetah. And such claims may be clearly false. Reasons are things to which at least some people might be able to respond, and no one could respond to a reason to be twenty years younger.

Similar claims apply to our beliefs and desires. When it would be better for us if we had some belief or desire, we have object-given reasons to want to have this belief or desire, and to cause ourselves to have it, if we can. It is not worth claiming that we also have state-given reasons to *have* this belief or desire. And as I argue in Appendix A, we have other reasons to reject such claims.

6 Hedonic Reasons

Our object-given reasons to want some possible event are all provided by facts about this event. Such reasons are *telic* when they are provided by the facts that make some possible event good as an end, or worth achieving for its own sake. Such reasons are *instrumental* when they are provided by the fact that some event would have good effects, by being a means to some good end.

Telic reasons are *intrinsic* when they are provided by facts about some possible event's intrinsic properties or features, or what this event would *in itself* involve. We might have such reasons, for example, to want to make someone feel less lonely, or to see the sublime view from the summit of some mountain, or to understand how life or the Universe began. We might also have *extrinsic* telic reasons to want some possible event, which would be provided by facts about this event's relation to other events. But such reasons do not need to be separately considered, since such events would be *extrinsically* good by making some longer sequence of events, of which they were one part, *intrinsically* better.

Different objective theories partly disagree about which facts give us intrinsic telic reasons. Such theories may appeal to different views about well-being, or about which kinds of life are most worth living. These theories may also disagree about *whose* well-being we have reasons to care about, and try to promote. According to *Rational Egoism*, for example, each of us has reasons to care about and promote only our own well-being. According to *Rational Impartialism*, we always have most reason to care equally about everyone's well-being. We ought, I believe, to reject both these views. Nor should we assume that object-given reasons are provided only by facts about our own or other people's well-being. There may be other things that are worth achieving. Of this great variety of object-given reasons, it will be enough to consider here, as our examples, the reasons that are provided by certain facts about our hedonic well-being. These hedonic reasons are, I believe, widely misunderstood.

When we want something, we are often responding to the features of this thing that give us reasons to want it. But we have some desire-like states that are not, in this way, responses to reasons. Three examples

are the instinctive states of hunger, thirst, and lust. Another important set of mental states, though they are often assumed to be desires, are better regarded as being in a separate category. These are the *hedonic likings* and *dislikings* of certain actual present sensations that make our having these sensations pleasant, painful, or in other ways unpleasant, or in which their pleasantness or unpleasantness partly consists.

It is sometimes claimed that these sensations are in themselves good or bad in the sense that their intrinsic qualitative features, or what they *feel* like, gives us reasons to like them or dislike them. But we do not, I believe, have such reasons. Nor could these likings or dislikings be either rational or irrational. That is clearest in the case of some sensations that some people love and others hate, such as the sensations that we can give ourselves by eating milk chocolate, taking strenuous exercise, and having cold showers. Some of these likings or dislikings are odd. Many people hate the sound of squeaking chalk. I hate the feeling of touching velvet, the sound of buzzing house-flies, and the flattening, deadening effect of most overhead lights. The oddness of these dislikes does not make me less than fully rational. Whether we like, dislike, or are indifferent to these various sensations, we are not responding or failing to respond to any reasons.

Similar remarks apply, I believe, to many aesthetic experiences. It is sometimes claimed that we have reasons to enjoy, or be thrilled or in other ways moved by, great artistic works. In many cases, I believe, this claim is false. We can have reasons to *want* to enjoy, or to be thrilled or moved by, these artistic works. But these are not reasons to *enjoy*, or to *be* thrilled or moved by, these works. We do have reasons to admire some novels, plays or poems, given the importance of some of the ideas that they express. But poetry is what gets lost in the translation, even if this translation expresses the same ideas. And we never have reasons to enjoy, or be moved by, great music. If we ask what makes some musical passage so marvellous, the answer might be 'Three modulations to distant keys'. This answer describes a *cause* of our response to this music, not a *reason*. Modulations to distant keys are like the herbs, spices, or other ingredients that can make food delicious. When someone neither enjoys nor is moved by some great musical work, this person is not in any way less than fully rational, by failing to

respond to certain reasons. In comparing music with food in this way, I am not belittling music, ranking it below novels, plays, or poems. Without music, Nietzsche plausibly (though falsely) said, life would be an error. But music is also the lost battlefield and graveyard of most general aesthetic theories.

Since these claims are controversial, we can return to those non-aesthetic sensations that people like or dislike. Though these sensations are not in themselves good or bad, they are parts of complex mental states that *are* good or bad. When we are in pain, what is bad is not our sensation but our conscious state of having a sensation that we dislike. If we didn't dislike this sensation, our conscious state would not be bad. What these sensations feel like may in part depend on whether we dislike them. Such sensations might be claimed to be in themselves bad when their quality is affected in certain ways by our disliking them. On this view, it would still be true that, if we didn't dislike these sensations, neither they nor our conscious state would be bad, nor would we be failing to respond to some reason.

When we are having some sensation that we intensely like or dislike, most of us also strongly want to be, or not to be, in this conscious state. Such desires about such conscious states we can call *meta-hedonic*. Many people fail to distinguish between hedonic likings or dislikings and such meta-hedonic desires. But these mental states differ in several ways. What we dislike is some sensation. What we want is not to be having a sensation that we dislike. Our desire could be fulfilled either by our ceasing to have this sensation, or by our continuing to have it but ceasing to dislike it. No such claims apply to dislikes, which, unlike desires, cannot be fulfilled or unfulfilled.

Another difference involves time. Suppose that some flame is moving towards our hand, threatening us with great pain in the near future. Most of us would strongly want to avoid this future pain. But we cannot now *dislike* this future pain. Nor can we now like some future pleasure. Unlike our meta-hedonic desires, our hedonic likings or dislikings cannot be aimed at the future, or at what is merely possible. That is another reason why I do not call these mental states desires.

If we call these states desires, we should remember that, given the differences between these states and our other desires, true claims about

these states may not apply to our other desires. There are some other important and often ignored differences between these states and our meta-hedonic desires.

First, many people believe that our desires can *create* or *confer* value or disvalue. Korsgaard, for example, writes that something can be 'objectively good as an end *because* it is desired for its own sake'. On this view, we create value by valuing things, and things matter by mattering to us. This view may seem to be supported by the examples of pleasure and pain. Our hedonic likings and dislikings *do*, as I have said, make some of our conscious states good or bad. If we fail to distinguish between these likings or dislikings and our meta-hedonic desires, we may believe that these desires make their objects good or bad. That may seem to support the general view that our desires can create value.

Korsgaard's remarks provide one example. To illustrate her claim that something can be good 'because it is desired for its own sake', Korsgaard writes: 'chocolate gets its value from the way it affects us. We *confer* value on it by liking it.' Such examples do not, I believe, show that our *desires* can create or confer value, or disvalue, by making what we want to have, or to avoid, good or bad. Our future pleasures or pains are not made to be good or bad by our present desires to have these pleasures, and to avoid these pains. And when we are in great pain, by having some sensation that we intensely dislike, what makes our conscious state bad is our intense dislike, not our present desire not to be having the sensation that we dislike. Since our meta-hedonic desires do not make their objects good or bad, the examples of pleasure and pain do not decisively, or even, I believe, strongly support the view that our other desires have such value-creating power. Though it is good to have sensations that we *like*, nothing is good merely because we *want* this thing.

There is another important difference between these two kinds of mental state. Unlike our hedonic likings or dislikings, our meta-hedonic desires *are* responses to reasons, since we can have strong reasons for and against having such desires. This difference is enough to show that we should distinguish these two kinds of mental state. When we are experiencing intense pleasure, by having some sensation that we intensely like, we have no reason to be liking this sensation. If we did not like this sensation, we would not be being irrational, or making any

mistake. But we have strong reasons to want to be having, and to go on having, sensations that we intensely like. We have even stronger reasons to want not to be in agony, by having sensations that, for no reason, we intensely dislike.

7 Irrational Preferences

Our desires are rational, I have claimed, when we want events whose features give us reasons to want them. Our desires are not rational, and are in the old phrase *contrary to reason*, when we want some event that we have reasons *not* to want, and no reasons, or only weaker reasons, to want. When some desire is clearly and strongly contrary to reason, this desire is irrational. Other such desires are merely less than fully rational. There is no sharp borderline here, since irrationality is a matter of degree.

Suppose, for example, that we must choose which of two possible ordeals we shall later undergo. If one of these ordeals would be much more painful, this fact gives us a strong reason to prefer the other. If we have no other relevant reason, it would be contrary to reason, and in this way irrational, knowingly to prefer the more painful ordeal.

Most preferences of this kind involve our attitudes to time. Consider first an imagined man who has an attitude that we can call *Future Tuesday Indifference*. This man cares about his own future pleasures or pains, except when they will come on any future Tuesday. This strange attitude does not depend on ignorance or false beliefs. Pain on Tuesdays, this man knows, would be just as painful, and just as much *his* pain, and Tuesdays are just like other days of the week. Even so, given the choice, this man would now prefer agony on any future Tuesday to slight pain on any other future day. That some ordeal would be much more painful is a strong reason *not* to prefer it. That this ordeal would be on a future Tuesday is *no* reason to prefer it. So this man's preferences are strongly contrary to reason, and irrational.

Consider next some man who has a *bias towards the next year*. This imagined man cares equally about his future well-being throughout the next year, but he cares only half as much about his well-being in later years. Rather than having five hours of pain eleven months from now,

he would prefer to have nine hours of pain twelve months from now. Such preferences are also irrational. If we would have some future pain just over rather than just under a year from now, that is no reason to care now about this pain only half as much.

No one has these attitudes to time. But many of us have what I have called the *bias towards the near*. Unlike these two imagined attitudes, this bias does not draw wholly arbitrary distinctions. But suppose that, because you have this bias, you want some ordeal to be briefly postponed, though you know that this postponement would make your ordeal much worse. Rather than having one minute of slight pain later today, you prefer to have one hour of agony tomorrow. This preference would also be, though more weakly, irrational. Many people often act on less extreme preferences of this kind, thereby making their lives go worse.

These claims may seem too obvious to be worth making. Who could possibly deny that the nature of agony gives us reasons to want to avoid being in agony, and that the nature of happiness gives us reasons to want to be happy?

Such claims are denied by some great philosophers, and in many recent accounts of rationality. And, as we shall see, such claims *must* be denied by those who accept subjective theories about reasons.

3

Subjective Theories

8 Subjectivism about Reasons

Subjective theories appeal to facts about our present desires, aims, and choices. On the simplest subjective theory, which we can call

> *the Desire-Based Theory*: We have a reason to do whatever would fulfil any of our present desires.

For subjective theories to be plausible, however, they must admit that some desires do not give us reasons. Return to the case in which you want to run away from an angry, poisonous snake because you believe falsely that this act would save your life. If you had reasons to fulfil all of your present desires, your desire to run away would give you a reason for acting. But you have no reason to run away, since standing still is your only way to save your life.

There are two ways to explain why your desire to run away gives you no reason for acting. Subjectivists might claim that

> (A) reasons are provided only by desires that depend on true beliefs.

You have no reason to run away, (A) implies, because your desire depends on the false belief that this act would save your life. Remember next that our desires are *telic* when we want some event as an end, or for its own sake, and *instrumental* when we want some event as a means to some end. Our *aims* are often the telic desires that we have decided to

try to fulfil. You want to run away merely as a means of saving your life. So Subjectivists might instead claim that

(B) reasons are provided only by telic desires, or aims.

You have no reason to run away, (B) implies, because this act would not help you to fulfil or achieve any such desire or aim.

(A) may seem more plausible than (B). When instrumental desires depend on false beliefs, that may seem to make these desires in one way mistaken, which could be why such desires provide no reasons. When such desires do not depend on false beliefs, they may not be in any way mistaken.

Subjectivists can defend (B), however, in a different way. Suppose that I want to eat the two remaining apples that are on some tree. I also want to climb a ladder so that I can reach the higher apple. Suppose next that this tree's owner allows me to eat only one of these apples, and lets me choose which apple I shall eat. If instrumental desires gave us reasons, I would have more reason to choose the higher apple. If I chose the lower apple, I would then fulfil only my desire to eat this apple. If I chose the higher apple, I would fulfil not only my desire to eat this other apple, but also my instrumental desire to climb this ladder so that I can reach this apple. But this reasoning is obviously mistaken. Since I want to climb this ladder, not for its own sake, but only as a means of reaching this apple, I have no further, independent reason to fulfil this desire. My reason to climb this ladder derives entirely from, and adds nothing to, my reason to fulfil my desire to eat this higher apple.

As this example shows, instrumental desires do not provide reasons. On the simplest plausible subjective theory, which we can call

> *the Telic Desire Theory*: We have most reason to do whatever would best fulfil or achieve our present telic desires or aims.

This theory correctly implies that you have no reason to run away from the angry snake. Your aim is to save your life, and running away would not achieve this aim. There is no need to appeal to the fact that your desire to run away depends on a false belief.

In some cases, however, our *telic* desires or aims depend on false beliefs. I might want to hurt you, for example, because I falsely believe that you deserve to suffer, or because I want to avenge some injury that I falsely believe you have done me. Subjectivists ought to deny that this desire gives me a reason. When they consider such cases, many Subjectivists claim that reasons are provided only by telic desires or aims that are *error-free*, in the sense that they do not depend on false beliefs. According to what we can call

> *the Error-Free Desire Theory*: We have most reason to do whatever would best fulfil or achieve our present error-free telic desires or aims.

There are some obvious ways to improve this theory. If no reasons are provided by desires that depend on false beliefs, we can plausibly say the same about desires that depend on ignorance. This distinction is not deep. In the imagined case in which I want to hurt you, there are two ways in which my desire might be ill-grounded. I might believe falsely that you have intentionally injured me; or, though believing truly that you have injured me, I might not know that your aim was to save me from some greater injury. There is little difference between these versions of this case. If my desire to hurt you provides no reason when, and because, it depends on a false belief, this desire seems equally to provide no reason when it depends on ignorance.

If desires that depend on ignorance provide no reasons, we can plausibly take a further step. Subjectivists can claim that, just as we do *not* have reasons to fulfil those of our actual telic desires that we would *not* now have if we knew more, we *do* have reasons to fulfil the telic desires that, if we had greater knowledge, we *would* now have. As before, this distinction is not deep. If I learnt that you had good motives for injuring me, I might not only cease to wish you ill, but also come to wish you well. If that is true, Subjectivists might claim, I have a reason now to treat you well.

If we appeal to what we would want if we knew more, we might next carry this idea to its limit. According to

> *the Informed Desire Theory*: We have most reason to do whatever would best fulfil the telic desires or aims that we would now have if we knew all of the relevant facts.

Any fact counts as *relevant*, some writers claim, if our knowledge of this fact would affect our desires. But this criterion is too wide. As Gibbard remarks, if we knew and vividly imagined the full facts about what is going on in the innards of our fellow-diners, we might lose our desire to eat. And if we learnt certain facts about man's inhumanity to man, we might become so depressed that we would lose our desire to live. The Informed Desire Theory would then implausibly imply that, even though we actually want to eat and to stay alive, we have no reason to fulfil these desires. To avoid such implications, some Subjectivists claim that, for some fact to count as *relevant*, it is not enough that our knowledge of this fact would affect our desires. On such views, when we are choosing between several possible acts, what are relevant are only facts about these acts and their possible outcomes.

The Informed Desire Theory needs another revision. It is sometimes true that, if we were fully informed, that would change our situation in some way that altered both our desires and what we had reasons to do. If Subjectivists claim that our reasons are provided, not by our actual desires, but by our hypothetical informed desires, these people may be led in such cases to implausible conclusions. Suppose, for example, that we want to learn certain important facts. If we knew these facts, we would lose this desire. But that should not be taken to imply that we have no reason to act on this desire, by trying to learn these facts. Some Subjectivists therefore claim that we should try to fulfil the desires that, if we were fully informed, we would want ourselves to have in our actual uninformed state.

Some other Subjectivists appeal, not to what would best fulfil or achieve our desires or aims, but to the choices or decisions that we would make after carefully considering the facts. These people also make claims about how it would be rational for us to make such decisions. According to what we can call

> *the Deliberative Theory*: We have most reason to do whatever, after fully informed and rational deliberation, we would choose to do.

This form of Subjectivism can be easily confused with Objectivism, since such theories can be stated in deceptively similar ways. Subjectivists and Objectivists might both claim that

> (C) what we have most reason to do, or decisive reasons to do, is the same as what, if we were fully informed and rational, we would choose to do.

But this claim is ambiguous. Subjectivists and Objectivists may both claim that, when we are trying to make some important decision, we ought to deliberate in certain ways. We ought to try to imagine fully the important effects of our different possible acts, to avoid wishful thinking, to assess probabilities correctly, and to follow certain other procedural rules. If we deliberate in these ways, we are *procedurally* rational.

Objectivists make further claims about the desires and aims that we would have, and the choices that we would make, if we were also *substantively* rational. These claims are *substantive* in the sense that they not about *how* we make our choices, but about *what* we choose. There are various telic desires and aims, Objectivists believe, that we all have strong and often decisive object-given reasons to have. To be fully substantively rational, we must respond to these reasons by having these desires and aims, and trying to fulfil or achieve them if we can. Deliberative Subjectivists make no such claims. These people deny that we have such object-given reasons, and they appeal to claims that are only about procedural rationality.

Though these two groups of people might both accept (C), they would explain (C) in different ways. According to these Subjectivists, when it is true that

> (D) if we were fully informed and procedurally rational, we would choose to act in some way,

this fact makes it true that

> (E) we have decisive reasons to act in this way.

Objectivists claim instead that, when it is true that

> (E) we have decisive reasons to act in some way,

this fact makes it true that

> (F) if we were fully informed and both procedurally and substantively rational, we would choose to act in this way.

To illustrate these claims, we can suppose that, unless I stop smoking, I shall die much younger, losing many years of happy life. According to all plausible objective theories, this fact gives me a decisive reason to want and to try to stop smoking. If I were fully informed and substantively rational, that is what I would choose to do. What we ought rationally to choose, Objectivists believe, depends on what we have such reasons or apparent reasons to want and to do.

Suppose next that, after fully informed and procedurally rational deliberation — or what we can now call *ideal* deliberation — I would choose to stop smoking. Deliberative Subjectivists would then agree that I have a decisive reason to stop smoking. On this view, however, the inference runs the other way. Instead of claiming that what we ought to choose depends on our reasons, these Subjectivists claim that our reasons depend on what, after such deliberation, we would choose. If I have decisive reasons to stop smoking, that is *because* I would choose to act in this way.

As this example shows, these theories are very different. These Objectivists appeal to normative claims about what, after ideal deliberation, we have *reasons* to choose, and *ought rationally* to choose. These Subjectivists appeal to psychological claims about what, after such deliberation, we *would in fact* choose.

Different subjective theories sometimes disagree about what we have reasons to do. We can here ignore such disagreements, and consider only cases in which these theories agree. In such cases, we know all of the relevant facts, and the act that would best fulfil our present telic desires or aims is also what we would choose to do after ideal deliberation. We can then say that, according to

Subjectivism about Reasons: Some possible act is

> what we have most reason to do, and what we should or ought to do in the decisive-reason-implying senses,

just when, and because,

> this act would best fulfil our present fully informed telic desires or aims, or is what, after ideal deliberation, we would choose to do.

There is another disagreement between some subjective theories that we can note but then ignore. Suppose that, given the relevant facts, all subjective theories imply that I have a decisive reason to stop smoking. This reason, some of these theories claim, is given by the fact that

> (1) this act would best fulfil my present fully informed telic desires.

According to some other subjective theories, this reason is given by the fact that

> (2) stopping smoking would lengthen my life.

But (2) gives me this reason, these theories claim, only because (1) is also true. My reason to stop smoking is given by the fact that this act would lengthen my life, but this fact gives me this reason only because I want to achieve this aim. Similar claims apply to the fact that

> (3) after ideal deliberation, I would choose to stop smoking.

According to Deliberative Subjectivists, we have decisive reasons to do whatever, after ideal deliberation, we would choose to do. But my reason to stop smoking cannot be plausibly claimed to be given by the fact that this is what, after such deliberation, I would choose to do. Some of these people therefore claim that (2) is the fact that gives me my reason, but that (2) gives me this reason only because (3) is also true.

In assessing subjective theories, it will be enough to consider *what* these theories imply that we have reasons to do, ignoring these disagreements about which are the facts that give us these reasons.

When I say that, on these theories, reasons are provided by certain facts about our desires, aims, or choices, I shall also mean that these are among the facts that make it true that we have these reasons.

Subjectivism about Reasons is now very widely accepted. Many people take it for granted that we have subject-given reasons. Korsgaard for example writes that, if some act 'is a means to getting what you want . . . no one doubts that this is a reason'. Williams writes: 'Desiring to do something is of course a reason for doing it.' In many books and articles, Subjectivism is not even claimed to be the best of several views, but is presented as if it were the only possible view. So it is of great importance whether this view is true.

9 Why People Accept Subjective Theories

We ought, I believe, to reject all subjective theories, and accept some objective theory. Our practical reasons are all object-given and value-based.

Since so many people believe that *all* practical reasons are desire-based, aim-based, or choice-based, how could it be true that, as objective theories claim, there are *no* such reasons? How could all these people be so mistaken?

There are several possible partial explanations, because there are several ways in which our reasons may seem to be based on some of our desires, aims, or choices. First, as I have said, what we want is often something that is worth doing or achieving. In such cases, these two kinds of theory at least partly agree, since we have value-based object-given reasons to try to fulfil such desires.

Second, we often have such desires because we believe that we have such reasons. We are often motivated by the belief that some act or outcome would be good or best, in the reason-implying sense. When our desires depend on our beliefs that we have such reasons, we may fail to distinguish between these desires and these beliefs.

Third, some people accept desire-based theories about well-being. According to some of these theories, the fulfilment of some of our

present desires would be in itself good for us. If that were true, we would have value-based reasons to fulfil these desires.

Fourth, we can rightly appeal to our desires or aims when we describe our *motivating* reasons, or why we acted as we did. This may lead us to assume that our desires or aims can also give us *normative* reasons. And some people do not distinguish between these two kinds of reason.

Fifth, there is a superficial sense in which our desires or aims can be truly claimed to give us normative reasons. For example, I might truly claim that I have a reason to leave some meeting now, because I want to catch some train, or because my aim is to catch this train, and leaving now is my only way to fulfil this desire, or achieve this aim. But this desire-based or aim-based reason would be *derivative*, since this reason's normative force would derive entirely from the facts that gave me my reasons to want to catch this train, or to have this aim. If I had no reason to want to catch this train, or to have this aim, I would have no reason to leave now. When I claim that no reasons are provided by our desires or aims, I am referring to our primary, non-derivative reasons.

Sixth, when we could fulfil *other people's* desires, or help these people to achieve their aims, these facts may give us *non*-derivative reasons to act in these ways. When other people have some desire or aim that they have no reason to have, these people may have no reason to try to fulfil this desire or achieve this aim. But *we* may have such reasons. In helping other people to fulfil or achieve their desires or aims, we respect these people's autonomy, and avoid paternalism. Other people's desires, aims, or choices are often, in this respect, like votes, which should be given just as much weight even when the voters have no reason to vote as they do. Many people accept desire-based or choice-based theories because they are democrats, liberals, or libertarians, who believe that we should not tell other people what they ought to want, or choose, or do. Nozick, for example, claims that a substantive value-based theory 'opens the door to despotic requirements, externally imposed'.

Seventh, when we have some aim, and we believe that some possible act would be the only or the best way to achieve this aim, it may be true that we ought rationally to act in this way. Some people assume that, in such cases, we must have a reason to do what we ought rationally

to do. But that is not so. When we believe falsely that some act would achieve our aim, we may have no reason to act in this way. Though you ought rationally to run away from the angry snake, you have no reason to run away.

Eighth, when people claim that we have reasons to fulfil our present desires, they are often thinking of our desires for future activities or experiences that we believe we would enjoy. When these beliefs are true, we have reasons to fulfil these desires. But these reasons are provided, not by the fact that we would be fulfilling these desires, but by the fact that we would enjoy these future activities or experiences. If we would *not* enjoy these activities or experiences, we may have no reason to fulfil these desires. When children want something that they later get but don't enjoy, their parents sometimes say, 'See! You didn't *really* want that.' Such claims are false, since these children *did* want these things, and the truth is rather that their desires didn't give them reasons. Similar claims apply to our desires to avoid what we believe would be painful, or unpleasant. When people claim that our desires give us reasons, it is often such facts about what we would enjoy, or find painful or unpleasant, that they really have in mind. Such facts give us reasons that are *hedonic* rather than *desire-based*.

Ninth, some people mistakenly believe that hedonic reasons *are* desire-based. When these people think about sensations that are painful or unpleasant, they do not distinguish between our dislike of these present sensations and our meta-hedonic desires not to be having sensations that we dislike. It is our dislike, I have claimed, that makes our conscious state bad, and gives us our reason to try to end our pain, or our unpleasant state. Since these people do not distinguish between our dislike and our meta-hedonic desire, they believe that this desire gives us this reason. Similar claims apply to pleasures, and to some other good or bad conscious states.

Tenth, we have many reasons for acting that we wouldn't have if we didn't have certain desires. But these reasons are provided, not by the facts that our acts would fulfil these desires, but by certain other facts that causally depend on our having these desires. When we have some desire, for example, that may cause it to be true that this desire's fulfilment would be pleasant. In many cases, this fact would merely

give us a further reason to fulfil this desire, since what we want would be in itself worth achieving. But such cases take their clearest form when we have no such reason to have some desire. When we play some kinds of game, for example, such as games without rewards whose outcomes depend on luck, we have no reason to want to win. But if we do want to win, that may make it true that we would enjoy winning, and this second fact would then give us a reason to try to fulfil this desire.

In describing such cases, we can draw another distinction. According to subjective theories, some facts give us reasons in a way that depends on our having some desire. This dependence is *normative*. On some views, for example, my reason to stop smoking is given by the fact that this act would lengthen my life, but this fact gives me a reason only because I want to achieve this aim. This reason's normative force is claimed to derive from the fact that I have this desire, so this reason is desire-based. The value-based reasons that I have just described are quite different. When the fulfilment of some desire would give us pleasure, this fact gives us a value-based hedonic reason to do what would fulfil this desire. This reason may *causally* depend on our having this desire, since this act may give us pleasure only because we have this desire. But this reason would not *normatively* depend on our having this desire. If some act would give us pleasure, this fact gives us a reason to act in this way, whether or not this pleasure causally depends on our having some desire.

We have many other reasons that causally depend on our having some desire. Unfulfilled desires may, for example, be distressing, or distracting. Such facts give us reasons to fulfil these desires. As before, these would often merely be further reasons, since what we want would often be worth achieving. But such cases may involve desires that we have no such reasons to have. We may be distracted, for example, by wanting to know or remember some trivial fact, or by some obsessive or compulsive desire. I am sometimes distracted by a strangely affectless desire to cut my fingernails. It can be best to get rid of such desires by fulfilling them.

Suppose next that we must choose between two or more good possible aims, none of which would be more worth achieving than any of the

others. Some examples are choices between different possible careers, or research projects, or between doing voluntary work for different aid agencies, or political campaigns. If there is one of these possible aims that we most strongly want to achieve, this fact may give us reasons to adopt this aim. But these reasons would again be given, not by the fact that our strongest desire is to achieve this aim, but by certain other facts that would depend on our having this desire. If one of these aims seems most appealing, for example, that may give us reasons to believe that we would find this aim's achievement most rewarding. The thought of this aim's achievement may give us pleasure in advance. And our strongly wanting to achieve this aim may make it easier for us to make the efforts and sacrifices that would be needed to achieve this aim. We may need such desires in our darkest hours, when we are losing energy or hope. As before, it would be these other facts, and not our desire itself, that would give us reasons to adopt and try to achieve this aim.

Similar claims apply to our decisions and aims. When we have decided to try to fulfil some desire, thereby making its fulfilment one of our aims, this decision may give us a further reason to try to fulfil this desire, thereby achieving this aim. But this reason would not be provided merely by the fact that we have made this decision and adopted this aim. This reason would be provided by the fact that, if we do not act on this decision, we shall be less likely to achieve this aim, and more likely to waste our time. In some cases, however, neither is true, since we have nothing better to do than to reconsider some decision. If we have woken up in the middle of the night, for example, reconsidering our decision to adopt some aim may be less boring than simply waiting to drift back to sleep. In such cases, the fact that we have adopted some aim gives us no reason to keep and to try to achieve this aim, since this fact gives us no reason not to change our mind, and adopt some other aim instead.

We have many reasons to fulfil our desires or aims that are provided, not by the fact that we would be fulfilling these desires or aims, but by such other *desire-dependent* or *aim-dependent* facts. As before, when people claim that our desires or aims give us reasons, it is often such other facts that they really have in mind.

Since there are all these many ways in which our desires, aims, or choices can seem to give us reasons for acting, it is not surprising that so many people accept subjective theories. Many of these people have various true or plausible beliefs about which are the facts that give us reasons, and they have merely failed to see that these beliefs do not in fact support any subjective theory. Though these people may believe that they are Subjectivists, that is not really true. When these people make Subjectivist claims, they are misdescribing their view.

10 Analytical Subjectivism

There is another way in which some people have come to accept subjective theories about reasons. We can call some normative claim

substantive when this claim both

(a) states that something has some normative property,

and

(b) is *significant*, by being a claim with which we might disagree, or which might be informative, by telling us something that we didn't already know.

Two examples are the claims that it is bad to be in pain and irrational to care less about the further future.

As both Kant and Sidgwick warn, when we think about normative questions, we can be easily misled by claims that seem substantive but are merely *concealed tautologies*. In Kant's words:

> There is no science so filled with tautologies as ethics.

An *open* tautology uses the same words twice, in some way that does not make any significant claim, but tells us only that something is what it is, or that if something has a certain property, this thing has this property. Two examples are the claims that

(1) happiness is happiness,

and that

(2) acts that produce happiness produce happiness.

Some open tautologies can be used to suggest significant claims. Two examples are 'Business is business' and 'War is war'. When people make such claims, they intend to remind us that something is distinctively different from other things, and must be judged in its own terms. In business or war, these people may intend to suggest, ordinary moral standards do not apply. These suggested claims would be substantive. But most open tautologies are trivial. It is not worth claiming that happiness is happiness, desires are desires, beliefs are beliefs, and hope is hope.

Rather than using the same words twice, a *concealed* tautology uses different words or phrases with the same meaning. One example is the claim that

(3) felicity is happiness.

Since 'felicity' means 'happiness', (3) means the same as (1). (3) is not a substantive claim, though we might use (3) to tell someone what the word 'felicity' means. Consider next the claim that

(4) acts that produce happiness are felicific.

Since 'felicific' means 'produces happiness', (4) is another concealed tautology, whose two open forms would be

(2) acts that produce happiness produce happiness,

and

(5) acts that are felicific are felicific.

As before, these are not substantive claims. Everyone who understands these claims would accept them, because they are so obviously true. And everyone could consistently accept these claims whatever else they believe. (4) differs in these ways from the claim that

(6) acts that produce happiness are good.

Since 'good' does *not* mean 'produces happiness', (6) *is* a significant, substantive claim, which sometimes conflicts with many people's beliefs. Many people believe, for example, that cruel acts that give happiness to sadists are not in any way good.

Return now to subjective theories about reasons. Some people use the words 'reason', 'should', and 'ought' in what we can call *subjectivist* or *internal* senses. We can call these people *Analytical Subjectivists*. When some people, for example, say that

> (7) we have most reason to act in some way,

they mean that

> (8) this act would best fulfil our present fully informed telic desires.

This subjectivist sense of the phrase 'have most reason' we can call the *desire-fulfilment* sense. Some of these people also claim that

> (9) we have most reason to do what would best fulfil our present fully informed telic desires.

Since these people use the phrase 'have most reason' in the desire-fulfilment sense, (9) is not a substantive claim, but a concealed tautology, one of whose open forms would be the claim that

> (10) the act that would best fulfil our present fully informed telic desires is the act that would best fulfil these desires.

Everyone could accept this trivial claim, whatever else they believe. Similar claims apply to other subjectivist or internal senses of 'reason', 'should', and 'ought'. Though Analytical Subjectivists do not make substantive claims about what we have reasons to do, or about what we should or ought to do, these people make some other important claims, which I discuss in Chapters 24 and 30.

For Subjectivists about Reasons to make substantive claims, they must use the words 'reason', 'should', and 'ought' in the indefinable, normative

senses that I discussed in Section 1. It is these substantive, non-analytical subjective theories that, in these chapters, I am discussing.

It will be enough to consider cases in which different subjective theories agree. In such cases, we know all of the relevant facts, and the act that would best fulfil our present telic desires or aims is also what we would choose to do after ideal deliberation. Our deliberation is *ideal* when it is fully informed and procedurally rational. In discussing these theories, I shall make some claims that are only about desire-based reasons, but most of these claims would also apply to aim-based and choice-based reasons.

When making these claims, I shall use the word 'desire' in a wide sense, which covers any state of being motivated, or of wanting something to happen and being to some extent disposed to make it happen, if we can. My claims do not apply, however, to various complex states that involve desires. When we love someone, for example, we are motivated to act in certain ways. We care greatly about this person's well-being, and we want to do what would be best for him or her. Though our loving someone partly consists in our having such desires, we have strong reasons, I believe, to care about, and try to promote, the well-being of those we love. Such reasons are provided, not by the desires involved in loving someone, but by the ways in which love is in itself good, and by various other facts about our relations to those we love, such as facts about shared histories, or commitments, or reasons for gratitude, or by the facts that are involved in romantic or erotic love, or love for our parents, children, or other close relatives. To illustrate this distinction, we can suppose that I meet several strangers, all of whom need my help. If I had a strong desire to help one of these strangers, perhaps because I like her face, that would at most give me only a weak reason to help this stranger rather than any of the others. Love, in its various forms, is very different from such a desire.

11 The Agony Argument

Subjective theories can have implausible implications. Suppose that, in

> *Case One*, I know that some future event would cause me to have some period of agony. Even after ideal deliberation, I

> have no desire to avoid this agony. Nor do I have any other desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony.

Since I have no such desire or aim, all subjective theories imply that I have no reason to want to avoid this agony, and no reason to try to avoid it, if I can.

This case might be claimed to be impossible, because my state of mind would not be *agony* unless I had a strong desire *not* to be in this state. But this objection overlooks the difference between our attitudes to present and future agony. Though I know that, when I am later in agony, I shall have a strong desire not to be in this state, I might have no desire now to avoid this future agony.

It might next be claimed that my predictable future desire not to be in agony gives me a desire-based reason now to want to avoid this future agony. But this claim cannot be made by those who accept subjective theories of the kind that we are considering. These people do not claim, and given their other assumptions they could not claim, that facts about our *future* desires give us reasons.

Some other theories make that claim. A value-based objective theory about *reasons* might be combined with a desire-based subjective theory about *well-being*. On such a view, even if we don't now care about our future well-being, we have reasons to care, and we ought to care. These reasons are value-based in the sense that they are provided by the facts that would make various future events good or bad for us. But if our future well-being would in part consist, as this view claims, in the fulfilment of some of our future desires, these *value-based* reasons would be reasons to act in ways that would cause these future *desires* to be fulfilled. It might be similarly claimed that we have value-based reasons to fulfil other people's desires, because such acts would promote the well-being of these other people. Though these theories claim that we have reasons to fulfil these desires, these value-based objective theories about reasons are very different from the desire-based subjective theories that we are now considering.

We can also imagine a temporally neutral desire-based theory. On this view, what we have most reason to do, at any time, is whatever

would best fulfil all of our desires throughout our life, whether or not these acts would be good for us. According to a similar, personally neutral theory, what we have most reason to do is whatever would best fulfil everyone's desires, whether or not these acts would be good for anyone. These imagined theories are also very different from the subjective theories that we are now considering.

According to these theories, it is only certain facts about our own *present* desires, aims, or choices that give us reasons, or on which our reasons depend. We are supposing that, in *Case One*, I have carefully considered all of the relevant facts about my possible future period of agony. Since I have no present desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony, all subjective theories imply that I have no reason to want to avoid this agony. Similar claims apply to my acts. Even if I could easily avoid this agony—perhaps by moving my hand away from the flames of some approaching fire—I have no reason to act in this way. Such a reason would have to be provided by some relevant present desire, and I have no such desire.

Some *Analytical* Subjectivists would accept this conclusion. If these people claimed that I would have no reason to avoid this agony, their claim would not be normative, but a concealed tautology, which merely repeats my description of this imagined case. These people would mean only that, after ideal deliberation, I am not motivated to move my hand away from the approaching fire. We could all agree that, in this trivial and misleading sense, I would have no reason to act in this way.

We are discussing the views of *Non*-Analytical Subjectivists. These people use the phrase 'a reason' in the normative sense that we can also express with the phrase 'counts in favour'. These Subjectivists agree that it would make sense to claim that I have a reason to want and to try to avoid this future agony. But these people's theories imply that, since I have no relevant present desire, I have no such reason. No fact counts in favour of my wanting and trying to avoid this agony. Similar claims apply to other such cases. According to these Subjectivists, when we have no relevant present desires, we would have no reason to want to avoid some period of future agony.

We can now argue:

> We all have a reason to want to avoid, and to try to avoid, all future agony.
>
> Subjectivism implies that we have no such reason.

Therefore

> Subjectivism is false.

We can call this *the Agony Argument*.

Some Subjectivists might claim that we can ignore this argument, because my example is purely imaginary. Every actual person, they might say, wants to avoid all future agony.

This reply would fail. First, we are asking whether subjective theories imply that we all have a *reason* to want to avoid all future agony. To support the claim that we all have such a reason, it is not enough to claim that everyone *has* this desire. These Subjectivists would also have to claim that, when we have some desire, this fact gives us a reason to have it. As we shall see, that is an indefensible claim.

Second, it seems likely that some actual people do not want to avoid all future agony. Many people care very little about pain in the further future. Of those who have believed that sinners would be punished with agony in Hell, many tried to stop sinning only when they became ill, and Hell seemed near. And when some people are very depressed, they cease to care about their future well-being.

Third, even if there were no such actual cases, normative theories ought to have acceptable implications in merely imagined cases, when it is clear enough what such cases would involve. Subjectivists make claims about which facts give us reasons. These claims cannot be true in the actual world unless they would also have been true in possible worlds in which there were people who were like us, except that these people did not want to avoid all future agony, or their desires differed

from ours in certain other ways. So we can fairly test subjective theories by considering such cases.

Subjectivists might reply that, even in such possible worlds, there would be some telic desires that everyone must have, because without these desires these people could not even be rational agents, who can act for reasons. To be such agents, Williams suggests, we must have 'a desire not to fail through error', and some 'modest amount of prudence'. But such claims are irrelevant here. We could be agents who act for reasons without wanting to avoid all future agony.

Subjectivists might next claim that, if some theory has acceptable implications in all or most actual cases, this fact may give us sufficient reasons to accept this theory. We might justifiably accept such a theory even if there are some unusual or imagined cases in which this theory's implications seem to be mistaken.

Many theories of many kinds can be plausibly defended in this way. For such a defence to succeed, however, we must be able to claim that there are no other, competing theories which have more acceptable implications. And Subjectivists cannot make that claim. When subjective theories are applied to actual people, these theories often have plausible implications. But that is because most actual people often have desires that they have object-given reasons to have, because they want things that are in some way good, or worth achieving. In many such cases, subjective theories have the same implications as the best objective theories. In trying to decide which theories are best, we must consider cases in which these two kinds of theory disagree. That is how, for similar reasons, we must decide between different scientific theories. Such disagreements take their clearest form in some unusual actual cases and some imaginary cases. So Subjectivists cannot claim that we can ignore these cases, or that we can give less weight to them. On the contrary, these are precisely the cases that we have *most* reason to consider. In their claims about such cases, subjective theories are, I am arguing, much less plausible than the best objective theories. And if these objective theories are more plausible whenever these two kinds of theory disagree, these objective theories are clearly better.

There is another possible reply. Deliberative Subjectivists appeal to what we would want and choose after some process of informed and *rational* deliberation. These people might argue:

> (A) We all have reasons to have those desires that would be had by anyone who was fully rational.
>
> (B) Anyone who was fully rational would want to avoid all future agony.

Therefore

> We all have a reason to want to avoid all future agony.

As I have said, however, such claims are ambiguous. Objectivists could accept (B), because these people make claims about *substantive* rationality. According to objective theories, we all have decisive reasons to have certain desires, and to be substantively rational we must have these desires. These reasons are given by the intrinsic features of what we might want, or might want to avoid. We have such a decisive object-given reason to want to avoid all future agony. If we did not have this desire, we would not be fully substantively rational, because we would be failing to respond to this reason.

Subjectivists cannot, however, make such claims. On subjective theories, we have no such object-given reasons, not even reasons to want to avoid future agony. Deliberative Subjectivists appeal to what we would want after deliberation that was *merely procedurally* rational. On these theories, *if* we have certain telic desires or aims, we may be rationally required to want, and to do, what would best fulfil or achieve these desires or aims. But, except perhaps for the few desires without which we could not even be agents, there are no telic desires or aims that we are rationally required to have. We can be procedurally rational whatever else we care about, or want to achieve. As one Subjectivist, Rawls, writes:

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently.

So Subjectivists cannot claim that anyone who is fully rational would want to avoid all future agony.

It might be objected that, in making these remarks, I have underestimated what Subjectivists can achieve by appealing to claims about procedural rationality. Smith, for example, claims that

> (C) we are rationally required not to have desires or preferences that draw some arbitrary distinction.

By appealing to this 'minimal principle', Smith writes, Subjectivists can explain the irrationality of many desires and preferences, such as the preferences of my imagined man who cares about what will happen to him except on any future Tuesday. This man's preferences are irrational, Smith claims, because they draw an arbitrary distinction. It would be similarly arbitrary, Subjectivists might claim, not to want to avoid all future agony.

Subjectivists cannot, however, make such claims. Our preferences draw arbitrary distinctions when, and because, what we prefer is in no way preferable. It is arbitrary to prefer one of two things if there are no facts about these things that give us any reason to have this preference. My imagined man would prefer to have one of two similar ordeals if, and because, this ordeal would be on a future Tuesday. To explain why this preference is arbitrary, we must claim that

> (1) if some ordeal would be on a future Tuesday, this fact does not give us any reason to care about it less.

Unlike my imagined man, most of us would always prefer to have one of two ordeals if, and because, this ordeal would be less painful. To explain why *this* preference is *not* arbitrary, we must claim that

> (2) if some ordeal would be less painful, this fact *does* give us a reason to care about it less.

(1) and (2) are claims about object-given reasons. Since Subjectivists deny that we have such reasons, these people cannot appeal to such claims, or to the 'minimal principle' that Smith states with (C).

Smith also claims that

> (D) we can be rationally required to have some desire when, and because, our having this desire would make our set of desires more coherent and unified.

To illustrate this requirement, Smith supposes that we want to help only some of the people whom we know to be in desperate need. Our desires would be more coherent, and would 'make more sense', Smith claims, if we wanted to help all of these people. But this claim assumes that

> (3) whenever someone is in desperate need, this fact gives us a reason to want to help this person.

If such facts did not give us such reasons, our desires would not be less coherent, or make less sense, if we wanted to help only some of these people. And (3) is another claim about object-given reasons, to which Subjectivists cannot appeal.

Consider next Smith's claim that we can be rationally required to have a more unified set of desires. Mere unity is not a merit. Our desires would be more unified if we were monomaniacs, who cared about only one thing. But if you cared about truth, beauty, and the future of humanity and I cared only about my stamp collection, your less unified set of desires would not be, as Smith's claim seems to imply, less rational than mine. Smith might reply that my set of desires would be more impressively unified if I had several coherent desires. But if I also wanted to collect match-boxes, drawing pins, ticket stubs, and plastic cups, your less unified set of desires would still be more rational than mine. And this appeal to coherence would again assume that we have object-given reasons to have our desires. Subjectivists deny that we have such reasons.

There are other problems. If we don't care about some of our future agony, our desires would be more coherent if we didn't care about any of our future agony. For all these reasons, Subjectivists cannot claim that, if we were procedurally rational, we would want to avoid all future agony.

Since Subjectivists cannot defend this claim, my earlier conclusion stands. Subjectivists must claim that, in *Case One*, I would have no reason to want to avoid my future period of agony. As I have said, we can argue:

> We all have a reason to want to avoid, and to try to avoid, all future agony.
>
> Subjectivism implies that we have no such reason.

Therefore

> Subjectivism is false.

Some Subjectivists might now bite the bullet, by denying that we have this reason. In *Case One*, these people might say, though the approaching flames threaten to cause me excruciating pain, this fact does not count in favour of my wanting and trying to move my hand away. But that is hard to believe.

We can next remember why Subjectivism has these implications. Since Subjectivists deny that we have object-given reasons, they must agree that, on their view,

> (E) the nature of agony gives us no reason to want to avoid being in agony.

We can argue:

> The nature of agony does give us such a reason.

Therefore

> Subjectivism is false.

These arguments are, I believe, decisive.

Subjectivists might protest that, in denying (E), we are not *arguing* against their view, but are merely rejecting this view. If that is so, our claim could instead be that everyone ought to reject this view, since

(E) is a very implausible belief. Subjectivists are not *Nihilists*, who deny that we have any reasons. These people believe that we have reasons for acting. If we can have some reasons, nothing is clearer than the truth that, in the reason-implying sense, it is bad to be in agony. It can be hard to remember accurately what it was like to have sensations that were intensely painful. Some of the awfulness disappears. But we can remember such experiences well enough. According to Subjectivists, what we remember gives us no reason to want to avoid having such intense pain again. If we ask 'Why not?', Subjectivists have, I believe, no good reply.

4

Further Arguments

12 The All or None Argument

We have reasons, I have claimed, to have certain telic desires, such as a reason to want to avoid all future agony. We can now ask whether, as Subjectivists claim, our telic desires give us reasons.

Suppose that, in

> *Case Two*, I want to *have* some future period of agony. I am not a masochist, who wants this pain as a means to sexual pleasure. Nor am I repentant sinner, who wants this pain as deserved punishment for my sins. Nor do I have any other present desire or aim that would be fulfilled by my future agony. I want this agony as an end, or for its own sake. I have no other present desire or aim whose fulfilment would be prevented either by this agony, or by my having my desire to have this agony. After ideal deliberation, I decide to cause myself to have this future agony, if I can.

Subjective theories here imply that I have a decisive reason to fulfil my desire and act on my decision, by causing myself to be in agony. If there is a fire nearby, and I shall have no other way to fulfil my desire, I would have a decisive reason to thrust my hand into this fire. That is hard to believe.

In response to this objection, Subjectivists might reply that *Case Two* cannot be coherently imagined. Some writers claim that, if we really believed that it would be *us* who would later be in agony, and we also

understood what this agony would be like, it is inconceivable that we might want ourselves to be later in this state. But this claim is false. We can want what we know will be bad for us. It makes sense to suppose that someone wants to have some future period of agony, for its own sake. Nor could Subjectivists claim that, if we had this desire, that would make it impossible for us to be rational agents, who act for reasons.

Though it is conceivable that someone might want future agony for its own sake, this case *is* hard to imagine. This fact may seem to weaken this objection to subjective theories.

The opposite is true. This fact *strengthens* this objection. If we find it hard to imagine that anyone might have this desire, that is because we assume what objective theories claim. The nature of agony, we believe, gives everyone very strong reasons to want *not* to be in this state. According to subjective theories, we have no such object-given reasons. If that were true, it would *not* be hard to imagine that someone might want, for its own sake, to have some future period of agony. We could at most claim that this desire would be unusual, like the bizarre sexual desires that some people have. This case is hard to imagine because the awfulness of agony gives everyone such clear and strong reasons *not* to have this desire. It is hard to believe that anyone could be so irrational.

In an attempt to answer this objection, Subjectivists might now revise their view. They might claim that

> (F) for some desire or aim to give us a reason, we must have some reason to have this desire or aim.

If Subjectivists could appeal to (F), they could claim that, since I have no reason in *Case Two* to want to have some future period of agony, their theory does not imply that I have any reason to fulfil this desire.

To assess this reply, we can suppose that, in

> *Case Three*, I want to *avoid* some future period of agony.

Could Subjectivists claim that I have some reason to have this desire?

We are supposing that, in our examples, we know all of the relevant facts, and we have gone through some process of ideal deliberation. Subjective theories imply that, in such cases,

> (G) for us to have a reason to have some desire or aim, we must have some present desire or aim that gives us this reason.

There is one straightforward way in which we might be claimed to have some desire-based or aim-based reason to want to avoid some future period of pain. Subjective theories imply that

> (H) if some possible event would have effects that we want, or would help us to achieve some aim, this fact gives us a reason to want this event as a means to these effects, or to the achievement of this aim.

Suppose that, if I have a headache while I am playing in some chess match this afternoon, my pain would distract me, and would deny me the victory that I want. Subjective theories then imply that I have a reason to want to avoid this headache as a means of helping me to win this game, thereby fulfilling my desire. But we can suppose that, in *Case Three*, I have no such instrumental reason to want to avoid my future period of agony. Since this period would be fairly brief, my avoiding this agony would not have any other effects that I want, or help me to fulfil or achieve any of my other present desires or aims. On these assumptions, (H) does not imply that I have any reason to want to avoid this agony.

Subjectivists might also claim that

(I) when it is true either that

> (a) our *having* some desire or aim would have effects that we want,

or that

> (b) we *want* to have this desire or aim,

> these facts give us a reason to have this desire or aim, or at least give us a reason to cause ourselves to have or to keep this desire or aim, if we can.

But in *Case Three* I might have no such reasons. Suppose first that I cannot avoid my future period of agony. Partly for this reason, my desire to avoid this agony has no effects that I want. And this desire has some effects that I don't want, since it fills me with anxiety about what lies ahead. For these reasons, I don't want to have this desire. On these assumptions, (I) does not imply that I have any reason to have or to keep this desire.

Since I have no *other* present desire or aim that gives me any desire-based or aim-based reason to want to avoid this agony, Subjectivists might now claim that this desire *itself* gives me such a reason. To defend this claim, Subjectivists might say that

> (J) when we have some present fully informed desire or aim, this fact gives us a reason to have this desire or aim.

If (J) were true, all such desires or aims would be rationally self-justifying. My desire to avoid this agony would give me a reason to have this desire. But if I wanted to *be* in agony, this fact would give me a reason to want to be in agony. If I wanted to waste my life, this fact would give me a reason to want to waste my life. *Whatever* we want, our having such informed desires would give us reasons to have them. Since these claims are clearly false, Subjectivists must reject (J). Because Subjectivists cannot appeal to (J), these people must agree that, in this version of *Case Three*, my desire to avoid my future agony gives me no reason to have this desire. Since I have no other present desire or aim that gives me any reason to have this desire, these people must now admit that, on their view, I have no reason to want to avoid this agony.

Suppose next that, in a different version of this case, I *could* avoid this future agony. My having this desire would then lead me to do what would avoid this agony, thereby fulfilling this desire. This fact might be claimed to give me a desire-based reason to have this desire. Subjectivists might say that

> (K) if our having some fully informed desire would lead us to do what would fulfil this desire, this fact would give us a reason to have this desire.

But if (K) were true, all such fulfillable desires would be rationally self-justifying. If our wanting to be in agony would lead us to thrust our hand into some fire, this fact would give us a reason to want to be in agony. If our wanting to waste our lives would lead us to waste our lives, this fact would give us a reason to want to waste our lives. Since these claims are clearly false, Subjectivists must reject (K). These people must again admit that, on their view, I have no reason to want to avoid my future period of agony. So subjective theories imply that, in both versions of *Case Three*, I have no reason to have this desire.

There are many actual cases of this kind. When we want to avoid some future period of agony, or lesser pain, it is often true that, even after ideal deliberation, we would have no other present desire or aim whose fulfilment would be prevented by this future pain, and no present desire or aim that could be claimed to give us a desire-based or aim-based reason to want to avoid this pain. So subjective theories imply that we often have no reason to want to avoid some future period of pain.

Similar claims apply to many other kinds of case. When we want ourselves or others to have some future period of happiness, or we have other good or rational aims, it is often true that, even after ideal deliberation, we would have no other present desire or aim that would be fulfilled by the achievement of these aims, and no other desire or aim that could be claimed to give us a reason have these aims. That is often true because we want such things for their own sake, not as a means of fulfilling other desires. So subjective theories imply that we often have no reason to want ourselves or others to have such periods of happiness, and no reason to have several other good or rational aims.

Return now to the claim that

> (F) for some desire or aim to give us a reason, we must have some reason to have this desire or aim.

We have seen that, in *Case Three*, I have no desire-based or aim-based reason to have my desire to avoid my future agony. So if Subjectivists accepted (F), they would have to claim that my desire to avoid this agony does not give me any reason for acting. Even if I could easily fulfil this desire by moving my hand away from the flames of some approaching fire, I would have no reason to act in this way. This claim contradicts all subjective theories, and is clearly false. So Subjectivists cannot appeal to (F).

There is another reason why Subjectivists cannot claim that, for some desire to give us a reason, we must have some reason to have this desire. On these people's theories, as we have seen, any such reason would have to be provided by some other desire. For this other desire to give us this reason, (F) implies, we must have some reason to have this desire. On subjective theories, this reason would also have to be provided by some *other* desire, and so on for ever. We could not have any such beginningless chain of desire-based reasons and desires. Any such chain must begin with, or be grounded on, some desire that, according to these theories, we have no reason to have. So if these Subjectivists appealed to (F), they would have to conclude that none of our desires give us reasons, thereby denying their theory's main claim.

Since Subjectivists cannot appeal to (F), they must admit that, on their theories,

> (L) we have most reason to do what would best fulfil or achieve our present fully informed telic desires or aims, *whatever* we want, and whether or not we have *any reason* to have these desires or aims.

Similar claims apply to the choices that we would make after ideal deliberation.

We can now return to *Case Two*, in which I want to have some future period of agony, not as a means, but as an end, or for its own sake. I have no other present desire or aim that would be either fulfilled or prevented by this future agony, or by my desire to have this agony. After ideal deliberation, I have decided to cause myself to have this agony, if I can. Since Subjectivists must accept (L), they must admit that, on

their view, I have most reason to cause myself to be in agony for its own sake. This act would best fulfil my present fully informed telic desires, and is what, after ideal deliberation, I have chosen to do. If there is a fire nearby, and I have no other way to fulfil my desire, I would have a decisive reason to thrust my hand into this fire. That is very hard to believe. Given my description of this case, there are, I believe, no facts that count even weakly in favour of my thrusting my hand into this fire. And I would have decisive reasons not to cause myself to be in agony in this way.

There could be many other, similar cases. According to subjective theories, if we had such informed desires to hit our howling baby, or to smash some malfunctioning machine, these facts would give us reasons to hit our baby and smash this machine. If what we most wanted and chose was to frustrate all of our future desires, this fact would give us a decisive reason to frustrate all of these desires. If what we most wanted and chose was to waste our lives, and to achieve other bad or worthless aims, these facts would give us decisive reasons to waste our lives, and to try to achieve these bad or worthless aims. These claims are also very hard to believe. These implications of subjective theories give us decisive reasons, I believe, to reject all such theories.

Subjectivists might reply that, though *these* desires and choices would not give us any reasons for acting, that does not show that *no* desires or choices give us reasons. These people must admit that, in *Case Two*, my desire to be in agony gives me no reason for acting. But Subjectivists might claim that, in *Case Three*, my desire *not* to be in agony *does* give me a reason. These people might similarly claim that, though we would have no reasons to fulfil our desires if what we wanted was to suffer in other ways, to waste our lives, or to achieve other bad or worthless aims, we *do* have reasons to fulfil our desires when what we want is to be happy, to live productive and worthwhile lives, or to achieve other good aims.

Subjectivists *cannot*, however, make such claims. These claims appeal to differences between the reason-giving features of the *objects* of these desires or aims. If we make such claims, we have moved to an objective theory, which appeals to such object-given reasons. Subjectivists cannot distinguish in these ways between desires or aims that do or don't give

us reasons. We are considering cases in which we know all the relevant facts. In such cases, we can argue:

> If we have desire-based reasons for acting, all that would matter is *whether* some act would fulfil the telic desires that we now have after ideal deliberation. It would be irrelevant *what* we want, or would be trying to achieve.

Therefore

> Either all such desires give us reasons, or none of them do.
>
> If all such desires gave us reasons, our desires could give us decisive reasons to cause ourselves to be in agony for its own sake, to waste our lives, and to try to achieve countless other bad or worthless aims.
>
> We could not have such reasons.

Therefore

> None of these desires gives us any reason. We have no such desire-based reason to have any desire, or to act in any way.

We can call this *the All or None Argument*. Similar arguments apply to aim-based and choice-based reasons.

When we want to avoid agony, or to be happy, or we have other good or rational aims, we do indeed have reasons to try to fulfil these desires and achieve these aims. But these reasons are provided, not by the facts that these acts would fulfil or achieve these desires or aims, but by the features of what we want, or have as our aims, that make these events good or worth achieving.

Here is an overlapping argument for this conclusion. According to Objectivists, we have instrumental reasons to want something to happen, or to act in some way, when this event or act would have effects that we have some reason to want. As that claim implies, every instrumental reason gets its normative force from some other reason. This other reason may itself be instrumental, getting its force from some third

reason. But at the beginning of any such chain of reasons, there must be some fact that gives us a reason to want some possible event as an end, or for its own sake. Such reasons are provided by the intrinsic features that would make this possible event in some way good. It is from such telic value-based object-given reasons that all instrumental reasons get their normative force.

Subjectivists must reject these claims. According to these people, instrumental reasons get their force, not from some telic reason, but from some telic desire or aim. We can have desire-based reasons to have some desire, and we can have long chains of instrumental desire-based reasons and desires. But at the beginning of any of *these* chains, as we have seen, there must always be some desire or aim that we have no such reason to have. And as my examples help us to see, we cannot defensibly claim that such desires or aims give us reasons. I would have no reason to thrust my hand into the fire. We would have no reason to hit our howling baby, or to waste our lives, or to try achieve countless other bad or worthless aims. So subjective theories are built on sand. Since all subject-given reasons would have to get their normative force from some desire or aim that we have no such reason to have, and such desires or aims cannot be defensibly claimed to give us any reasons, we cannot be defensibly claimed to have any subject-given reasons. We cannot have any such reasons to have any desire or aim, or to act in any way.

13 The Incoherence Argument

Subjectivists might again protest that my arguments have appealed to merely imaginary cases. When applied to actual cases, these people might claim, subjective theories have acceptable implications.

As I have said, however, good theories about reasons must be able to be applied successfully to merely imaginary cases. Nor have I appealed only to such cases. I have argued that, in many actual cases, subjective theories imply that we have no reasons to want ourselves or others to avoid future periods of agony, or to have future periods of happiness, and no reason to have many other good aims. And though subjective theories often have acceptable implications, this fact does not support

these theories, since these theories have such implications only when they overlap with the best objective theories.

To illustrate this third point, let us compare two kinds of epistemic theory. According to

> *the reason-based theory*, we ought to believe what the facts that are known to us give us decisive reasons to believe.

According to an implausible imaginary theory, which we can call

> *the belief-based theory*, we ought to believe whatever, after considering the facts, we would in fact believe.

When applied to actual people, this belief-based theory would often have acceptable implications. Since most of us often believe what the facts that we have considered give us decisive reasons to believe, this belief-based theory often implies that we ought to believe what we have such decisive reasons to believe. But that is not what this theory claims. In its claims about what we ought to believe, this theory implies that we have no reasons to have our beliefs. When this belief-based theory has acceptable implications, that is because most actual people assume that they do have such reasons, and often have beliefs that respond to these reasons. So we should reject this theory.

Similar claims apply to theories about what we ought to do. According to what we can here call

> *objective reason-based theories*, we ought to try to achieve the aims which the facts that are known to us give us decisive reasons to have.

According to

> *subjective aim-based theories*, we ought to try to achieve the aims which, after considering the facts, we would in fact have.

When applied to actual people, these subjective theories often have acceptable implications. Since most of us often have the aims which the facts that we have considered give us decisive reasons to have, subjective theories often imply that we ought to try to achieve these aims. But that is

not what these theories claim. In their claims about what we ought to do, these theories imply that we have no reasons to have our aims. When subjective aim-based theories have acceptable implications, that is because most actual people assume that they do have such reasons, and often have aims that respond to these reasons. So we should reject these theories. These theories can seem plausible, we might say, only because most people do not believe what these theories claim.

Many Subjectivists do not fully believe what their own theory claims. We have been discussing cases in which we know all of the relevant facts. In many cases, however, we do not know all these facts. Many Subjectivists claim that, in these other cases,

> (M) what we have most reason to do is whatever would best fulfil, not our actual present telic desires or aims, but the desires or aims that we would now have, or would want ourselves to have, if we knew and had rationally considered all of the relevant facts.

Many of these people also claim that

> (N) when we are making important decisions, we ought if we can to try to learn more about the different possible outcomes of our acts, so that we can come to have better informed telic desires or aims, and can then try to fulfil these desires or aims.

Subjectivists cannot, I believe, coherently make these claims. When we ought to try to find out and rationally consider certain facts, that is because these facts might give us certain reasons. Juries, for example, ought to consider the facts that might give them reasons to believe that some accused person did, or did not, commit some crime. We can similarly claim that, when we are deciding which outcomes we shall try to bring about, we ought in important cases to try to discover, and rationally consider, what these outcomes would be like. But if we make this claim, we are assuming that

> (O) these possible outcomes may have intrinsic features that would give us object-given reasons to want either to produce or to prevent these outcomes, if we can.

And (O) is what *Objectivists* believe. Subjectivists deny (O). According to these people, no such features of possible outcomes ever give us such reasons. If that were true, we would have no reason to try to discover, and rationally consider, what these outcomes would be like. So these people cannot coherently assert (N).

Nor can they coherently assert (M). If (O) were false, as Subjectivists claim, we would have no reason to believe that what we have most reason to do is whatever would best fulfil, not our actual present desires or aims, but the desires or aims that we would now have if we had rationally considered all of the facts about the possible outcomes of our acts. Subjectivists cannot call these the *relevant, reason-giving* facts, since these people deny that these facts give us reasons. And if these facts could not give us reasons to have these desires or aims, we would have no reason to accept (M). We would have no reason to believe that these better informed desires or aims have any higher reason-giving status, or are desires or aims that we have more reason to try to fulfil.

Some Subjectivists make the weaker claim that

> (P) we have reasons to fulfil only those of our present telic desires or aims that are error-free, in the sense that these desires do not depend on false beliefs.

To defend this claim, however, these people would also have to appeal to (O), which Subjectivists cannot do. If we had no object-given reasons, as these people believe, we would have no reason to want to know more about what we want, either by getting new true beliefs, or by losing our present false beliefs.

Some Subjectivists recognize these implications of their theories. When Korsgaard defends the view that our rationally choosing something makes this thing good, she writes that this view

> frees us from assessing the rationality of a choice by means of the . . . task of assessing the thing chosen: we do not need to identify especially rational ends.

To choose rationally, on Korsgaard's view, we needn't assess the merits of what we choose, since nothing *has* any such merits, by having any

reason-giving features. But most Subjectivists do not see that, given their assumptions, we have no reason to try to have and to fulfil such better informed desires or aims. If Subjectivists cannot appeal to (M), (N), or (P), as I have just argued, that undermines the subtler and more plausible versions of Subjectivism, such as the Deliberative Theory and the Informed and Error-Free Desire Theories. These theories are incoherent, since they assume both that

> (Q) our desires, aims, or choices give us reasons only if we would still have these desires and aims, or make these choices, if we had true beliefs about all the relevant intrinsic features of what we want,

and that

> (R) these features give us no reasons to want these things.

If these features gave us no such reasons, that would undermine the claim that, for our desires to give us reasons, they must be desires that we would still have if we had true beliefs about these features. We can call this *the Incoherence Argument against Subjectivism*. This objection, we can note, is quite separate from my earlier arguments, since this objection makes no appeal to claims about which facts give us reasons.

The Incoherence Argument does not apply to the simpler, Telic Desire Theory, which claims only that

> (S) we have most reason to do whatever would best fulfil or achieve our actual present telic desires or aims.

We have such reasons, this theory claims, whether or not our telic desires or aims rest on false beliefs. These Subjectivists can coherently claim that

> (T) we ought to try to discover the facts about how we can best fulfil our present telic desires or aims.

These people can make this claim because (T) does not assume that the things we want, or the possible outcomes of our acts, may have intrinsic reason-giving features. On the Telic Desire Theory, the relevant facts do

not include facts about what these outcomes would be like, except when these are facts about what would best fulfil our actual present desires or aims. These Subjectivists can also coherently claim that

> (U) if we *want* to have such better informed desires or aims, we ought to try to discover the facts about what the different possible outcomes would be like, so that we can have such desires or aims.

These people might then claim that, since most of us *do* want to have such better informed desires or aims, (U) implies that most of us ought to try to have them. But as before, these claims would not support the Telic Desire Theory. Most of us want to have better informed desires or aims because we believe what objective theories claim. The possible outcomes of our acts, we believe, may have features that would give us reasons.

Though the Telic Desire Theory is not incoherent, it has several implausible implications which have led many Subjectivists to move to the other subjective theories discussed above. And my other objections apply. On this theory, we often have no reason to want to avoid future agony, or to be happy, and we might have decisive reasons to cause ourselves to be in agony for its own sake, to waste our lives, and to try to achieve other bad or worthless aims.

The Incoherence Argument, I have claimed, undermines the subtler and more plausible versions of Subjectivism. There is another, more positive way to state what this argument shows. When many Subjectivists appeal to claims about what we would want or choose if we knew all the facts about the possible outcomes of our acts, these people rightly assume that these outcomes may have reason-giving features. Most of these people assume, for example, that we have object-given reasons to want to be happy, and to avoid agony. These people are not really Subjectivists. When these people make Subjectivist claims, they are not correctly stating what they actually believe.

One such person, I believe, is Frankfurt. In deciding what to care about, Frankfurt writes, we don't need to understand what is *important*. It is enough to understand what is *important to us*. To illustrate his view, Frankfurt imagines a group of people whose health is threatened by background radiation. Suppose, he writes, that

> someone genuinely does not care a bit about his health . . . In that case, the level of background radiation is not important to him. It truly does not matter to him; he has no reason to care about it.

Frankfurt here assumes that, if we don't care about something, we have no reason to care about this thing. On this view, we don't need to understand what is important because *nothing* is important. The truth is only that some things are important *to certain people*, in the sense that these people care about these things.

Giving another illustration, Frankfurt writes:

> Suppose that what a person cares about is avoiding stepping on the cracks in the sidewalk. No doubt he is committing an error of some kind in caring about this . . . his error consists in caring about, and thereby imbuing with genuine importance, something which is not worth caring about.

This last phrase may seem to imply that some things are worth caring about, and others aren't. But Frankfurt continues

> The reason it is not worth caring about seems clear: it is not important to the person to make avoiding the cracks in the sidewalk important to himself.

If it *was* important to this person to make avoiding the cracks important to him, this remark implies, this person would not be committing any kind of error. Avoiding the cracks would then have genuine importance for this person, and be something that was worth caring about. Whatever we care about, Frankfurt elsewhere claims, we thereby answer the 'question of how to live'.

It may be objected, Frankfurt writes, that

> an empirical account of what people actually care about . . . would miss the whole point of our original concern with the problem of what sort of life one *should* live.

Frankfurt replies

> It is not the factual question about caring that misses the point, but the normative one. If we are to resolve our difficulties and hesitations in settling on a way to live, what we need most fundamentally are not reasons or proofs. It is clarity and confidence.

While he is arguing that we should reject the normative question, Frankfurt imagines someone who first decides what he cares about, and then

> wonders whether he has got it right . . . he becomes concerned about whether he really should care about the things that, as a matter of fact, he does care about.

This man's concern, Frankfurt claims, is misguided. Though different people care about different things, we 'do not need to decide who is right'. By caring about or loving something, Frankfurt adds, we can give meaning to our lives, even if we recognize that what we love is bad.

Though this last claim is true, Frankfurt's other claims are, I believe, mistaken. Frankfurt writes

> Love is itself, for the lover, a source of reasons.

We might similarly claim

> Hate is, for the hater, a source of reasons.

Hitler's hatred of the Jews gave the last part of his life the kind of meaning that Frankfurt describes. And Hitler had the 'clarity and confidence' that Frankfurt claims to be more important than having reasons. It matters greatly, I believe, whether Hitler had reasons to do what he did, and whether, in our loves or hates, it is we or Hitler who got things right.

Frankfurt defends his claims in a puzzling way. When someone wonders whether he has got things right, Frankfurt writes,

> (V) 'he is asking . . . whether there may not be better reasons for him to live in some other way instead'.

If we ask this normative question, Frankfurt argues, 'we are bound to find ourselves helplessly in a spin'. No attempt to answer this question could possibly succeed. Some people try to answer such questions, since they believe that certain ends or aims are good in themselves, having *inherent* or intrinsic value. These people claim that

> (W) when some end 'has inherent value . . . there is . . . some reason for choosing it', and that when some end 'has greater inherent value than anything else', this reason is decisive.

But this claim, Frankfurt writes,

> (X) 'does not so much as address, much less answer, the question of how a person's final ends are appropriately to be established. Even if the claim were correct it, would still provide *no account at all* of how people are to select the ends that they will pursue.'

(X) implies that, even if (W) were true, (W) would not even *address* the question of which ends we are to choose. If claims (V) and (W) use the word 'reason' in its normative sense, which we can also express with the phrase 'counts in favour', Frankfurt's (X) would be clearly false. If some fact counts decisively in favour of our selecting some end, that could help us to decide which end to select. This suggests that Frankfurt is not using the word 'reason' in its normative sense. But why then does he claim to be discussing the *normative* question whether we might have *better* reasons to live in some other way?

This puzzle can, I believe, be solved. In these passages, Frankfurt misdescribes his real view. When we are deciding how to live, Frankfurt also writes, we need to decide which of several possible ends we shall try to achieve. These ends include

> personal satisfaction, pleasure, glory, creativity, spiritual depth, and conformity with the requirements of morality.

Frankfurt's list does not include

> personal dissatisfaction, pain, dishonour, futility, spiritual shallowness, and immorality.

Given Frankfurt's list of the possible ends that we need to consider, Frankfurt seems to believe that some ends have intrinsic value in the reason-giving sense, and that others don't. And Frankfurt seems to use the concept of an object-given reason to have some end or aim. The truth seems to be only that, like Hume and some other great philosophers, Frankfurt is not fully aware of the way in which his responses to such reasons guide his thoughts about what we should care about, and do.

When Frankfurt rejects appeals to intrinsic value, he writes:

> There is among philosophers a recurrent hope that there are certain final ends whose unconditional adopting might be shown to be in some way a requirement of reason.

On what seems to be Frankfurt's real view, there are many intrinsically good ends, but no ends have supreme value. Nor are there precise truths about which ends are most worth achieving. We often have to choose between many good ends or aims, none of which is clearly better than the others, and in such cases there is no end that reason requires us to choose. These plausible claims are very different from the view that no ends are in themselves good. Frankfurt, I suggest, is not a Nihilist about intrinsic goodness, but a Pluralist.

Frankfurt also rightly assumes that reason may require us to adopt some good ends. We need to have goals, Frankfurt claims, and we need productive work, so that our lives are not empty of meaning. We can sometimes give our lives meaning, by getting ourselves to have some goal that is worth achieving. In such cases, Frankfurt assumes, we have decisive reasons to give our lives meaning by adopting some such goal.

When Frankfurt criticizes what he calls overly *rationalist* theories, he makes the important claim that we don't need reasons for loving people. But this claim does not imply that, as Frankfurt sometimes suggests, love is the source of all our reasons. Though we don't need reasons for loving *particular* people, we have reasons to try to love some people, since love is in itself good. Love in this way differs from hate. And

though love is the source of some of our reasons, these are often reasons to do what is good for the people whom we love, by making their lives go in ways in which we and they have other reasons to want these lives to go. Love takes its simplest and clearest form, as Frankfurt notes, in the love of parents for their young children. Parents have strong reasons to hope that their children will get things right, adopt good ends, and will not care much, except briefly and for fun, about avoiding cracks in sidewalks.

14 Reasons, Motives, and Well-Being

We can now return to the ways in which events or outcomes can be good or bad. Of two possible events, one would be

> *better* in the *impartial-reason-implying* sense if this is the event that, from an impartial point of view, everyone would have more reason to want, or to hope will happen.

According to subjective theories about reasons, no events could be in this sense better than others, since there are no events that, from an impartial point of view, everyone would have more reason to want. It could not be better, for example, if some child's life were saved. There have been many people whose fully informed desires would not be better fulfilled when any child's life were saved. And even if everyone had such desires, subjective theories do not imply that everyone has *reasons* to have these desires, by having reasons to want any child's life to be saved. But that is what is meant by the claim that, in this impartial-reason-implying sense, it would be better if some child's life were saved.

Events can also be better *for* particular people, in the sense of making these people's lives go better, or contributing more to their well-being. Theories about well-being can differ in two ways, since they can use the phrase 'good for' in different senses, and they can make different claims about what would be good for people in these senses. On all plausible theories, everyone's well-being consists at least in part in being happy, and avoiding suffering. But different theories make partly conflicting claims about what else would be good or bad for people.

When we call some possible life

> 'best for someone' in the *reason-implying* sense, we mean that this is the life that this person would have the strongest self-interested reasons to want to live, and the life that other people would have the strongest reasons to want or hope, for this person's sake, that this person will live.

As I have said, 'self-interested' does not mean 'selfish'. Even the most altruistic people have reasons to care about their own future well-being.

If we accept some subjective theory about reasons, we cannot use 'best for someone' in this reason-implying sense. Subjective theories imply that there are no self-interested reasons. Such reasons are provided by facts about the intrinsic features of future events that would make these events good or bad for us. Subjectivists deny that we have such reasons.

Some Subjectivists claim that we can have a different kind of self-interested reason. According to these people, since most of us do care about our future well-being, most of us have *desire-based* self-interested reasons. These Subjectivists also claim that, since most of us care about morality, most of us have desire-based moral reasons. On this view, however, if we don't have these desires, we have no such reasons. In my imagined *Cases One* and *Two*, I would have no self-interested reason to try to avoid my future agony. And given Hitler's desires, Hitler may have had no moral reason not to commit mass murder. Though Subjectivists are free to use words as they wish, it is misleading to call such desire-based reasons *self-interested* or *moral*. As most of us use these words, no good theory is about self-interested reasons unless this theory implies that we all have self-interested reasons to try to avoid being in agony. And no good theory is about moral reasons unless this theory implies that we all have moral reasons not to commit mass murder. So we can justifiably claim that, according to subjective theories, there are no self-interested or moral reasons.

Of those who accept subjective theories about reasons, many use 'best for someone' in some sense that differs from the reason-implying sense.

One example is the definition proposed by Rawls when he presents his *thin theory of the good*. On this definition,

> a person's good is determined by what is for him the most rational plan of life.

Some life would be best for someone, Rawls writes, if this life would fulfil the plan that this person

> would adopt if he possessed full information. It is the objectively rational plan for him and determines his real good.

If we call some life

> 'best for someone' in this *present-choice-based* sense, we mean that this is the life that, after fully informed and procedurally rational deliberation, this person would adopt, or choose.

Though it is a normative question which kinds of deliberation are procedurally rational, and in other ways ideal, it is a psychological question what, after such deliberation, someone would in fact choose. On such views, there are no telic desires or aims that we are all rationally required to have, except perhaps those desires without which we could not even deliberate, choose what to do, and act. The most rational plan of life for someone, Rawls writes, is the plan

> which would be chosen by him with full deliberative rationality, *that is*, with full awareness of the relevant facts and after a careful consideration of the consequences.

We can be deliberatively rational in Rawls's sense whatever we have as our aims or ends. Rawls elsewhere claims that, from the fact that someone is *ideally rational*, we can infer nothing about what this person does or would want, or approve. There is nothing, Rawls assumes, that we have any object-given reasons to want as an end.

To illustrate his theory of the good, Rawls imagines a man whose chosen plan is to spend his life counting the numbers of blades of grass in various lawns. Rawls writes that, on his theory, 'the good for this man is indeed counting blades of grass'. This imagined man, Rawls assumes,

would enjoy spending his life in this way. But on Rawls's theory, that assumption is not needed. It would be enough that, after rationally considering the relevant facts, this man would in fact choose this plan of life. For another example, consider

> *Blue's Choice*: After such ideal deliberation, *Blue's* strongest desire is that the rest of his life consists only of unrelieved suffering. Blue therefore chooses some plan that would give him such a life.

On Rawls's theory, the best life for Blue would consist of unrelieved suffering.

This example might be claimed to be unrealistic, because no one would choose a life of unrelieved suffering. As I have said, however, it is irrelevant whether such cases actually occur. Rawls does not assume that any actual person would choose to spend his life counting blades of grass, and Rawls rightly applies his theory to his merely imagined man. Any acceptable normative theory must be able to be applied successfully to such imaginary cases. And though it is hard to believe that anyone would choose a life of unrelieved suffering, that is because it is hard to believe that anyone could be so irrational as to choose a life that is so obviously bad in the reason-implying sense. On Rawls's view, however, no life could be bad for someone in this sense, since we have no object-given reasons. In Rawls's words, 'There is no way to get beyond deliberative rationality.'

My example is, in one way, no objection to Rawls's theory of the good. When Rawls claims that some life would be best for someone, or would be this person's real good, he is using these phrases in his proposed present-choice-based sense. Rawls means that this is the life that, after ideal deliberation, this person would in fact choose. Blue, we have supposed, would choose a life of unrelieved suffering. So Rawls would be *right* to claim that, in his proposed sense, this is the life that would be best for Blue. That is merely another way of saying that this is the life that, after such deliberation, Blue would choose.

Rawls intends, however, to be claiming more than this. Rawls's proposed sense of 'best for someone' is intended to replace the ordinary sense of this phrase, by giving us a clearer way of saying everything that

we might want to say. And Rawls, I believe, would want to say that it would be better for Blue if Blue's life did not consist of unrelieved suffering.

Rawls could make that claim if he used 'best for someone' in some other sense. Since Rawls is a Subjectivist about Reasons, he cannot use 'best for someone' in the reason-implying sense. But this phrase is often used in other senses. When people call some possible life 'best for someone', some of them mean that

> this is the possible life in which this person would have the greatest sum of happiness minus suffering,

and others mean that

> this is the possible life in which this person's desires at different times would be best fulfilled.

We can call these the *hedonistic* and *temporally-neutral desire-based* senses of the phrase 'best for someone'. Rawls could truly claim that, in these senses, it would be bad for Blue to have his life of unrelieved suffering. This life would be hedonically very bad for Blue. And though such a life would best fulfil Blue's desires at the time when he chooses this life, his desires in the rest of his life would be much less well fulfilled.

There is, however, little point in claiming that, in these senses, this life would be bad for Blue. In the hedonistic sense, this claim would be another concealed tautology, whose open form would be the trivial claim that, if Blue's life contained more suffering, it would contain more suffering. In the temporally-neutral desire-based sense, this claim would be fairly trivial, since it would mean only that, if Blue's life contained more suffering, his desires would be less well fulfilled. Similar remarks apply to other cases. When people use 'best for someone' in either of these senses, they cannot have substantive normative beliefs about which lives would be best for people.

These people *could* have such beliefs if they accepted some objective theory about reasons, so that they could *also* use 'best for someone' in the reason-implying sense. They might then claim:

> (V) If some possible life would be best for someone in both the hedonistic and the temporally-neutral desire-based sense, these facts would make this the life that would be best for this person in the reason-implying sense.

This means:

> (W) If some possible life would both give someone the most happiness, and be the life in which this person's desires would on the whole be best fulfilled, these facts would make this the life that this person would have the strongest self-interested reasons to want, and to try to live, and the life that other people would have the strongest reasons to want or hope, for this person's sake, that this person will live.

This claim *is* substantive, and plausible. But if we accept some subjective theory about reasons, we cannot make such claims.

Subjectivists about Reasons might use other senses of 'best for some one'. But that would not help them to avoid implausible conclusions. Blue's strongest desire and chosen aim, after ideal deliberation, is a life of unrelieved suffering. Subjective theories unavoidably imply that

> (X) even if a life of unrelieved suffering would be, in other senses, bad for Blue, this is the life that Blue has most reason now to give himself, if he can.

If Blue could now ensure that he will have such a life, by getting himself enslaved to some cruel master, or committing some crime for which the punishment is endless hard labour, this would be what, on subjective theories, Blue has most reason to do, and what, if he knew the facts, he ought rationally to do.

Similar claims apply to actual cases. Subjective theories imply that we have no object-given reasons to want ourselves or others to live happy lives, and no such reasons to have any other good aim. And, as I have argued, Subjectivists cannot defensibly claim that we have *subject-given* reasons to have such aims, or to care about anything for its own sake. Such reasons would have to be provided by some desire or aim that we

have no reason to have, and such desires or aims cannot be defensibly claimed to give us any reasons. So we can now conclude that, on these widely accepted views, *nothing matters*.

Some Subjectivists would admit that, on their view, nothing matters in an impersonal sense. It is enough, these writers claim, that some things matter to particular people. But this reply shows how deep the difference is between the two kinds of theory that we have been considering. According to objective theories, some things matter in the normative sense that we have *reasons* to care about these things. When Subjectivists claim that some things matter to particular people, they mean only that these people *do* care about these things. That is not a normative but a merely psychological claim. We all know that people care about certain things. We hoped that philosophers, or other wise people, would tell us more than that.

As well as implying that nothing matters, subjective theories cannot even defensibly claim that we have any reasons for acting. As I have argued, our desires, aims, and choices cannot be defensibly claimed to give us any such reasons.

15 Arguments for Subjectivism

These bleak views are seldom defended. Most Subjectivists take it for granted that reasons are provided by certain facts about our desires or aims.

Of those who defend subjective theories, some appeal to a version of the claim that 'ought' implies 'can'. These people argue:

> (1) For us to have a reason to do something, it must be true that we *could* do it.
>
> (2) We couldn't do something if it is true that, even after ideal deliberation, we would not want to do this thing, or would not be motivated to do it.

Therefore

> For us to have a reason to do something, it must be true that after such deliberation, we *would* be motivated to do this thing.

But (2) is not relevantly true. Suppose I say, 'You ought to have helped that blind man cross the street', and you say, 'I couldn't have done that'. If I ask 'Why not?', it would not be enough for you to reply, 'Because I didn't want to'. Except in certain special cases, we *could* do something, in the relevant sense, if nothing stops us from doing this thing except the fact that we don't want to do it.

Some Subjectivists argue:

> (3) If we have some normative reason, we might act for this reason.
>
> (4) If we acted for this reason, we would be motivated to act in this way.
>
> (5) Since we would be motivated to act in this way, this reason would be desire-based.

Therefore

> All reasons for acting are desire-based.

But (5) is false. We cannot defensibly claim that, whenever people are motivated to act for some reason, this reason must be subject-given and desire-based *rather* than object-given and value-based. That claim would have to assume that, for some reason to be object-given and value-based, it must be impossible for anyone to be motivated to act for this reason. And that assumption would be absurd. If some act would achieve some aim that is good or worth achieving, we might be motivated to act for this reason.

These Subjectivists might reply

> (6) Whenever we act, we are motivated to act in this way, so we always have some desire-based reason for acting as we do.

Therefore

> (7) All reasons for acting are desire-based, even if some of these reasons might also be claimed to be value-based.

Therefore

> In our account of practical reasons, it is enough to appeal to some subjective desire-based theory.

But (6) either confuses normative and motivating reasons, or claims that, whenever we act, we thereby give ourselves a normative reason for acting as we do. That claim would falsely assume that, any act however crazy would partly justify itself. In taking (6) to imply (7), this argument also falsely assumes that we cannot have reasons on which we fail to act. And (7) falsely assumes that value-based reasons might also be desire-based.

There is another, much more important line of thought that leads many people to be Subjectivists. These people make some meta-ethical assumptions that I discuss in Part Six, and shall mention only briefly here. On the best objective theories, the fact that we have some reason is an *irreducibly normative* truth. Of those who accept subjective theories, many are *Metaphysical Naturalists*, who believe that there cannot be such facts or truths. According to these Naturalists, all properties and facts must be of the kinds that are investigated by the natural and social sciences. Irreducibly normative truths are incompatible, these people assume, with a scientific world-view.

Most of these Naturalists accept *reductive* desire-based or aim-based accounts of reasons for acting. According to some Analytical Subjectivists, when we claim that someone has a reason to act in some way, we *mean* that this act would fulfil one of this person's telic desires, or is what, after informed deliberation, this person would choose to do, or we mean something else of this kind. According to some other Naturalists, though the *concept* of a reason is irreducibly normative, the *fact* that someone has a reason is, or consists in, some such causal or psychological fact.

These reductive subjective theories can seem plausible if, like many people, we regard *normativity*, or the normative force of any reason,

as some kind of *motivating* force. We may then believe that we should identify reasons for acting with certain facts about what would fulfil our present desires, or about how we might be motivated to act. This may seem to be the best or the only way in which, as Metaphysical Naturalists, we can explain the normativity of these reasons. As some of these people write:

> For the philosophical naturalist, concerned to place normativity within the natural order, there is nothing plausible for normative force to be other than motivational force . . .
>
> there seems nothing for value to be, on deepest reflection, wholly apart from what moves, or could move, valuers, agents for whom something can matter.

Object-given value-based reasons cannot be regarded in such ways, since we have such reasons even if we would *not* be moved or motivated to act upon them.

Of the writers who give such reductive accounts, most claim to be describing normative reasons. But on such views, I believe, there aren't really any normative reasons. There are merely causes of behaviour. Things matter only in the sense that some people care about these things, and these concerns can move these people to act.

Such Naturalist accounts of reasons are, I believe, deeply mistaken. I defend this belief in Part Six, but I shall make one remark here. If Metaphysical Naturalism were true, we could not have reasons to have any particular beliefs. Such epistemic reasons are also irreducibly normative, and are therefore open to the same Naturalist objections. So it could not be true that we *ought* to accept Naturalism, nor could we have any reasons to accept this view. For us to be able to argue rationally about whether Naturalism is true, Naturalism must be false.

Naturalism, I believe, *is* false, and some things matter in the quite different sense that we have reasons to care about these things.

5

Rationality

16 Practical and Epistemic Rationality

We can now turn from reasons to rationality. As I have said, when we are aware of facts that give us certain reasons, we ought rationally to *respond* to these reasons. We respond to decisive reasons when our awareness of the reason-giving facts leads us to believe, or want, or try to do what we have these reasons to believe, or want, or do. We are irrational, or less than fully rational, insofar as we fail to respond to decisive reasons in these ways. To *fail* to respond to some reason, we must be aware of the facts that give us this reason.

While reasons are given by facts, what we can rationally want or do depends on our beliefs. If we have certain beliefs about the relevant, reason-giving facts, and what we believe would, if it were true, give us some reason, I am calling these *beliefs whose truth* would give us this reason. Such beliefs give us an *apparent* reason. When such beliefs are true, this apparent reason is also a real reason. These beliefs include assumptions of which we are not consciously aware, such as the assumption that some act would not harm ourselves or others. When we are ignorant, or have false beliefs, it may be rational for us to want, or do, what we have no reason to want, or do. We are then responding to what *merely appear* to be reasons. We ought rationally to respond to apparent reasons even if, because our beliefs are false, these reasons are not real.

We can next look more closely at how the rationality of our desires and acts depends on our beliefs. My claims about our desires would also apply to our aims. Our desires and acts *causally* depend on our

beliefs when we have these desires, and act in these ways, because we have these beliefs. Some desire might causally depend on some wholly irrelevant belief. We can imagine my wanting to go to sleep because I believe that 7 is a prime number. But if my desire directly depended on this belief, I would be mentally ill, or have some kind of local brain damage. 7's being a prime number gives me no reason to want to go to sleep. In most cases, when some desire depends on some belief, this relation is not merely causal. I may want to go to sleep because I believe that, unless I get some sleep, I shall perform badly in some interview tomorrow. Since this desire would be a rational response to what I believe, this desire would be not only caused by, but also *justified* by, my belief. I shall now briefly defend one view about how our desires and acts can be, or fail to be, justified by our beliefs.

The rationality of some of our desires depends only on their *intentional objects*, which are the possible events that we want, with the features that we believe these events would have. Such desires are rational when we want events whose features give us reasons to want them. It is always rational, for example, to want to avoid being in pain. The rationality of our other desires depends in part on our other beliefs about what we want. It is rational, for example, to want to take some medicine that we believe would both be safe and relieve our pain. Similar claims apply to our acts. The rationality of our acts depends on what we are intentionally doing, and may also depend on our other beliefs about what we are doing. On this view:

> (A) Our desires and acts are rational when they causally depend in the right way on beliefs whose truth would give us sufficient reasons to have these desires, and to act in these ways.

We can add:

> (B) In most cases, it is irrelevant whether these beliefs are true, or rational. Some of the exceptions involve certain normative beliefs.

> (C) When our beliefs are inconsistent, some of our desires or acts may be rational relative to some of our beliefs, but

> irrational relative to others. When we have no beliefs about the relevant, reason-giving facts, there may be nothing that we ought rationally to do.
>
> (D) Our having some desire is in one way rational when and because this desire itself is rational. But in some cases we could rationally cause ourselves to have some irrational desire. Our having this desire would then be, in a different way, rational. It could also be rational to cause ourselves to act irrationally. I discuss such cases further in Appendices A and B.

To be fully rational, we also need to meet certain rational requirements, such as requirements not to have contradictory intentions, and to intend to do what we believe that we ought to do. I shall not discuss these requirements here.

Many people would reject some of these claims. Our desires are irrational, Hume suggests, just when these desires causally depend on false beliefs. But false beliefs can be rational, and so can desires that depend on false beliefs.

On a much more widely held view, our desires are irrational just when they causally depend on *irrational* beliefs. To assess this view, we can suppose that I want to smoke because I want to protect my health and I believe that smoking is the most effective way to achieve this aim. I have this irrational belief because my neighbour smoked until he was aged 100, and I take this fact to outweigh all of the evidence that smoking kills. To simplify things, we can add that I don't enjoy smoking. I want to smoke only because I enjoy living, and I believe that smoking will prolong my life. Does the irrationality of my belief make my desire to smoke irrational?

It is best, I suggest, to answer No. What makes our desires rational or irrational is not the *rationality* of the beliefs on which these desires causally depend, but the *content* of these beliefs, or *what* we believe. Given my belief that smoking will protect my health, my desire to smoke is rational. I am wanting what, if my belief were true, I would have strong reasons to want. Suppose instead that I wanted to smoke because I had the rational belief that smoking would damage my health. On the

view that we are now discussing, since my desire to smoke would here depend on a rational belief, this desire would be rational. That is clearly false. It would be irrational for me to want to smoke because I believed that smoking would damage my health.

Suppose next that some hermit wants to live a life of complete solitude and self-inflicted pain, because he has the irrational belief that he would thereby please God. Given this man's belief, his desire is rational. And if this hermit wanted to live such a life because he had the rational belief that he would *not* thereby please God, his desire would not be rational.

Similar claims apply to our acts. In most cases, we act rationally when our acts depend on beliefs whose truth would give us sufficient reasons to act in these ways. Given my irrational belief that smoking will protect my health, it would be rational for me to smoke. Given this hermit's irrational belief that his life of self-inflicted pain would please God, he could rationally live such a life. Our claim should be only that, since these irrational beliefs are false, I and the hermit have no reasons to act in these ways.

Some people might object that, when they call some desire or act 'irrational', they *mean* that this desire or act causally depends on some irrational belief. If that is what these people mean, I cannot reject their claim that our desires or acts are irrational when they depend on irrational beliefs. But we ought, I believe, to use 'irrational' in its ordinary sense, to express strong criticism of the kind that we also express with words like 'foolish', 'stupid', and 'senseless'. And we ought, I suggest, to make different claims about which desires or acts deserve such criticism.

Of those who claim that the rationality of our desires depends on the rationality of our beliefs, many assume that we have no reasons to have our desires. Our desires can be rational or irrational, these people assume, only in the derivative sense that these desires causally depend on rational or irrational beliefs. But we do have reasons to have some of our desires. As Objectivists claim, we have reasons to want some events as ends; and, as Subjectivists also claim, we often have reasons to want what would be a means of achieving one of our ends or aims. Since we can have reasons to have our desires, the rationality of our desires should be claimed to depend on whether, in having these desires, we

are responding well to *these* reasons or apparent reasons. We should still claim that, when I want to smoke, *I* am being irrational, but the irrationality is in my belief, not my desire.

We have other reasons to reject the view that our desires or acts are irrational just when they causally depend on irrational beliefs. Such a view would be too narrow even when applied to beliefs. Suppose that, because I believe both that

(1) smoking protects my health

and that

(2) I am now smoking,

I believe that

(3) I am now protecting my health.

My belief in (3) may be in one way irrational, since this belief depends in part on my irrational belief in (1). In another way, however, my belief in (3) is rational. This belief is *rationally derived* from my beliefs in (1) and (2) in the sense that, if these other beliefs were true, that would give me a decisive reason to believe (3). Given my beliefs that I am now smoking and that smoking protects my health, it would be in one way irrational for me, if I asked myself this question, *not* to believe that I am now protecting my health. We might therefore claim that

(E) whether some belief is rational depends in part on whether this belief is rationally derived from some of our other beliefs, and in part on whether these other beliefs are rational.

The rationality of some of our beliefs depends in part on other things, such as their relations to our perceptual experiences. But when applied to many of our beliefs, (E) is roughly right.

We might make similar claims about our desires and acts. We often have some desire, or act in some way, because we have beliefs whose truth would give us sufficient reasons to have this desire, or to act in this way. Such desires or acts we can call *rationally supported* by these beliefs. And we might suggest that

> (F) whether some desire or act is rational depends in part on whether this desire or act is rationally supported by some of our beliefs, and in part on whether these beliefs are rational.

To vary my example, suppose that I want to go to some crowded and noisy party because I believe that I shall enjoy it. This belief is irrational because I ought to have learnt by now that I never enjoy such parties. On the view expressed by (F), given the irrationality of my belief, my desire to go to this party is in one way irrational. In another way, however, my desire is rational, since it is rationally supported by my beliefs. It is rational to want what I believe that I shall enjoy. And if I wanted to go to this party because I had the rational belief that I would *not* enjoy it, my desire would be in one way irrational.

Suppose next that *Green* does something because she has the irrational belief that this act will be certain to achieve her aims. *Grey* does something because she has the irrational belief that this act will be certain to frustrate her aims. According to (F), there is one way in which Green and Grey are both acting irrationally, since these people's acts both depend on irrational beliefs. But there is another way in which Green's act is rational and Grey's is not, since it is rational to do what we believe will achieve our aims, and irrational to do what we believe will frustrate our aims.

Though (F) is plausible, this view is not, I believe, the best. According to (F), our desires and acts are in one way irrational when and because we are failing to respond to some epistemic reason or apparent reason. My act would be in this way irrational when I smoke because I have the irrational belief that smoking will protect my health. But it would be misleading to call my act *practically* irrational, since my mistake is only my failure to respond to my *epistemic* reasons not to have this belief. It would also be misleading to call this act *epistemically* irrational, since it is not in *acting* in this way that I am failing to respond to these epistemic reasons.

We should not, I suggest, make either of these misleading claims. When some belief is epistemically irrational, this irrationality can be plausibly and usefully claimed to be *inherited* by any other belief that depends on this belief. But it is not worth claiming that some belief's irrationality is also inherited by any desire or act that depends on this

belief. Given the differences between epistemic and practical reasons, we should turn to another, simpler view. We should claim that only beliefs can be epistemically irrational. Using a different metaphor, we might say that, when some belief is epistemically irrational, this irrationality can, like a virus, *infect* some of our other beliefs. But with a few exceptions to which I shall soon turn, this irrationality cannot be transmitted over the gap between our beliefs and our desires or acts. Our desires and acts are best called irrational only when, in having some desire or acting in some way, we are failing to respond to clear and strongly decisive *practical* reasons or apparent reasons not to have this desire, or not to act in this way.

On this simpler view, the rationality of our beliefs depends on whether, in having these beliefs, we are responding well to epistemic or truth-related reasons or apparent reasons to have these beliefs. The rationality of our desires and acts depends on whether, in having these desires and acting in these ways, we are responding well to practical reasons or apparent reasons to have these desires and to act in these ways. We might respond well to either set of reasons or apparent reasons, while responding badly to the other set. We might be practically rational but epistemically irrational, or practically irrational but epistemically rational.

We can next consider briefly another widely held view. On this view, what is distinctive of epistemic rationality is the aim of reaching true beliefs. We are epistemically rational, and are responding to epistemic reasons, when we act in the ways that we believe will best achieve this epistemic aim. Though this view cannot be claimed to be false, it is not, I believe, the best view. As well as distinguishing more clearly between epistemic and practical rationality, it would be better to draw this distinction in a different way, and in a different place. The deep distinction here isn't between

> the aim of reaching true beliefs and other possible aims.

When we act in the ways that we believe would best achieve some rational aim, we are being practically rational, and we are responding to practical reasons or apparent reasons, whatever this aim may be. The deep distinction is between

> the voluntary acts with which we respond to practical reasons, and our non-voluntary responses to epistemic reasons.

Trying to reach the truth is an activity, in which we engage for practical reasons. When we are doing mathematics, for example, we may have practical reasons to check some proof, or to redo some calculation in a different way, to confirm the results of some earlier calculation. While we are responding to these practical reasons, by acting in these ways, we shall also respond in non-voluntary and more immediate ways to many epistemic reasons. While we are checking some proof, for example, we respond to epistemic reasons whenever we come to believe, that, since something is true, something else must be true. Coming to have such a particular belief is *not* a voluntary act. As I suggest in Appendix A, practical and epistemic reasons support answers to different questions, and cannot possibly conflict.

17 Beliefs about Reasons

We can have rational beliefs and desires, and act rationally, without having any beliefs about reasons. Young children respond rationally to certain reasons or apparent reasons, though they do not yet have the concept of a reason. Dogs, cats, and some other animals respond to some kinds of reason—such as reasons to believe that we are about to feed them—though they will never have the concept of a reason. And some rational adults seem to lack this concept, or to forget that they have it. Hume, for example, seems to forget this concept when he declares that no desires or preferences could be unreasonable.

If we have beliefs about which are the facts that give us reasons, our desires and acts are often rational responses to what we believe. But that is not always true. Most of us have wanted some things that we believed we had no reasons to want and strong reasons not to want. That is true of many exhausted parents who want to hit their howling babies, and it is true of me whenever I want to smash some malfunctioning machine. When we have some desire that we believe we have no reason to have, and some reasons not to have, our having this desire is not fully rational. Such desires, we can say, are *inconsistent* with, or fail to *match,* our normative beliefs.

I have claimed that, in *most* cases, our desires are rational if these desires depend upon beliefs whose truth would give us sufficient reasons to have these desires. I have also claimed that, in such cases, it is irrelevant whether our beliefs are true, or rational. These claims do not apply when our desires partly depend on certain *normative* beliefs. It may be relevant whether *these* beliefs are true, or rational. Suppose that we falsely and irrationally believe both that some fact gives us a reason to have some desire, and that this desire is rational. If these beliefs were true, we would have a reason to have this desire, and this desire would be rational. That does not make it true that we actually have such a reason, nor does it make this desire rational. Similar claims apply to our acts. If we falsely and irrationally believe that we have a reason to act in some way, or that some act would be rational, that does not give us such a reason, nor does it make this act rational. Practical rationality is not so easily achieved.

It might be objected that, when we have irrational beliefs about which are the facts that give us reasons, that does not make us *practically* irrational. Since these are *beliefs*, we are being *epistemically* irrational, since we are failing to respond to our epistemic reasons not to have these beliefs. And practical and epistemic rationality are, as I have claimed, quite different.

As before, however, that claim applies only to most cases. When our beliefs are about practical reasons, these kinds of rationality and reason overlap. As Scanlon notes, many of our desires can be more fully described as states of being motivated by the belief that something would be good, or worth achieving, in the reason-implying sense. Given this very close relation between these desires and beliefs, the rationality of these desires *does* in part depend on the rationality of these beliefs. And if we have irrational beliefs about practical reasons, and about what we ought rationally to want or to do, our having such beliefs makes us in one way practically irrational.

There is a similar overlap between practical reasons and certain epistemic reasons. We have a practical reason, for example, to want to avoid being in agony, and an epistemic reason to believe that we have this practical reason. The nature of agony both gives us this practical

reason, and gives us this epistemic reason by making it obviously true that we have this practical reason.

Our desires and acts can be rational, I have said, without our having any beliefs about which are the facts that give us reasons. It is enough if we are responding rationally to our awareness of the reason-giving facts, or we are acting on beliefs about non-normative facts whose truth would give us reasons. But when we have beliefs about which facts give us reasons, we are fully practically rational only if these beliefs are rational, and only if we also want, intend, and try to do whatever we believe that we have decisive reasons to want, intend, and try to do.

According to some writers, to be fully rational, we don't need to respond well to reasons, or apparent reasons. It is enough to meet certain rational requirements, such as the requirement to want or intend whatever we believe that we have decisive reasons to want or intend. Such views are, I believe, too narrow.

To illustrate this disagreement, suppose that

> *Scarlet* prefers one hour of agony tomorrow to one minute of slight pain on any other day of the next week,
>
> *Crimson* prefers one hour of agony tomorrow to one minute of slight pain later today,

and

> *Pink* prefers six minutes of slight pain tomorrow to five minutes of slight pain later today.

These people all have true beliefs about what it is like to be in agony and in slight pain, and about personal identity, time, and all the other relevant non-normative facts. But these people differ in some of their beliefs about reasons.

Scarlet we have met before. On Scarlet's view, we have reasons to care about what will happen to us, except on any future Tuesday. Since tomorrow is a Tuesday, Scarlet believes that he has decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain on any other day of the next week. Scarlet has this preference, so he chooses to have the agony.

Crimson's view is closer to the views that many actual people accept. Crimson believes that, though we have reasons to care about all of our future, we have much stronger reasons to care about our nearer future. Crimson therefore believes that he has decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain later today. Crimson has this preference, so he chooses to have the agony.

On Pink's view, we ought to be equally concerned about all the parts of our future, since mere differences in timing have no rational significance. Pink therefore believes that he has a decisive though weak reason to prefer five minutes of slight pain later today to six minutes of slight pain tomorrow. Despite having this belief, however, Pink prefers and chooses to have the slightly longer pain tomorrow.

When Scanlon discusses someone with Scarlet's preference, he writes that 'such a person would not be irrational, but only substantively mistaken'. We should call someone irrational, Scanlon suggests, only when this person 'fails to respond to what he or she acknowledges to be relevant reasons'.

If Scanlon is using the word 'irrational' in its ordinary sense, his claims are not, I believe, justified. Scarlet avoids one kind of irrationality, since Scarlet's preference matches his beliefs about reasons. But in failing to care about his future agony, Scarlet is failing to respond to a very clear and strong reason. And though his preference matches his normative belief, this belief is very irrational. It is crazy to believe that we have reasons to want to avoid agony except on any future Tuesday. These facts are enough, I believe, to make Scarlet's preference irrational.

Crimson's preference is less irrational, since this preference does not draw an arbitrary line, and it is not implausible to believe that we have reasons to care more about our nearer future. But Crimson's version of this view is much too extreme. It is irrational to believe that we have decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain later today. Since Crimson's preference matches his belief about his reasons, he too avoids one kind of irrationality. But in preferring this agony to this slight pain, Crimson is failing to respond to a clear and strongly decisive reason, and his preference matches his belief only because both are irrational.

Since Pink's preference does *not* match his beliefs about reasons, Pink is in one way less rational than Scarlet and Crimson. But this fact is outweighed, I believe, by two others. In having his preference, Pink is failing to respond to a much weaker reason. While Scarlet and Crimson prefer to have one extra hour of agony, Pink merely prefers to have one extra minute of slight pain. And unlike Scarlet and Crimson, Pink has rational beliefs about reasons. These facts, I believe, make Pink much the least irrational of these three people.

People are most *clearly* irrational, Scanlon claims, when they fail to respond to what they themselves acknowledge to be reasons. This claim is in one way true, since such people are less than fully rational even according to their own beliefs. If these people were accused of not being fully rational, they would plead guilty. But that does not justify the claim that only such people should be called irrational. On Scanlon's view, even if we often fail to respond to very clear and decisive reasons, we could avoid irrationality merely by having no beliefs, or false beliefs, about which facts give us reasons, and about which desires or acts are rational. We ought, I believe, to reject this view. Scarlet's attitude to future Tuesdays is irrational even though he believes it to be rational. And if we have rational beliefs about practical reasons, and we admit our failures to respond to these reasons, we may be less irrational than those who have irrational beliefs and much greater unadmitted faults.

Similar claims apply to beliefs. Our beliefs are irrational, on views like mine, when we are failing to respond to clear and strongly decisive epistemic reasons or apparent reasons not to have these beliefs. On a Scanlonian view, our beliefs are irrational only when we fail respond to what we believe to be relevant reasons. Suppose that, though I know that my chance of winning some lottery is only one in a billion, I regard this fact as giving me no reason to give up my belief that I shall win. And though I know that no one else would survive a bare-handed fight with ten hungry lions, I regard this fact as giving me no reason to give up my belief that I would survive such a fight. On a Scanlonian view, these beliefs would not be irrational, since I would be merely making substantive mistakes about which facts give me reasons. In having these beliefs, however, I would be failing to respond to clear and strongly decisive reasons. That is enough to make these beliefs irrational.

There is another version of the view that our desires and acts are irrational only when they fail to match our normative beliefs. According to some people, since there are no truths about reasons or about what is rational, we are irrational only when we ourselves *believe* that we are irrational. Many people make such claims about morality. According to these people, since there are no moral truths, everyone ought to do whatever they believe they ought to do, and no one acts wrongly except by doing what they believe to be wrong. Moral scepticism here leads to one of the inconsistent, self-undermining forms of relativism.

Most of us rightly reject such views. If I break some trivial promise or tell some trivial lie despite believing that these acts are wrong, my acts may be slightly wrong. But when some SS officer killed many civilians, believing these acts to be his duty, his acts were very wrong. It may be some defence that, unlike me, this man did not believe that his acts were wrong. But his acts were morally much worse than mine. Similar claims apply, I believe, when we are discussing rationality. Of my imagined people, only Pink fails to respond to what he believes to be a reason. But Scarlet and Crimson are irrational, while Pink merely fails to be fully rational.

I have rejected Scanlon's claim that, when people like Scarlet and Crimson prefer an hour of agony to a minute of slight pain, these people's preferences are not irrational. There may, however, be no disagreement here. I am using 'irrational' in its ordinary sense, to mean, roughly, 'deserves strong criticism of the kind that we also express with words like "foolish", "stupid", and "crazy" '. At one point Scanlon suggests that we should use 'irrational' in what he calls a narrower sense, which applies only to people who fail to respond to what they themselves believe to be reasons, or who are inconsistent in certain other ways. If Scanlon is using 'irrational' in this narrower sense, his view may not conflict with mine. When Scarlet prefers an hour of agony to a minute of slight pain, his preference is not, I agree, in *this* sense irrational. And Scanlon might agree that Scarlet is making a very great substantive mistake, and that, compared with Pink's preference for an extra minute of slight pain, Scarlet's preference for an hour of agony deserves much stronger rational criticism. If this is Scanlon's view, however, it would be misleading for him to say that only Pink's preference is irrational, since

that would suggest that *Pink's* preference deserves stronger criticism. We ought, I believe, to use 'irrational' in its ordinary, wider sense. If we believe that one of two preferences deserves much stronger rational criticism, we shouldn't say that only the other preference is irrational.

We can next look briefly at a different version of these imagined cases. Scarlet and Crimson, we can now suppose, are both Subjectivists about Reasons. Though these people have the preferences described above, they do not believe that they have any reason to have these preferences. On their view, we have no reasons to want anything as an end, or for its own sake, and what we have most reason to do is whatever would best fulfil our present fully informed telic desires. Since Scarlet and Crimson are both fully informed, and they both now prefer a future hour of agony to a future minute of slight pain, they both believe that they have most reason to choose to have the agony.

On these assumptions, these people's preferences and acts are still, I believe, irrational. In preferring an hour of agony to a minute of slight pain, Scarlet and Crimson are failing to respond to a clear and strongly decisive reason. But their beliefs may not be irrational. While it is crazy to believe that we have reasons to care about future agony except on any future Tuesday, it is not crazy to believe that all practical reasons are given by desires, and that we have no reasons to want anything for its own sake. And many people accept such subjective theories because they were taught to accept them, and their teachers didn't even mention any objective theory. Though subjective theories are, I believe, false, it may not be irrational for these people to accept such theories.

Unlike Scarlet and Crimson, moreover, many of these actual people have rational desires and preferences. Though these people believe that they have no reason to care about their future well-being, they do care. And they may care equally about the whole of their future, so that they would never postpone some ordeal if they believed that this would merely make this ordeal more painful. Such people respond rationally to the facts that give them reasons to care about their future well-being, and they *do*, in this way, respond to these reasons. Their mistake is only in their failing to believe, at the conscious level, that they have these reasons. Some Subjectivists may even have such beliefs, and act upon them in their non-academic lives, ignoring or rejecting these beliefs

only when they teach or write. (This is like the way in which many economists believe, but only when they teach or write, that interpersonal comparisons of well-being make no sense.)

18 Other Views about Rationality

We can next briefly consider some other views about the rationality of our desires, aims, and acts. When some people call some act 'rational', they mean that this act would be most likely to fulfil our present desires, or more precisely would *maximize our expected utility*. Some other people mean that this act would be likely to be best for us, thereby maximizing our expected utility in an older, temporally neutral sense. We can call these the *present-desire-based* and *egoistic* senses of 'rational'. When people use 'rational' in these senses, they can truly claim that we act rationally when we do what would maximize our expected utility, or what would be likely to be best for us. But these are not substantive claims, which might conflict with other views about what is rational. These claims merely tell us that we act in these ways when we act in these ways. To make substantive claims, we must use 'rational' and 'irrational' in other senses. It is best, I have claimed, to use these words in their ordinary senses, to express certain kinds of praise or criticism.

In their substantive claims about rationality, most writers mainly discuss how we ought rationally to try to fulfil our desires, or achieve our aims, in the many cases in which we don't know all of the relevant facts. Such questions, as I have said, have great practical importance, and have been well discussed by many people. Some of these people make conflicting claims about how it would be rational to act in such cases, and about how we can best respond to risks and to uncertainty. But these disagreements are not deep.

There has been much less discussion of *which* desires or aims are rational. When people discuss this more fundamental question, their disagreements have been deep.

On one common view, our desires are rational when our having them has good effects. But if some whimsical despot credibly threatens to

torture me unless, one hour from now, I want to be tortured, that would not make this desire rational. This despot's threat might make it rational for me to cause myself to have this irrational desire, if I can. My *having* this desire would then be, in one way, rational. But this desire *itself* would still be irrational. This would be a case of rational irrationality.

According to another common view, as I have said, our desires or aims are rational when and because they causally depend in the right way on rational beliefs. We ought, I have argued, to reject this view.

The rationality of our desires, some people claim, partly depends on certain other facts about their origin. Our desires are rational, these writers claim, if they were formed through autonomous deliberation, and irrational if they were formed in certain other ways, such as by indoctrination or hypnosis. We ought, I believe, to reject such views. Our desires may be rational even if we were hypnotized or indoctrinated into having them. If we care little about our future, for example, we might be hypnotized into having such rational concern. Or we might be indoctrinated into loving our enemies, and wanting to do at least one good deed in every day. Such love and such desires are, I believe, fully rational. Suppose next that, after autonomous deliberation, we want to starve ourselves to death, thereby losing what would have been a happy life, or we have some other desire for something that is wholly undesirable. The autonomous origin of these desires would not make either them, or us, rational. On the contrary, we would be *less* irrational if, rather than forming these desires through autonomous deliberation, we were made to have them by some form of outside interference, like hypnosis.

According to some other, similar views, the rationality of our desires depends, not on how we came to have them, but on what would cause us to lose them, or on whether they would survive certain tests. Our desires should be called rational, Brandt suggests, if these desires would survive our being given some course of cognitive or belief-based psychotherapy. On this account, our desires might be rational because we are incurably insane. That is not a helpful claim.

According to another group of views, our desires or preferences are irrational when they are *inconsistent*. Two beliefs are inconsistent if they could not both be true. This definition cannot be applied directly to

desires, since desires cannot be true. But two desires can be inconsistent, many writers claim, in the sense that these desires could not both be fulfilled.

Such inconsistency involves no irrationality. Suppose that, after some shipwreck, I could save either of my two children, but not both. Even when I realize this fact, I could rationally go on wanting to save both my children. If we know that two of our desires cannot both be fulfilled, that might make it irrational for us to *aim* or *intend* to fulfil both desires. But these desires may still be in themselves rational, and it may still be rational for us to have them. When our desires are, in this sense, inconsistent, that might make our having them unfortunate. As I have claimed, however, that does not make such desires irrational.

For inconsistency to be a fault, it must be defined in a different way. Though desires cannot be true or false, many desires depend on beliefs about what is good or bad, and these beliefs might be inconsistent, so that they could not all be true. Our desires might be claimed to be derivatively inconsistent when they depend on such inconsistent normative beliefs.

That would be true, it may seem, if we both wanted something to happen, and wanted it not to happen. In having these desires, we might seem to be inconsistently assuming that it would be both better and worse if this thing happened. But in most cases of this kind, we are assuming that some event would be in one way good and in another way bad. For example, I might want to finish my life's work, so as to avoid the risk of dying with my work unfinished, and also want *not* to finish my life's work, so that, while I am alive, I would still have important things to do. Such desires and normative beliefs involve no inconsistency. For two of our desires to be irrationally inconsistent in this belief-dependent way, these desires must depend on beliefs that the very same thing would be both good and bad in the very same way. It is not clear that it would be possible to have such beliefs and desires; but, if it were, the objection that appeals to inconsistency would here be justified.

When we turn to larger sets of preferences, there is more scope for inconsistency. We might prefer B to A, C to B, and A to C. Such preferences are called *cyclical*. If these were *mere* preferences which

did not depend on normative beliefs, it is not clear that such a set of preferences could be claimed to be irrational. This claim is often defended with the remark that, if we had such cyclical preferences, we could be exploited. We might be induced to pay three sums of money first to have B rather than A, then to have C rather than B, and then to have A rather than C. Our money would be wasted, since we would be back with A, where we started. But this objection appeals, not to any inconsistency in such a set of preferences, but to their bad effects. And if we had such preferences, that might have some good effects. Suppose that, whenever our situation changed in some way that we preferred, that change would give us some pleasure. If we had three such cyclical preferences about three easily changeable situations X, Y, and Z, this would be, in a minor way, good for us. We could go round and round this circle, getting pleasure from every move. This merry-go-round would be, hedonically, a perpetual motion machine.

Things are different when such preferences depend on certain normative beliefs. Suppose that we have these preferences because we believe that X is intrinsically better than Y, which is better than Z, which is better than X. Such beliefs would be inconsistent if, as we can plausibly and I believe truly claim, the relation *intrinsically better than* is *transitive*. On this view, just as I can't be taller than you if you are taller than someone who is taller than me, X can't be better than Y if Y is better than Z which is better than X. If such beliefs are inconsistent, that could be claimed to make such preferences derivatively irrational. Though cases that involve such preferences are theoretically very interesting, they do not, I believe, have much practical importance.

The rationality of our desires does not depend, I have claimed, either on their origin, or on their consistency with our other desires. Of those who propose these criteria, some may be misled by presumed analogies with beliefs. The rationality of most of our beliefs *does* depend either on their origin, or on their consistency with our other beliefs, or both. There are relatively few beliefs whose rationality depends only on their content: or *what* we believe. That is true of beliefs about some necessary truths or falsehoods, such as some mathematical or logical beliefs. Some belief is intrinsically irrational, for example, if what we believe is some obvious contradiction. But most of our beliefs are *empirical* and

contingent, in the sense that they are beliefs about how the observable spatio-temporal universe happens to be. There are some empirical beliefs whose rationality depends only on their content. Two examples may be Descartes' belief 'I exist,' and the more cautious Buddhist belief 'This is the thinking of a thought'. Perhaps these beliefs must be true, in a way that makes them intrinsically rational. But few empirical beliefs are of this kind. Some empirical beliefs—such as the belief of some psychotic person that he is Napoleon or Queen Victoria—might seem to be, simply in virtue of their content, irrational. But the irrationality of even these beliefs is still mostly a matter of their origin, and of whether they conflict with our other beliefs. The rationality of most empirical beliefs cannot depend only on their content, because such beliefs are true only if they match the world. What we can rationally believe about the world depends on our other beliefs, our perceptual experiences, and the other evidence available to us.

No such claims apply to our intrinsic telic desires. The rationality of these desires does not depend on how they arose, or on their consistency with our other desires. When we want something as an end, or for its own sake, the rationality of this desire depends only on our beliefs about this desire's object, or what we want. These desires are rational, as objective value-based theories claim, when they depend on beliefs whose truth would make their objects in some way good, or worth achieving. This is the central, fundamental truth that is either ignored or denied by most of the theories that we have been considering.

In rejecting these analogies between the rationality of our beliefs and our desires, I am not forgetting that many of our desires depend upon normative beliefs. These beliefs are about truths that are not empirical and contingent, but necessary. Undeserved suffering, for example, could not have failed to be in itself bad. For such normative beliefs to be rational, we do not need to have evidence that they match the actual world, since these beliefs would be true in any possible world.

6

Morality

19 Sidgwick's Dualism

Objective theories about reasons can differ in several ways. One difference is in the range of events that these theories claim to be good or bad in the reason-implying senses. One of two outcomes would be worse, some theories claim, only if it would be worse *for* one or more people. That, I shall argue, is not true. Nor is it only outcomes that are worth achieving, since some acts are in themselves good; and some things may be worth doing only for their own sake.

Objective theories also differ in their claims about whose well-being we have reasons to promote. We can next consider three such theories. According to

> *Rational Egoism*: We always have most reason to do whatever would be best for ourselves.

According to

> *Rational Impartialism*: We always have most reason to do whatever would be impartially best.

Some act of ours would be impartially best, in the reason-implying sense, if we do what, from an impartial point of view, everyone would have most reason to want us to do. On one view, what would be impartially best is whatever would be, on balance, best for people, by benefiting people most.

In his great, drab book *The Methods of Ethics*, Sidgwick qualifies and combines these two views. According to what Sidgwick calls

> *the Dualism of Practical Reason*: We always have most reason to do whatever would be impartially best, unless some other act would be best for ourselves. In such cases, we would have sufficient reasons to act in either way. If we knew the relevant facts, either act would be rational.

Of these three views, Sidgwick's, I believe, is the closest to the truth. According to Rational Egoists, we could not have sufficient reasons to do what would be worse for ourselves than some other possible act. That is not true. We might have such reasons, for example, when and because our act would make things go impartially much better. I would have sufficient reasons to injure myself if that were the only way in which some stranger's life could be saved. According to Rational Impartialists, we could not have sufficient reasons to do what would be impartially worse than some other possible act. That is not true. We might have such reasons, for example, when and because our act would be much better for ourselves. I would have sufficient reasons to save my own life rather than the lives of several strangers.

On Sidgwick's view, we have both impartial and self-interested reasons for acting, but these reasons are not *comparable*. That is why, whenever one act would be impartially best but another act would be best for ourselves, we would have sufficient reasons to act in either way. No reason of either kind could be outweighed by any reason of the other kind.

Some reasons are *precisely* comparable in the sense that there are precise truths about their relative weight or strength. According to some desire-based subjective theories, all reasons are precisely comparable, since there are precise truths about the relative strengths of all of our desires. According to value-based objective theories, when we must choose between two things that are very similar, such as two cherries or two copies of some book, we might have precisely equal reasons to choose—or, as we could better say, *pick*—either of these things. And when we are comparing reasons of certain kinds—such as reasons

that are provided by differences in the costs of what we might buy, or differences in the length of otherwise similar pleasures and pains—the strengths of these reasons may be precisely comparable. But when we compare most reasons, either of the same or different kinds, these reasons are much less comparable.

Two such dissimilar reasons might be provided by the greater length of one of two possible pains and the greater intensity of the other. If we must choose between one brief but intense pain and another pain that would be much longer but much less intense, one of these possible experiences might be worse, in the sense that we would have more reason to prefer the other. But there could not, I suggest, be any precise truth about the relative strength of these reasons. One of these pains could not, for example, be 2.36 times worse than the other. Even in principle, there is no scale on which we could precisely compare the strengths of our reasons to avoid two such different pains. These claims might be challenged, because the length and intensity of pains both contribute to badness of the same hedonic kind. But there are other, clearer cases. There are only very imprecise truths about the relative strength of many other different kinds of reason, such as economic and aesthetic reasons, or our reasons to keep our promises and to help strangers. Such reasons *are* comparable, however, since some weak reasons of either kind could be weaker than, or be outweighed by, some strong reasons of the other kind.

According to Sidgwick's Dualism, in contrast, impartial and self-interested reasons are *wholly* incomparable. *No* impartial reason could be either stronger or weaker than *any* self-interested reason. Views of this kind are hard to defend. Suppose that we are choosing between some architectural plans for some new building. When neither of two conflicting reasons outweighs the other, we could rationally act in either way. If economic and aesthetic reasons were wholly incomparable, it would therefore be true both that

> (1) we could rationally choose one of two plans because it would make this building cost one dollar less, even though this building would be very much uglier,

and that

> (2) we could also rationally choose one of two other plans because it would make this building slightly less ugly, even though this building would cost a billion dollars more.

We can perhaps imagine how one of these choices might be rational, since we might have reasons to give absolute priority either to this building's beauty, or to its cost. But it would be most implausible to claim that we could rationally make *both* these choices. As this example suggests, to defend Sidgwick's view that impartial and self-interested reasons are wholly incomparable, it is not enough to claim that these reasons are of different kinds.

Sidgwick's defence of his view appeals in part to the rational significance of personal identity. Given the unity of each person's life, we each have strong reasons, Sidgwick claims, to care about our own well-being, in our life as a whole. And given the depth of the distinction between different people, it is rationally significant that one person's loss of happiness cannot be compensated by gains to the happiness of others. Sidgwick here appeals to the *separateness of persons*, which has been claimed to be 'the fundamental fact for ethics'.

Sidgwick's Dualism also rests on what Nagel calls our *duality of standpoints*. We live our lives from our own personal point of view. But we can also think about the world, and all the people in it, as if we had the impartial point of view of some detached observer. When we ask what we have most reason to do, we reach different answers, Sidgwick claims, from these two points of view. From our own point of view, self-interested reasons are *supreme*, in the sense that we always have most reason to do whatever would be best for ourselves. From an impartial point of view, impartial reasons are supreme, since we always have most reason to do whatever would be impartially best.

Suppose next that one possible act would be impartially best, but that some other act would be best for ourselves. Impartial and self-interested reasons would here conflict. In such cases, we could ask what we had most reason to do all things considered. But this question, Sidgwick claims, would never have a helpful answer. We could never have more reason to act in either of these ways. 'Practical Reason' would be 'divided

against itself', and would have nothing to say, giving us no guidance. This conclusion seemed to Sidgwick deeply unsatisfactory.

Sidgwick's reasoning seems to be this:

> (A) When we try to decide what we have most reason to do, we can rationally ask this question either from our own personal point of view or from an imagined impartial point of view.
>
> (B) When we ask this question from our personal point of view, the answer is that self-interested reasons are supreme.
>
> (C) When we ask this question from an impartial point of view, the answer is that impartial reasons are supreme.
>
> (D) To compare the strength of these two kinds of reason, we would need to have some third, neutral point of view.
>
> (E) There is no such point of view.

Therefore

> Impartial and self-interested reasons are wholly incomparable. When such reasons conflict, no reason of either kind could be stronger than any reason of the other kind.

Therefore

> In all such cases, we would have sufficient reasons to do either what would be impartially best, or what would be best for ourselves. If we knew the facts, either act would be rational.

We can call this the *Two Viewpoints Argument*.

Sidgwick's view is, I believe, partly true. But we ought to reject this argument, and revise this view.

We should reject premise (A). It can be worth asking what we would have most reason to want, or prefer, if we were in the impartial position of some outside observer. By appealing to what everyone would have

such impartial reasons to want or prefer, we can more easily explain one important sense in which outcomes can be better or worse. But when we are trying to decide what we have most reason to do, we ought to ask this question from our actual point of view. We should not ignore some of our actual reasons merely because we would not have these reasons if we had some other, merely imagined point of view.

We should also reject (D). To be able to compare partial and impartial reasons, we don't need to have some third, neutral point of view. We can compare these two kinds of reason from our actual, personal point of view.

When we compare these reasons, we can next reject premise (B). On Sidgwick's view, we could rationally do what we knew would be only very slightly better for ourselves, and would be impartially very much worse. For example, we could rationally save ourselves from one minute of discomfort rather than saving a million people from death or agony. If we acted in such a way, the main reactions of others would be horror and indignation. But our question here is: Would this act be rational?

Some people would answer Yes. According to these people, if we knew that this act would best fulfil our present desires, or would be best for us, this act, however horrendous, *would* be rational. Of those who hold such views, however, many use 'rational' in either the present-desire-based sense or the egoistic sense. If these people claimed that this act would be rational, some of them would mean that, in doing what would best fulfil our present desires, we would be doing what would best fulfil these desires. Others would mean that, in doing what we would be best for ourselves, we would be doing what would be best for ourselves. We can ignore such trivial claims. When I ask whether this act would be rational, I am not using 'rational' in either of these senses. I am asking whether this act would deserve one kind of criticism. We act rationally, I believe, only when we have beliefs about the relevant facts whose truth would give us sufficient reasons to act as we do.

In my imagined case, we know the relevant facts. Would we have sufficient reasons to save ourselves from mild discomfort, rather than saving a million people from death or agony? The answer, I believe, is No. This horrendous act would not be rational.

Such acts would not be rational, we might add, because they would be morally wrong. Sidgwick assumes that our self-interested reasons cannot be weaker than, or be outweighed by, our reasons to avoid acting wrongly. We should reject this assumption.

We might also reject Sidgwick's claim that we could always rationally do whatever we knew would make things go best. As an *Act Consequentialist*, Sidgwick believes that such acts would always be morally right. Most of us reject this view, since we believe that certain acts would be wrong even if they would make things go best. The wrongness of such acts, we might claim, would often give us decisive reasons not to act in these ways.

I shall soon turn to questions about morality, and about our reasons to avoid acting wrongly. But we can first revise Sidgwick's view in other ways. This view overstates the rational importance of personal identity. Sidgwick rightly claims that we have reasons to be specially concerned about our own future well-being. But we have other, similar reasons. Of our reasons to care about our future, many are provided, not by the fact that this future will be *ours*, but by various psychological relations between ourselves as we are now and our future selves. Most of us have partly similar relations to some other people, such as our close relatives, and those we love. These are the people, I shall say, to whom we have *close ties.* Our relations to these people can give us reasons to be specially concerned about their well-being. We can have reasons to benefit these people that are much stronger than some of our reasons to benefit ourselves. So we should reject Sidgwick's claim that, when assessed from our personal point of view, self-interested reasons are supreme.

As well as having these *personal* and *partial* reasons to care about the well-being of ourselves and those to whom we have close ties, we also have *impartial* reasons to care about everyone's well-being. Some of Sidgwick's claims imply that we have such reasons only when we consider things from an impartial point of view. But that is not so. Imagining himself as an egoist, Nagel writes:

> Suppose I have been rescued from a fire and find myself in a hospital burn ward. I want something for the pain, and so

> does the person in the next bed. He professes to hope that we will both be given morphine, but I fail to understand this. I understand why he has reason to want morphine for himself, but what reason does he have to want *me* to get some? Does my groaning bother him?

This egoistic attitude would be, as Nagel remarks, 'very peculiar'. Unless we have been taught to accept some desire-based subjective theory, or we lack the concept of a truly normative reason, most of us rightly believe that we have some reason to want any stranger's pain to be relieved. And we have such impartial reasons even when our actual point of view is not impartial. As I have said, we can have reasons to benefit strangers that conflict with, and are much stronger than, some of our self-interested reasons. Rather than saving ourselves from some minor harm, we would have much stronger reasons to save many strangers from death or agony.

Sidgwick's view, however, is partly right. Our partial and impartial reasons are, I believe, only *very imprecisely* comparable. According to what we can call

> *wide value-based objective views*: When one of our two possible acts would make things go in some way that would be impartially better, but the other act would make things go better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either of these ways.

The word 'often' allows for various exceptions. Different wide value-based objective views make conflicting further claims about when it would *not* be true that we had sufficient reasons to act in either of these ways. We ought, I believe, to accept some view of this kind.

To illustrate one such view, we can suppose that, in

> *Case One*, I could either save myself from some injury, or act in a way that would save some stranger's life in a distant land,

and that, in

> *Case Two*, I could save either my own life or the lives of several distant strangers.

In both cases, on most people's views, I would be *morally* permitted to act in either way. If that is so, I would also be *rationally* permitted, I believe, to act in either way. In *Case One* I would have sufficient reasons either to save myself from some injury or to save this stranger's life. And I might perhaps have such reasons whether my injury would be as little as losing one finger, or as great as losing both legs. In *Case Two*, I would have sufficient reasons to save either my own life or the lives of the several strangers. And I might have such reasons whether the number of these strangers would be two or two thousand. Though my reason to save *two* strangers would be *much* weaker than my reason to save *two thousand* strangers, both these reasons might be neither weaker nor stronger than my reason to save my own life. If these claims are true, the relative strength of these two kinds of reason is very imprecise.

There is such great imprecision, we can claim, because these reasons are provided by very different kinds of fact. Our impartial reasons are *person-neutral*, in the sense that these reasons are provided by facts whose description need not refer to us. One example is the fact that some event would cause great suffering. We all have reasons to regret anyone's suffering, and to prevent or relieve this person's suffering if we can, whoever this person may be, and whatever this person's relation to us. We have such reasons to prevent or to regret the suffering of any *sentient* or conscious being. When we are in pain, as Nagel writes,

> the pain can be detached in thought from the fact that it is mine without losing any of its dreadfulness . . . suffering is a bad thing, period, and not just for the sufferer . . . *This experience* ought not to go on, *whoever* is having it.

Our personal and partial reasons are, in contrast, *person-relative*. These reasons are provided by facts whose description must refer to us. We each have such reasons to be specially concerned about the well-being both of ourselves and of those other people who are in certain ways related to us. Though I would have reasons to prevent both my own

pain and the pain of any distant stranger, my relation to *myself*, and to *my* pain, is very different from my relation to that stranger, and to that stranger's pain. That is why these reasons are so imprecisely comparable.

According to some wide value-based views, when we are choosing between morally permissible acts, our reasons to give ourselves some benefit are always stronger than, or outweigh, our reasons to give the same benefit to strangers; but this difference is very imprecise. On one such view, we are rationally required to give to our own well-being more weight than we give to any stranger's well-being, but this greater weight could be as little as twice as much or as great as a hundred or a thousand times as much.

These views are, I believe, too egoistic. We could often rationally give equal or even greater weight to some stranger's well-being. Suppose that, like Nagel, I am in pain in some hospital ward, and the only dose of morphine belongs to me. I would have sufficient reasons, I believe, to give this morphine to the stranger in the next bed. And I would have such reasons even if this stranger's pain was less bad than mine.

Such acts are rational, it might be claimed, only when we are denying ourselves some fairly small benefit. Suppose instead that, in

> *First Shipwreck*, I could use some life-raft to save either my own life or the life of a single stranger. This stranger is relevantly like me, so our deaths would be, for each of us, as great a loss.

When the stakes are as high as this, we may seem to be rationally required to give significant priority, or much greater weight, to our own well-being. If that is true, I would not have sufficient reasons to save this stranger rather than myself. This act, even if morally admirable, would not be fully rational.

I am inclined to believe that this act *might* be fully rational. This stranger's well-being matters just as much as mine. And if I gave up my life to save this stranger, this act would be generous and fine. These facts might, I believe, give me sufficient reasons to act in this way.

There is, I must admit, a strong objection to this view. I believe that, as Sidgwick claims, we have reasons to be specially concerned about our own well-being. And in this imagined case, my death would be impartially as bad as the stranger's death. Since I would have *equal* impartial reasons to save either myself or this stranger, my self-interested reasons might be claimed to break this tie, or tip the scale, giving me decisive reasons, all things considered, to save myself.

These reasons may not, however, be decisive. Even when the stakes are very high, we may not be rationally required to give any priority to our own well-being. We might defensibly accept a revised version of Sidgwick's view. According to what we can call

> *Wide Dualism*: When we are choosing between two morally permissible acts, of which one would be better for ourselves and the other would be better for one or more strangers, we could rationally either give greater weight to our own well-being, or give roughly equal weight to everyone's well-being.

Different versions of this view make different further claims. Though such views do not rationally *require* us to give greater weight to our own well-being, they may *permit* us to give *much* greater weight to our own well-being. And they *do* require us *not* to give much greater weight to any stranger's well-being. On some versions of this view, for example, I could rationally save one of my fingers rather than saving some stranger's life, but I could *not* rationally save some *stranger's* finger rather than saving *my* life. In permitting us to give such great priority to our own well-being, but requiring us *not* to give such great priority to the well-being of strangers, Wide Dualism recognizes and endorses our reasons to be specially concerned about our own well-being.

Suppose next that, in

> *Second Shipwreck*, I could save either some stranger's life or the life of someone to whom I have close ties, such as one of my children, or some friend.

As Wide Dualists could claim, I could not rationally choose to save this stranger. I ought morally to give priority to my child. I would have other strong non-moral reasons to act in this way, such as the reasons that are involved in my love for my child or friend. And if I saved this stranger rather than my child or friend, this act would *not* be generous and fine.

Similar claims might apply to *First Shipwreck*. I might have young children who depend on me, or have other obligations to certain other people. That might make it wrong for me to save some stranger rather than myself, since I could not then care for my children, or fulfil these other obligations. This stranger might have similar obligations that his death would cause to be unfulfilled, but those obligations would not be mine. And if my death would be bad for those who love me and are loved by me, that would give me other decisive reasons to save my life. So in this version of *First Shipwreck*, I would be rationally required to save myself.

Suppose next that I have no such reason-giving and obligation-involving ties to certain other people. I am inclined to believe that, in this other version of this case, I could rationally choose to give up my life to save this stranger. In such cases, we may be rationally permitted to ignore our reasons to be specially concerned about our own well-being. But we need not here decide whether that is true, or whether my act, though morally admirable, would be less than fully rational.

20 The Profoundest Problem

We can now turn to the relations between reasons and morality. According to

> *Moral Rationalism*: We always have most reason to do our duty. It could not be rational to act in any way that we believe to be wrong.

According to

> *Rational Egoism*: We always have most reason to do what would be best for ourselves. It could not be rational to act in any way that we believe to be against our own interests.

Many people accept both these views. Most of these people believe that duty and self-interest never conflict, since each of us will have some future life in which, if we have done or failed to do our duty, we shall get the happiness or suffering that we deserve. That is claimed by most of the world's great religions.

Sidgwick doubted that we shall have some future life, and he thought it to be likely that, in some cases, duty and self-interest conflict. If there are such cases, Sidgwick claims, that would raise 'the profoundest problem in ethics'.

Sidgwick's problem was in part that Moral Rationalism and Rational Egoism both seemed to him intuitively very plausible, but that, if duty and self-interest sometimes conflict, these views cannot both be true. If we had to choose between two acts, of which one was our duty but the other would be better for ourselves, these views imply that we would have most reason to act in each of these ways. That is inconceivable, or logically impossible. Just as we could not keep most of our money in each of two different wallets, we could not have most reason to act in each of two different ways. So if duty and self-interest sometimes conflict, we would have to reject or revise at least one of these views.

When they consider these alternatives, some writers reject Moral Rationalism. Reid, for example, claims that, if it would be against our interests to do our duty, we would be 'reduced to this miserable dilemma, whether it be best to be a knave or a fool'. We would be knaves if we didn't do our duty, but fools if we did. Other writers reject Rational Egoism. According to these people, we could never have sufficient reasons to act wrongly, not even if that was our only way to save ourselves from great pain or death.

Sidgwick found such claims incredible. Rather than rejecting one of these views, he revised them both. According to another version of Sidgwick's Dualism, which we can call

> *the Dualism of Duty and Self-Interest*: If duty and self-interest never conflict, we would always have most reason both to do our duty and to do what would be best for ourselves. But if we had to choose between two acts, of which one was our duty but the other would be better for ourselves, reason would give

> us no guidance. In such cases, we would not have stronger reasons to act in either of these ways. If we knew the relevant facts, either act would be rational.

Partly because he accepted this view, Sidgwick passionately hoped that duty and self-interest never conflict. If there are such conflicts, he writes,

> the whole system of our beliefs as to the intrinsic reasonableness of conduct must fall . . . the Cosmos of Duty is thus really reduced to a Chaos, and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure.

These magnificently sombre claims are, however, overstatements. Sidgwick believed that in most cases duty and self-interest do not conflict. Sidgwick's view implies that, in these many cases, we would have most reason to do our duty, at no cost to ourselves. In such a world, the cosmos of duty would not be a chaos. Nor would our whole system of beliefs about what is reasonable conduct fall if we concluded that, when duty and self-interest conflict, we could reasonably, or rationally, act in either way. But it would be bad if, in such cases, we and others would have sufficient reasons to act wrongly. The *moralist's problem*, we might say, is whether we can avoid that conclusion. And it would be disappointing if, in such cases, reason gave us no guidance. We may hope that, in at least some of these cases, there would be something that we had most reason to do. The *rationalist's problem*, we might say, is whether that is true.

These problems might take other forms. Sidgwick assumes that, if we had sufficient reasons to act wrongly, these reasons would be self-interested. We should not make that assumption, since we can have other strong reasons to act wrongly. Some of these reasons are personal and partial, but not self-interested. We might have sufficient reasons to act wrongly, for example, if some wrong act was our only way to save from great pain or death, not ourselves, but our close relatives, or other people whom we love.

We might also have strong impartial reasons to act wrongly. As an Act Consequentialist, Sidgwick claims that we ought always to do whatever

would make things go best. Most of us reject this view, since we believe that some acts would be wrong even if they would make things go best. It might be wrong to kill someone, for example, even when that is the only way in which many other people's lives could be saved. Even if this act would be wrong, however, the fact that we would be saving many people's lives, thereby making things go best, might be claimed to give us sufficient reasons to act in this way. If that were true, this would be another kind of case in which we could rationally act wrongly.

There is a third possibility. On Sidgwick's view, we always have sufficient reasons to do our duty, and to avoid acting wrongly. We can call this view *Weak Moral Rationalism*. If we are Subjectivists about Reasons, we must reject this view. Rawls for example claims that, if our present informed desires would be best fulfilled by acting unjustly, we would not have sufficient reasons to do what justice requires. According to such subjective theories, we might have no reason to do our duty, and decisive reasons to act wrongly. It might then be *irrational* for us to do our duty.

To cover these various possibilities, we can revise Sidgwick's description of what he calls 'the profoundest problem'. When we are choosing between different possible acts, we can ask:

> Q1: What do I have most reason to do? Do I have sufficient or decisive reasons to act in any of these ways?
>
> Q2: What ought I morally to do? Would any of these acts be wrong?

These questions might, it seems, have conflicting answers, since we might sometimes have sufficient or decisive reasons to act wrongly. Our problem is to decide whether we do or could have such reasons, and, if that is true, what further conclusions we should draw.

In considering these questions, it will help to distinguish between two conceptions of normativity. On the *reason-involving* conception, normativity involves reasons or apparent reasons. On the *rule-involving* conception, normativity involves requirements, or rules, that distinguish between what is *correct* and *incorrect*, or what is *allowed* and *disallowed*. Certain acts are required, for example, by the law, or by the code of

honour, or by etiquette, or by certain linguistic rules. It is illegal not to pay our taxes, dishonourable not to pay our gambling debts, and incorrect to eat peas with a spoon, to spell 'committee' with only one 't', and to use 'refute' to mean 'deny'. Such requirements or rules are sometimes called 'norms'.

These conceptions of normativity are very different. On the rule-involving conception, we can create new normative truths merely by introducing, or getting some people to accept, some rule. Legislators can create laws, and anyone can create the rules that define some new game. When Shakespeare wrote, there were regularities but no rules about the spellings of English words. Later writers of English have created such rules. In contrast, on the reason-involving conception, there is normativity only when there are normative reasons or apparent reasons. We cannot create such reasons merely by getting people to accept some rule.

These conceptions may conflict. When there are such rules or requirements, we may have reasons to follow them. But these reasons are mostly provided, not by the mere existence or acceptance of these rules, but by certain other facts, most of which depend on some people's acceptance of these rules. If we drive on the correct side of the road, we shall be less likely to crash. If we use words with their correct spelling and meaning, that may make us seem better educated, and help us to be understood. When there are no such reason-giving facts, we may have no reason to follow some rule or requirement. We may have no reason, for example, to follow some fashion, or to refrain from violating some taboo. When I was told, as a child, that I shouldn't act in certain ways, and I asked why, it was infuriating to be told that such things are *not done*. That gave me no reason not to do these things.

Many of these claims do not apply to *moral* requirements. On some views, it is we who create these requirements. That is true, I believe, only in limited and often superficial ways. What we can create are only the particular forms that, in different communities, more fundamental, universal, and uncreated requirements take. For example, it is true everywhere that some people ought to care for those other people who cannot care for themselves, such as young children and those who are disabled by disease or old age. In most communities it is mostly close relatives who have such responsibilities. But that is not true everywhere.

There are also various uncreated rational requirements. For example, if we believe that we have decisive reasons to act in some way, we might be rationally required either to act in this way, or to give up this belief. And if we believe that some act is our only way to achieve some aim, we might be rationally required either to act in this way or to give up this aim.

Moral requirements often conflict with requirements of other kinds. We can be legally required, for example, to act wrongly. And many men have believed that, though it would be morally wrong to fight some duel, it would be dishonourable not to fight. Most of us would believe that, in these two kinds of case, moral requirements are more important. These requirements are often called *overriding*. But it would be trivial to claim that moral requirements are *morally* more important, or *morally* overriding. Legal requirements are *legally* overriding, and the code of honour is overriding in this code's terms. To be able to make significant claims about the relative importance of these conflicting requirements, we need some impartial, neutral criterion.

Reasons provide such a criterion. We can compare the strengths of our reasons to follow these requirements. The men who fought duels had at most weak reasons to follow the code of honour, and they had strong moral reasons not to fight. And when we are legally required to act wrongly, we may have decisive moral reasons to break the law. Moral requirements may thus be more important in the reason-implying sense than the requirements of the code of honour, or the law.

It would be similarly trivial to claim that rational requirements are rationally overriding. So we should ask whether we have reasons to follow these requirements. It is a difficult question how much these requirements matter in the reason-implying sense. Following these requirements might be good, not in itself, but only as a means. And in appealing to claims about what matters in the reason-implying sense, we are not assuming that rationality matters.

We can next note one difference between moral and rational requirements. When we are deciding what to do, we often ought to ask whether any of our possible acts would be morally required, or wrong. But we need not ask which acts would be rational. That question arises only when we consider our own past acts, or the acts of others, and we ask

whether these acts make us or others open to certain kinds of criticism. Compared with questions about what we ought to do or have reasons to do, questions about rationality are much less important.

When we are deciding what to do, as I have said, we have two main questions:

Q1: What do I have most reason to do?

Q2: What ought I morally to do?

Of these questions, it is the question about reasons that is wider, and more fundamental. And if these questions often had conflicting answers, because we often had decisive reasons to act wrongly, that would undermine morality. For morality to matter, we must have reasons to care about morality, and to avoid acting wrongly. No such claim applies the other way round. If we had decisive reasons to act wrongly, the wrongness of these acts would not undermine these reasons.

These claims might be denied. When I claim that the wrongness of these acts would not undermine these reasons, I mean that we would still have these reasons. It might be similarly claimed that, even if we had decisive reasons to act wrongly, *morality* would not be undermined, since these acts would still be wrong.

This defence of morality would be weak. It could be similarly claimed that, even if we had no reasons to follow the code of honour, or the rules of etiquette, this code and these rules would not be undermined. It would still be dishonourable not to fight some duels, and still be incorrect to eat peas with a spoon. But these claims, though true, would be trivial. If we had no reasons to do what is required by the code of honour, or by etiquette, these requirements would have no importance. If we had no reasons to care about morality, or to avoid acting wrongly, morality would similarly have no importance. That is how morality might be undermined.

It might next be objected that, in making these claims, I am appealing to the reason-involving criterion of importance. I am assuming that something is important only when and because we or others have reasons to care about this thing. But I have not defended this criterion.

And like morality or the code of honour, the reason-involving criterion cannot support itself. Just as it would be trivial to claim that morality is *morally* important or that rationality is *rationally* important, it would be trivial to claim that reasons are important in the *reason-implying* sense.

As this objection rightly claims, we cannot show that reasons matter by appealing to claims about reasons. But justifications must end somewhere. And if reasons are fundamental, we should not expect that we could justify the reason-involving criterion of importance, by appealing to some other, deeper criterion.

Reasons *are*, I believe, fundamental. Something matters only if we or others have some reason to care about this thing. It would have great importance if morality did not in this sense matter, because we had no reason to care whether our acts were right or wrong. To explain and defend morality's importance, we can claim and try to show that we do have such reasons. Morality might have supreme importance in the reason-implying sense, since we might always have decisive reasons to do our duty, and to avoid acting wrongly. But if we defend morality's importance in this way, we must admit that the deepest question is not what we ought morally to do, but what we have sufficient or decisive reasons to do.

In the rest of this volume I shall mostly discuss morality. If reasons are more fundamental, as I have just claimed, it may seem that I should continue to discuss reasons. But we have sufficient reasons for turning to morality.

First, we can plausibly assume that we do have strong reasons to care about morality, and to avoid acting wrongly. In discussing morality, I shall in part be discussing these reasons. And these are among the reasons that most need discussing, because they raise some of the hardest questions.

Second, before we can judge the strength of our reasons to avoid acting wrongly, we must answer certain questions about which acts are wrong. One example is the question whether, as Act Consequentialists believe, we ought to sacrifice our life if we could thereby save the lives of several strangers. If that were true, we could more plausibly claim that we might have sufficient or even decisive reasons to act wrongly.

According to the overlapping sets of beliefs that most people accept, which Sidgwick calls *common sense morality*, we are morally permitted to give some kinds of strong priority to our own well-being. We might have no duty to sacrifice our life, even if we could thereby save very many strangers. If morality's requirements are in such ways much less demanding, it is less plausible to claim that we can have sufficient or decisive reasons to act wrongly.

There are other ways in which, when considering morality, we shall be considering reasons. On several plausible moral principles or theories, whether some act is wrong depends on what, in certain actual or imagined situations, we or others would have most reason or sufficient reason to consent to, or agree to, or to want, or choose, or do. To know what these principles and theories imply, we must answer questions about reasons. That is like the way in which, to know about the nature and properties of atoms, we must answer questions about sub-atomic particles.

7

Moral Concepts

21 Acting in Ignorance or with False Beliefs

Before we start to ask which acts are wrong, it will help to discuss what we mean by 'wrong', and what we are believing when we believe that some act is wrong. These questions are about *moral* senses of 'wrong', and the concepts that these senses express. We can ignore non-moral senses, such as the sense in which we might give the wrong answer to some question, or open some cereal packet at the wrong end.

It is often assumed that the word 'wrong' has only one moral sense. This assumption is most plausible when we are considering the acts of people who know all of the morally relevant facts. We can start by supposing that, when we think about such acts, we all use 'wrong' in the same sense, which we can call the *ordinary* sense. In many cases, however, we don't know all of the relevant facts, and we must act in ignorance, or with false beliefs. When we think about such cases, we can use 'wrong' in several partly different senses. Some of these senses we can define by using the ordinary sense. Some act of ours would be

> *wrong* in the *fact-relative* sense just when this act would be wrong in the ordinary sense if we knew all of the morally relevant facts,
>
> *wrong* in the *belief-relative* sense just when this act would be wrong in the ordinary sense if our beliefs about these facts were true,

and

> *wrong* in the *evidence-relative* sense just when this act would be wrong in the ordinary sense if we believed what the available evidence gives us decisive reasons to believe, and these beliefs were true.

Acts are in these senses *right*, or at least *morally permitted*, when they are not wrong, and they are what we *ought morally* to do when all of their alternatives would be in these senses wrong.

Some writers claim or assume that, even when we are considering the acts of people who don't know all of the morally relevant facts, it is enough to ask which of these people's acts would be wrong, or were wrong, in the ordinary sense. Other writers assume that one of the senses I have just defined *is* the ordinary sense. These assumptions are, I believe, mistaken. We ought to use 'wrong' in all these senses. If we don't draw these distinctions, or we use only some of these senses, we shall fail to recognize some important truths, and we and others may needlessly disagree.

To illustrate these points, we can suppose that, as your doctor, I must choose between different ways of treating you. I am a bad doctor, since I have various unjustified beliefs about what, given the evidence, are the likely effects of different treatments. I also have some reasons to wish that you were dead. This story could continue in several ways. Suppose that, in

> *Case One*, I give you some treatment that I believe and hope will save your life, but which kills you, as it was almost certain to do,

and that, in

> *Case Two*, I give you some treatment that I believe and hope will kill you, but which saves your life, as it was almost certain to do.

According to some people, it is enough to use 'right' and 'wrong' in their belief-relative senses. On this view, it is enough to claim that I

acted rightly in *Case One*, because I did what I believed would save your life, and that I acted wrongly in *Case Two*, because I did what I believed would kill you.

It is *not* enough to make these claims. We should also claim that, in *Case One*, I acted wrongly in the fact-relative and evidence-relative senses, since I killed you, as on the available evidence my act was almost certain to do. If I had asked some fully informed adviser what I ought to do, this person should not have told me that I ought to do what he or she knew would almost certainly kill you. We should similarly claim that, in *Case Two*, I acted rightly in the fact-relative and evidence-relative senses, since my act saved your life, as it was almost certain to do. I did what any fully informed adviser ought to have told me that I ought to do.

Suppose next that, though certain treatments nearly always cure people who have your particular disease, and certain other treatments would nearly always kill such people, your case is one of the unpredictable exceptions. And suppose that, in

> *Case Three,* I give you some treatment that is almost certain to kill you, but which saves your life, as I hoped and unjustifiably believed that it would,

and that, in

> *Case Four,* I give you some treatment that is almost certain to save your life, but which kills you, as I hoped and unjustifiably believed that it would.

According to some people, it is enough to use 'right' and 'wrong' in their evidence-relative senses. On this view, if some believer in sorcery tried to kill some enemy by sticking pins into a wax dummy, this person would not be acting wrongly. It is not wrong to stick pins into wax dummies, since there is no evidence that such acts do any harm. And I acted rightly, in *Case Four*, when I gave you a treatment that, on the available evidence, was almost certain to save your life. But I acted wrongly in *Case Three* when I gave you a treatment that was almost certain to kill you.

As before, it is not enough to make these claims. We should not say only that I acted rightly, in *Case Four*, since my act was almost certain to save your life. We should also claim that I acted wrongly in the

belief-relative and fact-relative senses, by *murdering* you. Murders should at least be mentioned.

Nor is it enough to say that, in *Case Three*, I acted wrongly by doing what was almost certain to kill you. We should also claim that I acted rightly in the fact-relative and belief-relative senses, since I intentionally saved your life. In failing to believe that my act would almost certainly kill you, I may be guilty of negligence, since I may have failed to read the recent medical journals, as I ought to have done. But it might instead be true that I conscientiously read these journals, and my mistake was only that I failed to believe what the evidence reported in these journals gave me decisive reasons to believe. Though I would then be at fault for medical incompetence, my failure to respond to these epistemic reasons would not be morally wrong.

According to some other people, it is enough to use 'right' and 'wrong' in their fact-relative senses. But suppose that, in

> *Case Five*, I give you some treatment that, as I justifiably believe, is almost certain to save your life, but which in fact kills you.

It is not enough to claim that, since I killed you, I acted wrongly. We should also claim that I acted rightly in the belief-relative and evidence-relative senses. It is morally important that I justifiably believed that my act was almost certain to save your life. Suppose instead that, in

> *Case Six*, I give you some treatment that, as I justifiably believe, will almost certainly kill you, but which in fact saves your life.

It is not enough to claim that, since I saved your life, I acted rightly. We should also claim that I acted wrongly in the belief-relative sense, because I believed that my act would kill you, as I intended it to do. Attempted murders should at least be mentioned.

It would be possible to draw these distinctions without using these different senses of 'right' and 'wrong'. We might use only the evidence-relative senses. We might then claim that, though I did not act wrongly in *Case Four* when I murdered you, I had morally decisive reasons not to act in this way, and my act was blameworthy, giving me reasons for remorse and giving others reasons for indignation. Or we might use

only the belief-relative senses. We might then claim that, though I did not act wrongly in *Case One* when I tried to save your life, I had morally decisive reasons not to act in this way, because my act killed you, as I should have known that it was almost certain to do. Or we might use only the fact-relative senses. We might then claim that, though I did not act wrongly in *Case Six* when I saved your life, my act was blameworthy, because I was trying to kill you. But if we use 'wrong' in only one of these three senses, we may be misunderstood by those who use 'wrong' in only one or both of the other two senses. We and others may mistakenly believe that we are disagreeing. When we consider cases in which people do not know all of the morally relevant facts, there is no one sense of 'wrong' that everyone uses. So it is best to distinguish and use all these three senses.

We can next ask which of these senses are most important. As some of my claims have implied, that depends on which questions we are asking. We can start with questions about blameworthiness, which we can take to include questions about reasons for remorse and indignation. What is most important here is what, when acting, people believe. We should claim that

> (A) when some act is wrong in the *belief-relative* sense, because this act would be wrong if the agent's non-moral beliefs were true, this fact makes this act blameworthy.

In *Cases Two*, *Four*, and *Six*, for example, I act in ways that I believe will kill you. These acts would all be wrong if my beliefs were true, since intentionally killing you would be wrong. So (A) rightly implies that these acts were all blameworthy.

It might be similarly claimed that

> (B) when some act is wrong in the *fact-relative* sense, because this act would be wrong if the agent knew the relevant facts, this fact makes this act blameworthy.

But we ought to reject this claim. Remember that, in

> *Case Five*, I kill you by doing what I justifiably believe will save your life.

Since this act would be wrong if I knew that it would kill you, (B) implies that this act was blameworthy. But that is clearly false. When I learn that I have killed you, I shall be appalled. But since I justifiably believed that my act would save your life, this act was not blameworthy. And I have no reason for remorse, nor do others have any reason for indignation.

Here is a wider objection to (B). Suppose that, in

> *Case Seven*, I save your life by doing what I justifiably believe will save your life.

It is clear that, in this case, my act was *not* blameworthy, since this act wasn't in any sense wrong. Though my act kills you in *Case Five* but saves your life in *Case Seven*, this difference is, from my point of view, entirely a matter of luck. In calling this difference a matter of *luck* from my point of view, I mean that I could not have known that one of these acts would kill you, and that this fact was in no way under my control. Though the difference between these cases is entirely a matter of luck, (B) implies that my act was blameworthy in *Case Five* but not in *Case Seven*. (B) therefore implies that

> (C) an act's blameworthiness might entirely depend on luck.

When children are learning what it is for acts to be blameworthy, some of them have beliefs that assume or imply (C). Some of these children believe, for example, that well-intentioned acts are blameworthy when these acts have bad effects, even if these effects were wholly unpredictable. And some adults have had similar beliefs, such as the belief that we can inherit blameworthiness and guilt for the sins of our ancestors. These sins were not under our control. But when we understand blameworthiness better, we realize that (C) is false. Since (B) implies (C), we ought also to reject (B). When some act is wrong in the fact-relative sense, this fact does not make this act blameworthy.

There are two alternatives to (C). According to what we can call

> *the Kantian view*, an act's blameworthiness cannot depend on luck.

According to

> *the semi-Kantian view*, an act's blameworthiness cannot depend *entirely* on luck. But when two acts are blameworthy in some way that does not depend on luck, one of these acts may be *more* blameworthy in some way that *does* depend on luck.

This view is in itself less plausible than the Kantian view, since it is hard to see how blameworthiness might *partly* depend on luck. But this semi-Kantian view is sometimes claimed to have more plausible implications. Return for example to

> *Case Two*, in which I save your life by doing what I believe will kill you,

and

> *Case Four*, in which I kill you by doing what I believe will kill you.

These acts are both wrong in the belief-relative sense, since if my beliefs were true these acts would both kill you, as I intend them to do. In the *fact*-relative sense, however, my act is wrong only in *Case Four*. Though my act kills you in *Case Four* but saves your life in *Case Two*, this difference is, from my point of view, entirely a matter of luck. So, on the Kantian view, these acts are equally blameworthy. According to some semi-Kantians, that is not so. These people believe that

> (D) when acts are blameworthy because they are wrong in the belief-relative sense, these acts are more blameworthy if they are also wrong in the fact-relative sense.

On this view, though my attempts to kill you are both blameworthy, my act is more blameworthy in *Case Four*, because this attempt succeeds. Though attempted murder is blameworthy, murder deserves more blame, and gives me and others reasons for greater remorse and greater indignation.

Some semi-Kantians might also claim that

> (E) when acts are blameworthy because they are wrong in the belief-relative sense, these acts are more blameworthy if they are also wrong in the *evidence*-relative sense.

But remember that, in

> *Case Four*, I kill you by giving you a treatment that, on the evidence, was almost certain to save your life, but which I unjustifiably believed would kill you.

Suppose next that, in

> *Case Eight*, I kill you by giving you a treatment that I justifiably believed would kill you.

These acts are both wrong in the belief-relative and fact-relative senses, since they both kill you, as I believed they would. (E) implies that, in *Case Eight*, my act is more blameworthy, because this act is also wrong in the evidence-relative sense. We ought, I believe, to reject this claim. Murder can be plausibly regarded as more blameworthy than attempted murder. But we cannot plausibly regard murder as more blameworthy if and because the murderer's beliefs about the likely effects of his act were epistemically justified, because these beliefs were better supported by the available evidence. The most that we could claim is that, if potential murderers have such justified beliefs, these people are more dangerous, because their attempts to kill other people are more likely to succeed. That is not a difference in blameworthiness.

On the Kantian view, all such attempts to kill are equally blameworthy, whether or not these acts succeed, or were likely to succeed. It is equally blameworthy to shoot someone and hit, to shoot someone and miss, and to stick pins into a wax dummy believing irrationally that this way of killing someone will succeed. We cannot deserve less blame merely because we are either less successful in hitting our intended target, or are epistemically irrational.

This Kantian view is, I believe, true. Though murder can be plausibly regarded as more blameworthy than attempted murder, this claim's

plausibility can be sufficiently explained, I believe, in other ways, some of which I mention in a note.

We can next define a fourth relevant sense of 'wrong'. Some act is

> *wrong* in the *moral-belief-relative* sense just when the agent believes this act to be wrong in the ordinary sense.

On one fairly plausible view, which we can call

> *the Thomist View*, when people believe that they are acting wrongly, that is enough to make their act wrong, even if this act would not otherwise be wrong.

Suppose, for example, that it would not be wrong to use artificial contraceptives, or to perform an early abortion, or to help someone to die in a swifter, better way. On this Thomist view, such acts *would* be wrong if they were done by people who mistakenly believed them to be wrong. As Thomists add, however, when people believe that some act would be *right*, that is *not* enough to make this act right. Conscientious SS officers often acted wrongly, even when they believed their acts to be right, or to be their duty.

Even if we reject this view, it seems clear that

> (F) in most cases, when someone acts in some way that this person believes to be wrong, that makes this act blameworthy.

Of the facts that can make acts blameworthy, this may be the most important. In some cases, however, people do what they believe to be wrong because they are half-aware that their act is not wrong, but morally required. One example may be Huckleberry Finn when he helped a runaway slave to escape. Some such acts may not be blameworthy. But in most cases, an act's blameworthiness depends on whether this act is wrong in the belief-relative and moral-belief-relative senses.

We can next ask which are the most important senses of 'ought', 'right', and 'wrong' when we are trying to decide what to do. In the cases that we have been discussing, and many others, the rightness of our acts

depends on the goodness of their effects or possible effects. It is often assumed that

> (G) in such cases, we ought to try to act in the way that would be right in the fact-relative sense, because this act would make things go best.

In my medical examples, (G) has acceptable implications. In trying to do what would save your life, I would be trying to do what would make things go best. But in many other cases (G) is false. Consider

> *Mine Shafts*: A hundred miners are trapped underground, with flood waters rising. We are rescuers on the surface who are trying to save these men. We know that all of these men are in one of two mine shafts, but we don't know which. There are three flood-gates that we could close by remote control. The results would be these:

	The miners are in	
	Shaft A	Shaft B
Gate 1	We save 100 lives	We save no lives
We close: Gate 2	We save no lives	We save 100 lives
Gate 3	We save 90 lives	We save 90 lives

Suppose next that on the evidence available and as we believe, it is equally likely that the miners are all in Shaft A or all in Shaft B. If we closed either Gate 1 or Gate 2, we would have a one in two chance of doing what would be right in the fact-relative sense, because our act would save all of these hundred people. If we closed Gate 3, we would have *no* chance of doing what would be in this sense right. But this is clearly what we ought to do, since by closing Gate 3 we shall be certain to save ninety of these people.

When I claim that we ought to close Gate 3, I am using 'ought' in the ordinary sense. This act is also what we ought to do in the more precise belief-relative and evidence-relative senses, since the hundred miners *are*, as we justifiably believe, equally likely to be in either shaft. Since it would be wrong for us to try to act rightly in the fact-relative sense

by closing either of the other gates, we ought to reject claim (G). On a rough statement of the true view, which we can call

> *Expectabilism*: When the rightness of some act depends on the goodness of this act's effects or possible effects, we ought to act, or try to act, in the way whose outcome would be *expectably-best*.

In calling some act's outcome 'expectably-best', we do *not* mean that we expect this act to produce the best outcome. In this example, the outcome would be expectably-best if we closed Gate 3, though this act would be certain *not* to produce the best outcome, as our act might do if instead we closed one of the other gates. To decide which of our possible acts would make things go *expectably-best*, we take into account both how good the effects of the different possible acts might be, and the probabilities, given our beliefs or the available evidence, that these acts would have these effects. When what matters is only the number of lives that are saved, some act's outcome would be expectably-best if this is the act that would save the greatest *expectable number* of lives. The expectable number that some act would save is the number of lives that this act might save, multiplied by the chance that this act would save these lives. In *Mine Shafts,* for example, if we closed either Gate 1 or Gate 2, the expectable number of lives saved would be 100 multiplied by a chance of one in two, or by 0.5. This number would be 50. If we closed Gate 3, this expectable number would be 90, since this act would be certain to save 90 lives.

We can similarly claim that, whenever we don't know what effects our acts would have, the expectable goodness of some act's effects is, roughly, the goodness of these possible effects multiplied by the chance that this act would have these effects. Expectabilism applies to all cases, including those in which we know which act would in fact make things go best. This act's outcome would be expectably-best.

I have just rejected the view that, when we don't know what effects our acts would have, we ought to try to do what would in fact make things go best. It is sometimes claimed that, if we reject this view, we cannot explain why we ought, in many cases, to try to discover more of the facts, so that we can make better informed decisions. But this claim is mistaken. We ought to try to get more information whenever acting in

this way would itself make things go expectably-best. In important cases, that is often true. In *Mineshafts*, if we could easily find out where the miners are, trying to find this out would make things go expectably-best, since we would then be very likely to save all these people.

There is another reason why, when we are trying to decide what to do, we can ignore the fact-relative senses of 'ought', 'right', and 'wrong'. We cannot try to do what is right in the fact-relative *rather than* the belief-relative sense. Suppose I believe that, to save your life, I must act in a certain way. Though I know that my belief might be false, I cannot try to do what *would in fact* save your life rather than doing what *I now believe* would save your life, since what *I now believe* is that acting in this way *would in fact* save your life. We cannot base our decisions on the facts except by basing our decisions on what we now believe to be the facts. In the same way, as Sidgwick points out, though we know that our moral beliefs may be mistaken, we cannot try to do what is really right rather than what, at the time of acting, we believe to be right.

I claimed earlier that, when we ask whether some act was blameworthy, or whether the agent has reasons for remorse and others have reasons for indignation, what is most important is whether this act was wrong in the belief-relative and moral-belief-relative senses. I have just claimed that, when we are choosing between different possible acts, we need not ask what we ought to do in the fact-relative sense. And when the rightness of our acts depends on the goodness of their effects, we ought to try to do, not what would in fact make things go best, but what on the evidence, or given our beliefs, would make things go expectably-best. These claims may seem to imply that it has little importance which acts are right or wrong in the fact-relative senses.

There is, however, one way in which these fact-relative senses can be claimed to be fundamental. As well as asking, in some actual case, whether some act would be wrong, we can ask wider questions about which moral beliefs are true, and which moral principles or theories we ought to accept and try to follow. We ought to try to answer some of these questions, or at least to think about some other people's answers. Though we cannot try to do what is really right rather than what we now believe to be right, we ought to try to have true moral beliefs, since we shall then be less likely to act wrongly.

In trying to answer such questions, it is best to proceed in two stages. We can first ask which acts would be wrong if we knew all of the morally relevant facts. These are questions about which acts would be wrong, in such cases, in what I have called the ordinary sense. But these are also questions about which acts would be wrong in the fact-relative sense. Acts are in this sense wrong when these acts would be wrong in the ordinary sense if we knew all of the relevant facts.

After answering these questions, we can turn to questions about what we ought morally to do when we don't know all of the relevant facts. These questions are quite different, since they are about how we ought to respond to risks, and to uncertainty. As in the case of non-moral decisions, though these questions have great practical importance, they are less fundamental. These are not the questions about which different people, and different moral theories, most deeply disagree. Given the difference between these two sets of questions, they are best discussed separately. So I shall often suppose that, in my imagined cases, everyone would know all of the relevant facts. We can then ask what we ought to do in the simplest, fact-relative sense. In many other cases these distinctions do not matter, so I shall often use 'best' to mean 'best or expectably-best'.

There is much more to be said about the relations between these and some other similar senses of 'ought' and 'wrong'. There are difficult questions, for example, about when and how people who have different beliefs, or are aware of different evidence, can disagree about what someone ought to do. My aim has been only to argue that we need to distinguish these senses, and to decide which senses are most relevant to the kind of moral question we are asking.

We can next return briefly to questions about what, in non-moral senses, we should do, or ought to do. These are questions, we can now say, about what we *ought practically* to do. We can call some possible act

> what we *ought practically* to do in the *fact-relative* sense just when and because this act is what we have decisive reasons, or most reason, to do.

This fact-relative sense of 'ought' is what I am calling the *decisive-reason-implying* sense. When we are considering cases in which people

know all of the relevant, reason-giving facts, it may be enough to use this sense of 'ought'. In many cases, however, people do not know, or have false beliefs about, these relevant facts. In such cases, we can call some act

> what we *ought practically* to do in the *evidence-relative* sense just when this act would be what we had decisive reasons to do, if we believed what the available evidence gives us decisive reasons to believe, and these beliefs were true.

We can similarly call some act

> what we *ought practically* to do in the *belief-relative* sense just when this act would be what we had decisive reasons to do, if our beliefs about these facts were true.

We can also call some act

> what we *ought practically* to do in the *normative-belief-relative* sense just when this act is what we believe that we ought practically to do, or what we believe that we have decisive reasons to do.

As well as asking what we ought to do in these four senses, we can ask which acts are *rational*. We ought, I have claimed, to use the words 'rational' and 'irrational' to express certain kinds of praise or criticism. Questions about rationality are, in several ways, like questions about blameworthiness. The answers depend, for similar reasons, on the agent's beliefs. On the view that I defended earlier,

> (H) we ought rationally to act in some way when this act is what we ought practically to do in the belief-relative or normative-belief-relative senses.

In the case of the angry snake, for example, you ought rationally to run away, given your false belief that this act would save your life. In some cases, some act might be rational relative to our beliefs about the reason-giving facts, but irrational relative to our normative beliefs, or vice versa.

According to some writers, whether we ought rationally to act in some way depends *only* on our normative beliefs, and our acts are irrational only if we are either failing to respond to what we believe to be decisive reasons, or failing to do what we believe that we ought to do. This is like the view that acts are blameworthy only if the agent believes them to be wrong. Such views are, I have claimed, too narrow. Acts can be blameworthy even if the agent believes them to be right, as in the case of the conscientious SS officer. We should similarly claim that, if we are aware of facts that give us what are clearly and strongly decisive reasons to act in some way, we ought rationally to act in this way even if we fail to believe that these facts give us such reasons. Similar claims apply to our desires and aims. When Scarlet prefers agony next Tuesday to mild pain on any other day, his preference is irrational even though he is not failing to respond to what he believes to be a reason.

22 Other Kinds of Wrongness

We should distinguish, I have just claimed, between several moral senses of 'ought', 'right', and 'wrong'. I defined these senses by using a single sense, which I have called the *ordinary* sense. We can now ask whether we can explain this ordinary sense, and whether there is more than one such sense.

It can be unclear, or indeterminate, what we should claim to be part of the meaning of some word. It is unclear, for example, whether it is part of the meaning of the word 'cheetah' that cheetahs are hunters and have claws, or part of the meaning of 'war' that wars have to be declared. If we decide to include more in our accounts of the meaning of our words, we shall more often claim that some word has several senses. We might, for example, claim that the word 'war' has two senses, one of which applies only to wars that have been declared. I have already distinguished several senses of 'wrong', and I shall now distinguish several others. On a different account, to which I shall return, there is only one moral sense of 'wrong'. It is worth considering both accounts, but we need not choose between them.

Though I shall discuss the English word 'wrong', our questions are about the *concept* wrong, which is what is meant by this English word,

and by words in other languages with sufficiently similar meanings. This concept refers to the *property* of wrongness. (When we claim that some word, phrase, or concept *refers to some property*, we are not thereby claiming that anything *has* this property. There are many properties that nothing has, such as the properties of being a Greek god, or a witch.) If there are different senses of 'wrong', these senses express different versions of the concept *wrong*, which refer to different kinds of wrongness.

Like the concept of *a reason*, and the decisive-reason-implying concepts *should* and *ought*, at least one version of the concept *wrong* is indefinable, in the sense that it cannot be helpfully explained in other terms. We can use this concept to define some other moral concepts. We can say that some act is

> *right*, or *morally permitted*, when this act would not be wrong,

and that some act is

> *our duty*, *morally required*, or what we *ought morally* to do, when it would be wrong for us *not* to act in this way.

We might instead define this version of the concept *wrong* by appealing to an undefined version of one of these other concepts. Some act would be wrong, we might say, when we ought not to act in this way. But though we can explain how these concepts are related to each other, this group of concepts all have a common element which we cannot helpfully explain merely by using words. Like the concept of *a reason*, and the decisive-reason-implying concept *should* or *ought*, these moral concepts must be explained in other ways, by getting people to think certain thoughts. To express this indefinable version of the concept *wrong*, I shall use the phrase '*mustn't-be-done*'.

These moral concepts, I shall assume, also have other, definable versions. For example:

> in the *blameworthiness* sense, 'wrong' means 'blameworthy',
>
> in the *reactive-attitude* sense, 'wrong' means 'an act of a kind that gives its agent reasons to feel remorse or guilt, and gives others reasons for indignation and resentment',

> in the *justifiabilist* sense, 'wrong' means 'could not be justified to others',
>
> in the *divine command* sense, 'wrong' means 'forbidden by God'.

These senses can be combined to form more complex senses. For example, when we claim that some act is wrong, we might mean that this act is blameworthy because such acts are unjustifiable to others. Or we might mean that this act mustn't-be-done because such acts are forbidden by God.

Some people use 'ought morally' and 'wrong' in reason-implying senses. In what we can call the *decisive-reason* senses,

> 'what we ought morally to do' means 'what we have decisive reasons to do',

and

> 'wrong' means 'what we have decisive reasons *not* to do'.

These senses are misleading, and should not be used. We often believe that we have decisive reasons to act in some way, though we do not believe that we ought morally to act in this way. And if Rational Egoists used these decisive-reason senses, they would claim that

> (I) we ought morally to do whatever would be best for ourselves.

But Rational Egoism is best regarded, not as a moral view, but as an external rival to morality. On this view, we always have decisive reasons to do whatever would be best for ourselves, whether or not these acts would be morally wrong.

In what we can call the *decisive-moral-reason* senses,

> 'what we ought morally to do' means 'what we have decisive *moral* reasons to do',

and

> 'wrong' means 'what we have such reasons not to do'.

These senses do not, I believe, have much importance. We already have the concept of what we have decisive reasons to do, and it adds little to claim that some of these reasons are moral reasons. It is also unclear which reasons should be called 'moral'. It is unclear, for example, whether our reasons to promote the well-being of others should all be called moral reasons. Whether we ought morally to act in some way cannot be helpfully claimed to depend on how we ought to answer such partly verbal questions.

In what we can call the *morally-decisive-reason* senses,

> 'what we ought morally to do' means 'what we have morally decisive reasons to do',

and

> 'wrong' means 'what we have such reasons not to do'.

Though these senses may seem very similar to the *decisive-moral-reason* senses, there are two important differences. First, when we ask whether we have morally decisive reasons to act in some way, we are not asking whether we have decisive reasons of the kind that should be called 'moral'. We are asking whether we have reasons to act in this way that *morally outweigh* any reasons that we may have not to act in this way. Second, to be able to state our moral beliefs by using 'wrong' in the decisive-moral-reason sense, we must believe that we always have decisive reasons not to act wrongly. But if we claim instead that we have *morally* decisive reasons not to act in some way, that leaves it open whether these reasons are also *non-morally* decisive, or decisive *all things considered*. We could use 'wrong' in this sense even if we believed that, in some cases, we might have sufficient or decisive reasons to act wrongly.

Some people seem to use

> 'what we ought morally to do' to mean 'what we have the strongest impartial reasons to do'.

Some act is in this sense wrong when we have stronger impartial reasons to do something else. We can call these the *impartial-reason-implying* senses of 'ought' and 'wrong'. There are, as I have said, similar senses of 'good', 'bad', and 'best'. According to some Act Consequentialists:

> We ought always to do whatever would make things go best.

If this claim uses both 'ought' and 'best' in these impartial-reason-implying senses, it would mean:

> (J) What we have the strongest impartial reasons to do is whatever would make things go in the way in which we all have the strongest impartial reasons to want things to go.

We can call this view *Impartial-Reason Act Consequentialism*. To express this sense of 'ought', we can use the phrase *ought-impartially*.

This sense of 'ought' differs significantly from more familiar moral senses. Sidgwick, for example, writes:

> the good of any one individual is of no more importance, from the point of view . . . of the Universe, than the good of any other . . . And . . . as a rational being I am bound to aim at good generally . . . not merely at a particular part of it . . . I ought not to prefer my own lesser good to the greater good of another.

When Sidgwick claims that he *ought* not to prefer his own lesser good, he does not seem to mean that such a preference would be blameworthy, or unjustifiable to others, or that such an act would give him reasons for remorse and give others reasons for indignation. Sidgwick seems to mean only that, when assessed from an impartial point of view, his reason to give himself some lesser good is weaker than, or outweighed by, his reason to give some greater good to someone else.

This kind of Consequentialism may be better regarded, not as a moral view, but as being, like Rational Egoism, an external rival to morality. Given this view's claim that we ought to sacrifice our lesser good for the greater good of others, it is much closer to morality. That makes this view, in some ways, a more serious rival. Impartial-Reason Act

Consequentialism may be accepted by many people who would reject Rational Egoism, because they regard their own well-being as what Sidgwick calls a 'narrow' and 'ignoble end'.

(J) may seem to be a trivial claim, which is close to a tautology. It is not, however, trivial to claim that acts can be right or wrong, and outcomes can be good or bad, in these impartial-reason-implying senses. On some widely accepted views about reasons, as I have claimed, there are no such acts or outcomes. And even if (J) were a tautology, Impartial-Reason Act Consequentialists could make other, substantive claims. If they are Hedonistic Utilitarians, for example, these people might claim

> (K) What we ought-impartially to do is whatever would produce the greatest sum of happiness minus suffering.

These people may believe that we all have strong reasons to act in this way. And they might not act upon, or even have, moral beliefs that involve any of the more familiar senses of 'ought morally' and 'wrong'. These people may be convinced that it matters greatly how well things go, and they may be strongly motivated and often moved to act in ways that prevent or relieve suffering. But they may be doubtful whether any acts are duties, or mustn't-be-done, and doubtful about blameworthiness, and about reasons for remorse and indignation. That is one way in which this form of Consequentialism is an external rival to morality.

According to some writers, as I have said, there is only a single moral sense of 'wrong', 'right', and 'ought'. It would be implausible to make this claim about one of the definable senses. If we can use 'wrong' in one definable sense, we can surely use it in others. Nor is there any one definable sense that can be plausibly claimed to be the only sense that everyone uses. We cannot even claim that everyone uses 'wrong' to mean 'what we have morally decisive reasons not to do', since some people never or seldom use the concept of a reason.

It would be more plausible to claim that everyone uses 'wrong' in the indefinable sense that I am expressing with the phrase 'mustn't-be-done'. The blameworthiness and reactive-attitude senses might be claimed to appeal implicitly to this indefinable sense, because the attitudes of blame, guilt, remorse, and indignation all involve the belief

that some act is wrong. In defining the morally-decisive-reason sense of 'wrong', we might have to use the word 'morally' indefinably. And some other definable senses might be claimed to express, not the belief that certain acts are wrong, but certain other beliefs about wrong acts. The divine command and justifiabilist senses might, for example, express the beliefs that acts are wrong, in the sense that they mustn't-be-done, when and because these acts are forbidden by God, or unjustifiable to others.

When some writers claim that words like 'wrong' and 'ought' have only one moral sense, they appeal to the fact that, even when we and other people have very different moral views, we regard ourselves as *disagreeing* with these other people. If we and others used these words in different senses, these writers claim, we could not be disagreeing with these other people, since we wouldn't be discussing the same questions.

This argument is weak. Different people may use 'wrong' or 'ought' in different definable senses that partly overlap. That may be enough to make disagreement possible. Suppose for example that, when I claim that some act is wrong, I mean that such acts are blameworthy because they are forbidden by God. When you claim that some act is wrong, you mean that such acts are blameworthy because they are unjustifiable to others. If I claimed that some act was wrong and you claimed that it wasn't, we would be disagreeing about whether this act was blameworthy. And when people use 'wrong' in such different senses, that may *increase* their disagreements. In the case just imagined, if we understood each other's use of 'wrong', you might believe that no acts are in my sense wrong, since you believe that no acts are blameworthy because they are forbidden by God. I might believe that no acts are in your sense wrong, since I believe that no acts are blameworthy because they are unjustifiable to others. We would then completely disagree, since each of us would reject all of the other's moral beliefs.

When different people in the same community use words like 'wrong' or 'ought' in such different, partly overlapping senses, these people have reasons to move to other, thinner senses, which they can all use. It would then be clearer when these people disagree, and what they are disagreeing about. In the case just imagined, if you and I both used 'wrong' to mean

'blameworthy', we would be able to agree that many acts are in this sense wrong, even though we disagreed about what makes these acts wrong.

In some cases, we can add, those who use 'wrong' or 'ought' in different senses may *not* be disagreeing. On Sidgwick's view, for example, I ought to give up my life if I could thereby save the lives of two strangers who are relevantly like me. If Sidgwick were using 'ought' in the blameworthiness or reactive-attitude senses, most of us would reject this claim. We would believe that, if I saved myself rather than these two strangers, my act would not be blameworthy, and I would have no reason to feel remorse, nor would these strangers or others have any reasons to be indignant. But Sidgwick might mean only that I would have stronger impartial reasons to save the two strangers. That claim would not conflict with other people's moral beliefs.

Consider next those cases in which the rightness of our acts depends on the goodness of their effects. In such cases, some people claim that

> (L) we ought to do what would make things go best,

and others claim that

> (M) we ought to do what would make things go expectably-best.

If (L) uses 'ought' in the fact-relative sense, and (M) uses 'ought' in the evidence-relative sense, these claims do not conflict, and we could accept them both. Nor would either claim conflict with a version of (M) that used 'ought' and 'expectably-best' in belief-relative senses.

There is another avoidable disagreement. According to some writers, we ought to do certain things, such as keeping our promises, saving people's lives, and doing what would make things go expectably-best. According to some other writers, we ought to *try* to do these things. We ought, I believe, to make both these claims. We should not claim only that we ought to *do* these things, since it is morally important whether we tried to do them. We may deserve no blame, for example, if we tried but failed to keep some promise, or to save someone's life. Nor should we claim only that we ought to *try* to do certain things, since it is often morally important whether our acts succeed. If our attempt to keep some promise fails, for example, it may be true that we ought to act in

some other way instead. When we claim that we ought to do something, we should often be taken to mean that we ought to do this thing or at least try to do it.

It is unimportant whether the various senses that I have described should be called different senses of 'wrong', which refer to different kinds of wrongness. It is enough to distinguish these senses, and the concepts that they express. We can then decide which of these concepts are most worth using.

In making that decision, we can return to the question of how much morality matters in the reason-implying sense. If some possible act would be wrong, does this fact give us a reason not to do it? If so, how strong are such reasons?

The answers depend in part on what we mean by 'wrong', and on the kind of wrongness to which our use of 'wrong' refers. Suppose first that, in claiming that some act is wrong, we mean that we have decisive moral reasons not to act in this way. These reasons would be provided by the facts that made some act wrong. Two examples might be the facts that some act would be a lie or would cause pointless suffering. On this view, the fact that

> (N) some act is wrong

would be the higher-order fact that

> (O) there are certain other facts that give us decisive moral reasons not to act in this way.

This higher-order fact would not give us a *further*, *independent* reason not to act in this way. Though we might claim that an act's wrongness always gives us a reason not to do it, this reason would be *derivative*, since its normative force would derive entirely from these other reason-giving facts. So if we used 'wrong' only in this decisive-moral-reason sense, we could claim that

> (P) when some act would be wrong, this fact would not give us any further reason not to act in this way.

On this view, it would have no practical importance whether some act would be wrong. When we were trying to decide what to do, it would always be enough to ask whether we had decisive reasons for or against acting in any of the possible ways. If we decided that we had such reasons, we could then ask whether these were *moral* reasons, so that our act was wrong in the decisive-moral-reason sense. But this would not be a question about what we ought to do, or had reasons to do. This question would be merely conceptual, like the questions of which are the kinds of reason that can best be called legal, or aesthetic. So we have little reason, I believe, to use this sense of 'wrong'.

Many people assume that an act's wrongness does give us strong or even decisive further reasons not to do it. If these people use 'wrong' in the decisive-moral-reason sense, their assumption would be false, in the way that I have just described. That does not show that these people cannot be using 'wrong' in this sense, since these people may not have seen the point that I have just made. But most of us, I believe, use 'wrong' in one or more other senses. And when certain acts would be wrong in these other senses, we *can* claim that the wrongness of these acts gives us further, independent reasons not to act in these ways.

Suppose first that we use 'wrong' in the indefinable sense. When we claim that some act is in this sense wrong, we are not claiming that this act has what Scanlon calls the 'purely formal, higher-order property' of having other, reason-giving properties. We are claiming that this act has the highly distinctive substantive property of being something that *mustn't-be-done*. Though I believe strongly that some acts are in several other senses wrong, it seems to me a more open question whether any acts have this indefinable property. But if they do, we could plausibly claim that, when some act mustn't-be-done, that gives us a very strong reason not to do it. This is one of the senses of 'wrong' with which it seems most plausible to claim that

> (Q) when some act would be wrong, this fact always gives us a decisive reason not to do it.

(Q) would be just as plausible, though for significantly different reasons, if we used 'wrong' to mean 'forbidden by God'.

If we use 'wrong' in the other definable senses, we could similarly claim that an act's wrongness gives us independent reasons not to do it. When some act would be blameworthy, unjustifiable to others, and is an act that would give us reasons for remorse and give others reasons for indignation, these facts would all give us further reasons not to act in this way. We should not, however, claim that these facts would always give us our *strongest* reasons not to act wrongly. If some act would cause great suffering, for example, that might give us a much stronger reason than the reasons given by the facts that this act would be blameworthy and unjustifiable to others.

As I have said, we need not choose between these senses of 'wrong', and the concepts that they express. It is worth using several of these concepts, asking, for example, which acts are wrong in the indefinable, justifiabilist, reactive-attitude, or blameworthiness senses. In the rest of this book I shall use 'ought morally' and 'wrong' vaguely, in some combination of these senses.

There are some deep and difficult questions about how we should understand these normative concepts, and about whether acts can have the properties to which these concepts refer. Except in Part Six, I shall say little about these *meta-ethical* questions. Such questions will be easier to answer when we have made more progress in our thinking about practical and epistemic reasons, and about morality. As Rawls and Nagel claim, our moral theories 'are primitive, and have grave defects', and 'ethical theory . . . is in its infancy'.

Rather than proposing a new moral theory, I shall try to learn from some existing theories, hoping to get somewhat closer to the truth. I shall start with Kant, because he is the greatest moral philosopher since the ancient Greeks. When Kant presents his famous formulas, his aim, he writes, is to find 'the supreme principle of morality'. I shall ask whether he succeeds.

PART TWO

PRINCIPLES

This page intentionally left blank

8

Possible Consent

23 Coercion and Deception

According to Kant's best-loved principle, often called

> *the Formula of Humanity*: We must treat all rational beings, or persons, never merely as a means, but always as ends.

To treat people as ends, Kant claims, we must never treat them in ways to which they could not consent. In explaining the wrongness of a lying promise, for example, Kant writes

> he whom I want to use for my own purposes with such a promise cannot possibly agree to my way of treating him.

Korsgaard comments:

> People cannot assent to a way of acting when they are given no chance to do so. The most obvious instance of this is when coercion is used. But it is also true of deception . . . knowledge of what is going on and some power over the proceedings are the conditions of possible assent.

O'Neill similarly writes:

> if we coerce or deceive others, their dissent, and so their genuine consent, is in principle ruled out.

Korsgaard concludes:

> According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrong-doing to others.

These remarks suggest this argument:

> It is wrong to treat people in any way to which they cannot consent.
>
> People cannot consent to being coerced or deceived.

Therefore

> Coercion and deception are always wrong.

It is sometimes right, however, to treat people in ways to which they cannot consent. When people are unconscious, for example, they cannot consent to life-saving surgery, but that does not make such surgery wrong.

Kant's claim, Korsgaard might say, applies only to acts whose nature makes consent impossible. Deception, unlike surgery, is such an act. For people to be able to consent to our way of treating them, they must know what we are doing. If people knew that we were trying to deceive them, we would be unable to deceive them. So we cannot possibly deceive people with their consent. This might be why, unlike surgery, deception is always wrong. But consider

> *Fatal Belief*: I know that, unless I tell you some lie, you will believe truly that *Brown* committed some murder. Since you could not conceal that belief from Brown, he would then murder you as well.

If I say nothing, you could reasonably complain with your dying breath that I ought to have saved your life by deceiving you. I could not defensibly reply that, since I could not have deceived you with your consent, this way of saving your life would have been wrong. My life-saving lie *would* be like life-saving surgery on some unconscious

person. Just as this person would consent to this surgery if she could, you would consent to my deceiving you. It is a merely technical problem that, if I asked you for your consent, that would make my deceiving you impossible. We could solve this problem if you had the ability to make yourself lose particular memories. After you had given your consent, you could deliberately forget our conversation, so that my lie could save your life. Since you would consent to my deceiving you if you could, my lie would be morally as innocent as some lie that was needed to give someone a surprise party.

Similar remarks apply to coercion. People could not consent to being coerced, it might be claimed, because if people gave consent they would not be being coerced, and if they were being coerced they could not freely give consent. But we can freely consent to being later coerced in some way. Before the discovery of anaesthetics, many people freely consented to being later coerced during painful surgery. And we can freely consent to some kinds of coercion even while we are being coerced. Most of us would vote in favour of everyone's continuing to be legally coerced, by threats of punishment, to pay fair taxes and obey good laws. I would consent to being coerced to be less untidy. Though deception and coercion are often wrong, what makes them wrong is not, I believe, the fact that these are acts whose nature makes consent impossible.

24 The Consent Principle

Return now to Kant's claim that

> (A) it is wrong to treat people in any way to which they cannot possibly consent.

People cannot consent, Korsgaard writes, 'when they are given no chance to do so'. O'Neill similarly writes, 'To treat others as persons we must allow them the *possibility* either of consenting to or of dissenting from what is proposed'. These remarks assume that Kant means

> (B) It is wrong to treat people in any way to which they cannot possibly consent because we have not given them the possibility of giving or refusing consent.

When we treat people in some way, they can often give or refuse consent in a *declarative* sense, by telling us or others that they do or don't consent. Korsgaard and O'Neill use 'consent' in a different and more important sense. People can give or refuse consent in this *act-affecting* sense if they have what Korsgaard calls 'power over the proceedings', because they will be treated in some way only if they consent. So we can restate (B) as

> *the Choice-Giving Principle*: It is wrong not to give other people the power to choose how we shall treat them.

If this were what Kant meant, we would have to reject Kant's claim, since the Choice-Giving Principle is clearly false. This principle mistakenly implies, for example, that we ought to let other people choose whether or not we give their student essays low grades, buy what they are trying to sell us, take back what they stole from us, report their crimes, or vote against them in some election. In most morally important cases, moreover, our choice between different possible acts would have significant effects on two or more people. We could not give to more than one of these people the power to choose how we shall act, as would be shown if two of these people made conflicting choices. So the Choice-Giving Principle also mistakenly implies that, in all these cases, whatever we did would be wrong.

There is, I believe, a better way to interpret Kant's remarks. Korsgaard and O'Neill assume that, when Kant claims

> (A) It is wrong to treat people in any way to which they cannot possibly consent,

he means

> (C) It is wrong to treat people in any way to which they cannot consent in the act-affecting sense because we have not given them the power to choose how we treat them.

I suggest that Kant means

> (D) It is wrong to treat people in any way to which they *could not* consent in the act-affecting sense, *if* we gave them the power to choose how we treat them.

It might be objected that, if we gave people this power, they *could* choose that we act in any of the possible ways, so there would never be any act to which these people could not consent. If this were the kind of impossibility that Kant had in mind, (D) would be trivial, since (D) would never imply that some act is wrong. But there is another kind of impossibility. When people say 'I cannot possibly consent to your proposal', they hardly ever mean that giving consent is not one of the choices that is open to them. These people nearly always mean that they could not *rationally consent*, because they have decisive reasons to refuse consent. Kant, I suggest, means

> (E) It is wrong to treat anyone in any way to which this person could not rationally consent.

I shall call this the *Consent Principle*.

We have several reasons to believe that Kant is appealing to this principle. While the Choice-Giving Principle is obviously false, the Consent Principle might be true, which makes it more likely to be what Kant means. When Kant claims that we could not do something, he often means that we could not rationally do this thing. Kant also writes that, if he treated someone wrongly, this person

> could not possibly agree to my way of treating him, *and so himself contain the end of this act*.

If Kant were claiming that we ought to let other people choose how we shall treat them, he would have no reason to add that, for our treatment of someone to be justified, this person must be able to 'contain the end of this act', by sharing this act's aim. When we let other people choose how we shall treat them, we are not acting with some aim that these people might be unable to share. Kant must mean that, when *we* are choosing how we shall treat other people, we ought always to act with some aim that these people would be able to share. Nor would it be enough if these people could *conceivably* share our aim, since many unjustifiable

aims could conceivably be shared. We ought to act only with some aim that other people could *rationally* share, so that they could rationally consent to our way of treating them.

Kant's remark about shared ends or aims, though helping to explain his claims about consent, is in itself less plausible. Even if other people could rationally share our aim, we may be acting wrongly if and because these people could not rationally consent to our way of achieving this aim. Though you could rationally share my aim that my tame tiger be fed, you could not rationally consent to being what my tiger eats. And even if other people could *not* rationally share our aim, we may not be acting wrongly if these people could rationally consent to our act. Though you could not rationally share my aim of reciting someone's name a thousand times, you could rationally consent to my reciting your name. So, compared with the question whether other people could rationally share our aims, it is more important whether these people could rationally consent to our acts.

Kant's claims about consent give us an inspiring ideal of how, as rational beings, we ought all to be related to each other. It is worth asking whether we could achieve this ideal. We cannot always let everyone choose how we treat them. But we might be able to treat everyone only in ways to which they could rationally consent. And if that is possible, Kant may be right to claim that this is how everyone ought always to act.

25 Reasons to Give Consent

Whether we could achieve Kant's ideal depends on which are the acts to which people could rationally consent. Rawls suggests that, in proposing the Consent Principle, Kant assumes that

> (F) people could rationally consent to some act if and only if, or *just when*, they could will it to be true that the agent's maxim is a universal law.

Rawls is referring here to another of Kant's proposed statements of the supreme principle of morality. According to Kant's

> *Formula of Universal Law*: It is wrong to act on any maxim that we could not will to be a universal law.

By our *maxims* Kant means, roughly, our policies and underlying aims. We need not yet consider in what sense maxims might be willed to be universal laws.

Kant does not, however, commit himself to (F). And this assumption would be a mistake. Suppose that I am your doctor, and I ask you whether you consent to my giving you some medical treatment. For it to be rational for you to consent, you would need to have beliefs about whether I am a well-qualified and conscientious doctor, and about what effects this and the other possible treatments would be likely to have. But you wouldn't need to have beliefs about whether I would be acting on some maxim, or policy, that you could will to be a universal law.

To support his suggestion that Kant assumes (F), Rawls appeals to Kant's remark that all of his various principles are merely different statements of 'precisely the same law'. Rawls takes this remark to imply that Kant's other principles 'cannot add to the content' of Kant's Formula of Universal Law. Rawls therefore proposes that we should try to interpret Kant's other principles in ways that make them add nothing, because they contain no other ideas.

Kant is a greater philosopher than this proposal assumes. Kant himself goes even further in underrating his achievements, since he denies that he is presenting even one new principle. The truth is that, in the cascading fireworks of a mere forty pages, Kant gives us more new and fruitful ideas than all the philosophers of several centuries. Of the qualities that enable Kant to achieve so much, one is inconsistency. If we ignore some of Kant's claims because they conflict with others, we may miss some of what Herman calls the 'untapped theoretical power and fertility' of Kant's ideas.

Kant's Consent Principle is one example. It is surprising that this principle has been so little discussed. This principle has great appeal, and is worth considering as a separate moral idea, not merely as another way of stating Kant's Formula of Universal Law. So in asking what this principle implies, I shall not assume (F).

When we ask whether someone could rationally consent to some act, our question should be about consent in the *act-affecting* sense. It is not worth asking whether people could rationally consent to being treated in some way, if their refusal of consent would be a mere declaration,

or protest, which would either make no difference to how others would treat them, or might make others treat them even worse. If that were true, it might be rational for these people not to protest, even if they were being treated in ways that were very bad for them, and very wrong.

Our question should also be about *informed* consent. When people do not know what effects some act might have, it is irrelevant whether they could rationally consent to this act. People could rationally consent to being grossly maltreated, if they did not know what was being done to them. For these reasons, we can restate the Consent Principle as

> CP: It is wrong to treat people in any way to which they could not rationally consent in the act-affecting sense, if these people knew the relevant facts, and we gave them the power to choose how we shall treat them.

We should be counted as *treating* people in some way when we know that our act, or one of its possible alternatives, would or might affect these people in some way, or be an act with which they would have some personal reason to be concerned. That could be true even when our way of acting would not causally affect these people. Two examples would be failing to save someone's life, or breaking a promise to someone who is dead.

When people know the relevant facts, they could rationally consent to some act just when these facts would give them sufficient reasons to consent. People have *sufficient* reasons to consent to some act when these reasons are not weaker than any reasons they might have to refuse consent. So the Consent Principle could be more briefly stated as

> CP2: It is wrong to treat people in any way to which they would not have sufficient reasons to consent in the act-affecting sense.

In stating this principle in these ways, I assume that we are rational insofar as we respond to reasons or apparent reasons. On some other views about rationality, CP and CP2 state different principles, which might have different implications. If you accept such a view, you should take the Consent Principle to be stated by CP2. When I ask whether

someone could rationally consent to some act, I shall be asking whether this person would have sufficient reasons to consent.

For the Consent Principle to succeed, it must both be in itself plausible, and have plausible implications. This principle must not require too many acts that seem to us to be clearly wrong, or *condemn*—in the sense of implying to be wrong—too many acts that seem to us to be clearly morally required. If this principle both implies and plausibly supports many of our best considered intuitive moral beliefs, we could justifiably use this principle to guide some of these beliefs, by revising or extending them.

What the Consent Principle implies depends on our assumptions about which facts give us reasons. If we assume either some desire-based subjective theory, or Rational Egoism, the Consent Principle would not be plausible, and would mistakenly condemn many permissible or morally required acts. Suppose, for example, that in

> *Earthquake*, two people, *White* and *Grey,* are trapped in slowly collapsing wreckage. I am a rescuer, who could prevent this wreckage from either killing White or destroying Grey's leg.

White, Grey, and I, we should assume, are all strangers to each other; nor do we differ in any other morally relevant way. We should make similar assumptions about my later imagined cases. If these are the only morally relevant facts, it is clear that I ought to save White's life. We can next suppose that, if I saved Grey's leg, that would be much better for Grey, and would much better fulfil Grey's present fully informed desires. According to both desire-based subjective theories, and Rational Egoism, Grey could not then rationally consent to my failing to save her leg, so the Consent Principle would mistakenly imply that it would be wrong for me to save White's life. Similar claims apply to countless other cases. There are countless right acts to which, according to both subjective theories and Rational Egoism, some people could not rationally consent. If we accept any of these theories, as many people do, we must reject the Consent Principle. That may be one reason why this principle has been so little discussed.

We ought, I have claimed, to accept some *wide value-based objective* theory. On such views, when one of two possible choices would make things go in a way that would be impartially better, but some other choice would make things go better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to make either choice. *Earthquake*, I believe, is one such case. If Grey could choose how I would act, she would have sufficient reasons, I believe, to make either choice. Grey could rationally choose that I save her leg, since this choice would be much better for her. But she would not be rationally required to make this choice. Grey could rationally choose instead that I save White's life. Grey could rationally regard White's well-being as mattering about as much as hers, and White's loss in dying would be much greater than Grey's loss in losing her leg.

White, in contrast, could not rationally choose that I save Grey's leg. We could often rationally choose to benefit some stranger, I believe, even if our choice would make us lose a somewhat greater benefit. But there is too great a difference between the possible benefits to White and Grey. White would not have sufficient reasons to give up her life so that I could save Grey's leg. So the Consent Principle rightly requires me to save White's life, since this is the only act to which both Grey and White could rationally consent.

Suppose next that, in

> *Lifeboat*, I am stranded on one rock, and five people are stranded on another. Before the rising tide drowns all of us, you could use a lifeboat to save either me or the five. We are all young, and would lose as much in dying.

Though some people would believe that you ought to give me some chance of being saved—which might be a chance of one in six or even one in two—most people would believe, more plausibly, that you ought to save the other five people.

If I could choose how you will act, could I rationally choose that you save the five rather than me? Some people would answer No. These people might agree that, if I chose to give up my life to save five strangers, this choice would be morally admirable. But this choice,

they believe, would also be irrational. On this view, since I could not rationally consent to your saving the five rather than me, the Consent Principle implies that it would be wrong for you to save the five. That is an unacceptable conclusion. So if we accept this view, we would have to reject the Consent Principle.

We ought, I believe, to reject this view. Though I could rationally choose that you save me, I could also rationally choose, I believe, that you save the five. I would have sufficient reason to give up my life if I could thereby save five strangers.

Could the five rationally consent to your saving me rather than them? The word 'consent' may be misleading here, since we may assume that each of the five could give consent only on her own behalf. But we should not make that assumption. When we apply the Consent Principle, we should ask whether, if each of the five could give or refuse consent to your act in the act-affecting sense, thereby choosing how you will act, this person could rationally choose that you save me rather than the five. The answer is clearly No. Suppose that *Green* is one of the five. Green would not have sufficient reasons to choose that you save me rather than saving *both* Green *and* four other people. Green would have both strong personal and strong impartial reasons not to make this choice. On these assumptions, the Consent Principle rightly implies that you ought to save the five, since this is the only act to which both I and each of the five would have sufficient reasons to consent.

As these examples suggest, whether we could rationally consent to some act depends in part on the benefits or burdens that would come to us or other people in the different outcomes that would be produced by this and the other possible acts. It makes a difference both how great these benefits or burdens would be, and to how many people they would come. It also makes a difference, I believe, how badly off we and the other people are. And it may make a difference whether we or the others are responsible for various features of our situation. That might be true, for example, if some of us have worked to produce the possible benefits, or are responsible, through negligence or recklessness, for the possible burdens. There may be other acts to which we would not have sufficient reasons to consent even though these acts would not impose any significant burden on us. We can have strong reasons, for example,

to refuse consent to other people's deciding how our lives will go, even when these people's decisions would not be bad for us.

Whenever people could not rationally give informed consent to being treated in some way, there must be facts about these acts which give these people decisive reasons to refuse consent. White, I have claimed, could not rationally consent to my saving Grey's leg rather than White's life, given the fact that White's loss would be so much greater than Grey's. This fact can also be plausibly claimed to make this act wrong. Similar claims apply to other cases. Whenever certain facts would give some people decisive reasons to refuse consent to being treated in some way, these facts would also provide moral objections to these acts.

For the Consent Principle to be true, these moral objections must be decisive, since this principle condemns all acts to which anyone could not rationally consent. For this much stronger claim to be defensible, it must be always or nearly always true that

> (G) there is at least one possible act to which everyone would have sufficient reasons to consent.

We can call (G) the *Unanimity Condition*. In cases in which there was no such act, the Consent Principle would mistakenly imply that whatever we did would be wrong. (G) is least likely to be true when

> (H) each of our possible acts would impose some very great burden on at least one person, or would deny at least one person some very great benefit.

Such people would have very strong reasons to refuse consent to being made to bear such burdens, or being denied such benefits. One such case is *Lifeboat*, in which either I or the five will be denied the benefit of being saved from an early death. In this case, I have claimed, (G) is true. Though I would have very strong reasons to choose that you save my life, these reasons would not be decisive. I would have sufficient reasons, I believe, to consent to your saving the five rather than me. If I would have such reasons, that strongly supports the view that, at least in cases in which the stakes are lower, there would be at least one possible act to which everyone could rationally consent.

I shall return to the question whether there would always be such an act. If that is true, we could argue:

> Whenever someone could not rationally consent to some act, there must be certain facts that give this person decisive reasons to refuse consent. These facts provide moral objections to this act.
>
> These objections must be significantly stronger than the objections to any other possible act to which everyone *could* rationally consent.
>
> Whenever there are significantly stronger moral objections to one of two acts, this act is wrong.

Therefore

> It is wrong to act in any way to which anyone could not rationally consent.

Though this argument is rough, it is enough to show that the Consent Principle is in itself plausible.

This principle also has many plausible implications, since it condemns many of the acts that are most clearly wrong, such as many acts of killing, injuring, coercing, deceiving, stealing, and promise-breaking. Many of these acts treat people in ways to which they would not have sufficient reasons to consent.

26 A Superfluous Principle?

According to some writers, nothing is achieved by appealing to the possibility of rational consent. These writers concede that it may always be wrong to treat people in ways to which they could not rationally consent. But what is morally important, these writers claim, is not the fact that these people could not rationally consent to these acts, but the various facts that give these people decisive reasons to refuse consent.

In considering this objection, we can first distinguish two aims that any moral principle might achieve. This principle might provide a

reliable *criterion* of wrongness, by truly telling us that all acts of a certain kind are wrong. This principle might also be *explanatory*, by describing one of the reasons why these acts are wrong, or one of the facts that make them wrong. According to the writers I have just mentioned, even if the Consent Principle is true, we do not need this principle as a criterion, nor is this principle explanatory.

This objection has most plausibility when we consider acts whose main effects would be on one person, with whom we cannot communicate and whose preferences we don't know. In such a case, we would have to make some decision on this person's behalf. Surgeons, for example, sometimes have to make decisions on behalf of their unconscious patients. In such cases, it may be enough to claim that we ought to try to do what would be best for this other person, or what would benefit this person most. It may not be worth adding that it would be wrong for us to act in any way to which this person could not rationally consent.

In most important cases, however, our choice between possible acts would have significant effects on two or more people. The view that I have just described might be widened to cover such cases. According to *Act Utilitarianism*, or

> *AU*: We ought always to do whatever would, on the whole, benefit people most, by giving people the greatest total sum of benefits minus burdens.

Act Utilitarians might claim that

> (I) everyone could rationally consent to all and only the acts that would, on the whole, benefit people most.

If (I) were true, AU and the Consent Principle would always *coincide*, by requiring all the same acts. These Utilitarians might then claim that AU is more fundamental, and that, since AU tells us how we ought always to act, the Consent Principle adds nothing to our moral thinking. But this claim would be false. If it were only these Utilitarian acts to which everyone could rationally consent, the Consent Principle would support AU. (I)'s truth would give us a further reason to believe that these acts were morally required, and a further reason to act in these ways.

(I) is not, I believe, true. There are many Utilitarian acts to which some people could not rationally consent, and many non-Utilitarian acts to which everyone could rationally consent. I shall give some examples later.

If the Consent Principle is true, this principle would be more than a reliable criterion of wrongness. Whenever someone could not rationally consent to being treated in some way, this fact would provide an objection to this act, and could be claimed to be one of the facts that would make this act wrong. The Consent Principle would have most importance when we must choose between many possible acts that would have significant effects on many people, whose interests or aims conflict. In such cases, if there is only one possible act to which everyone could rationally consent, this fact would give us a strong reason to act in this way, and might be enough by itself to explain why all the other possible acts would be wrong.

We have another reason to ask whether the Consent Principle is true. Even if we do not need to use this principle as a criterion of wrongness, it is worth asking whether we could achieve what I call *Kant's ideal*, by treating everyone only in ways to which they could rationally consent.

27 Actual Consent

It is often morally important whether people *actually* consent to being treated in some way, or whether, if they had the opportunity, these people *would in fact* consent. In such cases, it is not enough to ask whether people *could rationally* consent to some act. Some rapist might claim that his victim could have rationally consented to having sexual intercourse with him. Even if this claim were true, that would not justify this man's act. It may be objected that, since the Consent Principle does not require actual consent, this principle mistakenly ignores the moral importance of such consent.

That is not, however, true. Even if this man's victim could have rationally consented to having sexual intercourse with him, she could not have rationally consented to being raped, by having such intercourse forced on her despite her actual refusal of consent. In this and many other

kinds of case, we could not rationally consent to being treated in some way without our actual consent. Since the Consent Principle condemns all such acts, this principle does not ignore the moral importance of actual consent.

This principle might instead be claimed to give, implicitly, *too much* importance to actual consent. Consider

> *the Veto Principle*: It is wrong to treat people in any way to which they either do in fact, or would in fact, refuse consent.

Like the similar Choice-Giving Principle, this principle is clearly false. There are countless permissible or morally required acts to which some people either do or would refuse consent. In *Earthquake*, for example, even if Grey refuses her consent, I ought to save White's life rather than Grey's leg. And there is often no possible act to which everyone would in fact consent. Someone might now argue:

> It is wrong to treat people in any way to which they could not rationally consent.
>
> (J) No one could rationally consent to being treated in any way to which they either do in fact, or would in fact, refuse consent.

Therefore

> It is wrong to treat people in any way to which these people either do in fact, or would in fact, refuse consent.

If (J) were true, the Consent Principle would imply the Veto Principle. That would make the Consent Principle clearly false.

Should we accept (J)? It may be confusing to ask whether people could rationally consent to some act to which they actually refuse consent, since these people could not at the same time both give and refuse consent. To make our question clearer, we can appeal to another version of the Consent Principle. According to

> CP3: It is wrong to treat people in any way to which, if they had known the relevant facts, these people could not have rationally given, in advance, their irreversible consent.

Our consent to some act is *irreversible* when we know that, if we later withdrew our consent, that would make no difference to how we would later be treated.

There are many acts to which we could not rationally give such irreversible consent in advance. For example, we could seldom rationally give such consent in advance to sexual acts to which, at the time of these acts, we refuse consent. That would seldom be rational because the nature of most sexual acts is greatly affected by whether, at the time, both or all of the people involved actually consent.

There are also many acts, however, to which we *could* rationally give such irreversible consent. For us to have sufficient reasons to give such consent, it might have to be true both that

> (K) we have some reason to give irreversible consent, thereby restricting our future freedom,

and that

> (L) we shall not later learn some fact that might give us decisive reasons to regret that we earlier gave such consent.

But these conditions are often met. In many cases, for example, someone needs to know that someone else's consent is binding, and cannot be withdrawn. Suppose that, in *Earthquake*, once I had started to save White's life rather than Grey's leg, it would be dangerous for me to stop. Suppose next that Grey knows all of the relevant facts, and that Grey is just as able to make a good decision now as she will later be. On these assumptions, Grey could rationally make her decision now. We are not rationally required to postpone our decisions whenever we can. And Grey would have sufficient reasons, I have claimed, to choose that I save White's life rather than Grey's leg. If that is so, Grey would also have sufficient reasons to give irreversible consent to my later doing that. Grey could rationally say, 'Go ahead and save White's life, even if I later change my mind'.

When we apply the Consent Principle in the form stated by CP3, our aim is only to ask whether people could rationally consent to being treated in some way to which they in fact refuse consent. This question is easier to answer when we apply it to irreversible consent given in advance. In many actual cases, people would not in fact have sufficient reasons to give such consent in advance, thereby committing themselves in a way that would restrict their future freedom. But given the aims of our imagined thought-experiment, we can *suppose* that these people would have had sufficient reasons to make their decision in advance. Our question can be whether, on that supposition, these people would have had sufficient reasons to give their irreversible consent.

Grey, I have claimed, would have had such reasons. And in many other cases, I believe, people could rationally give such irreversible consent to being later treated in some way without their later actual consent. If that is true, we can reject premise (J) of the argument above. The Consent Principle does not imply the Veto Principle, and avoids at least the strongest objections to that principle.

Though we ought to reject the Veto Principle, we could plausibly accept a much weaker version of this principle. According to what we can call

> *the Rights Principle*: Everyone has rights not to be treated in certain ways without their actual consent.

When we claim that people have *rights* not to be treated in certain ways, we mean in part that, without these people's consent, such acts would be wrong. We can call these the *veto-covered* acts.

In stating this principle, it would often be hard to decide which are the acts that people have a right to veto. For this principle to be acceptable, these rights must be narrowly described. We should not, for example, claim that everyone always has a right not to be killed, since some killings are unavoidable, and some others are justified, as is true in some cases of self-defence. But we might claim that we all have certain more restricted rights, such as a right not to be killed for our own good without our consent. We might similarly claim that everyone has a right to veto what is done to their bodies, not only sexually but in other ways.

On one view, for example, everyone has a right not to be kept alive by medical treatments to which they refuse consent.

As well as condemning veto-covered acts to which people refuse consent, the Rights Principle should require us to give people the opportunity to refuse consent. When we cannot give people this opportunity, because we cannot communicate with them, we ought to try to treat these people only in those veto-covered ways to which, *if* they had the opportunity, they *would* consent. When people cannot consent to some act, but we know that they would have given or refused consent, these facts would have similar moral significance. When we ask whether people *would in fact* consent to some act, that is quite different from asking whether these people *could rationally* give such consent. We might know that certain people would not in fact consent to some veto-covered act, even though it would be irrational for them to refuse consent. In such cases, we might say, people have a right to be irrational, and to suffer the effects.

For consent to be morally significant, however, it must be given by people who have sufficient understanding of the relevant facts, and are able to consider these facts in a sufficiently clear-headed way. These conditions can be met by people who make some irrational decision. But the Rights Principle should not appeal to consent that is given by people who don't understand the most important relevant facts, or who are too young, or seriously mentally ill, or are affected by some other seriously distorting influence, such as being drunk, drugged, or threatened. Under such conditions, we can say, people cannot *validly* give or refuse consent.

When people cannot validly consent to some act, we might ask whether, *if* these people had been free from such distorting influences, they *would* have given such consent. But this question may be hard to answer. And there are other ways in which we could plausibly revise or extend the Rights Principle. Rather than appealing to the *hypothetical* consent that we believe that someone would have given at the time at which we act, we may be able appeal to this person's *actual* consent at some earlier time. In some cases, when people know that that they will later be affected by some distorting influence, they may validly give or refuse consent in advance to being later treated in some way. We may believe that we should later follow these earlier valid decisions. In some other cases, people cannot give valid consent at the time, and they

have neither given nor refused consent in advance. In such cases, we may believe that we ought to try to treat these people only in ways that they would later *retroactively endorse*, since they would later be glad that we acted as we did. Unlike the claim that people *would* have given valid consent, which could not be confirmed, many predictions of later endorsement could be either confirmed or shown to be false. That would provide a useful check on our use of such predictions to justify our acts.

We might next qualify the Rights Principle, so that it reflects the fact that the conditions for valid consent are matters of degree. When people are under some influence that to some extent distorts their judgment, though not so greatly as to make their decisions invalid, we may justifiably give these decisions less moral weight.

To illustrate some of these points, we can return to the view that everyone has a right not to have surgery performed on them without their consent at the time. This right is often claimed to be absolute, in the sense that it has no exceptions. But there are, I believe, some exceptions. Suppose that, in

> *Surgery*, to save *Green's* life, we must operate on her without anaesthetics. This operation would be very painful, but it would give Green many more years of worthwhile life. Green gives irreversible consent to this operation in advance, permitting us to use force, if necessary, if the pain later leads her to change her mind.

Before the discovery of anaesthetics, many people rationally gave such irreversible consent to life-saving surgery. If Green gave such consent, and the pain did later lead her to change her mind, we would be justified, I believe, in using force to complete this surgery. The Rights Principle should permit this act. We might however believe that, since great pain is a seriously distorting factor, Green's withdrawal of consent would not be valid.

Suppose next that, in a different version of this case, Green refuses to give such consent in advance. We may believe that this refusal is decisive, concluding that we ought to let Green die. But we might instead believe that Green's refusal should be regarded as invalid, or should be given less

weight, since the immediate prospect of great pain is another distorting factor, making it too difficult for people to make rational decisions. On one version of the Rights Principle, we could justifiably impose this surgery on Green if the pain of the surgery would be brief, and we also have strong reasons to believe that Green would later endorse our decision, being glad that we had saved her life despite her refusal of consent both at the time and in advance. We might know that, in such cases, most people endorse such surgery as soon as their worst pain is over.

In such cases, however, there is another, less obvious distorting factor. When we consider experiences that are painful, most of us have a strong *bias towards the future*. Once our pain is over, we care about it much less, or not at all. That makes it harder to justify imposing painful life-saving surgery by appealing to the fact that, after such surgery is over, almost everyone retroactively endorses such acts. Given our bias towards the future, we may underestimate the strength of the reasons that we earlier had to want to avoid what is now past pain.

Suppose next that, in

> *Depression*, Blue decides to kill herself. We have strong reasons to believe that, if we forcibly prevented Blue's act, Blue's depression would soon lift, and the rest of her life would go well.

Many of us would believe that we could justifiably override Blue's decision, and use force to prevent her from killing herself. If we accept the Rights Principle, we might claim that severe depression is a sufficiently distorting factor, so that Blue's refusal of consent is not valid. But if we made this claim, our standards of validity would be high, and would often fail to be met. People who are severely depressed may know the relevant facts, nor are they clearly incapable of making rational decisions. It would be more plausible to claim that, though Blue's depression does not make her refusal of consent invalid, it makes her less able to make rational decisions, so that Blue's refusal might be morally outweighed by her decisions at other times. For example, if Blue has frequent temporary depressions, she may have consented in advance to our later using force to prevent her from killing herself while she is depressed. That may be enough to justify our act, though

we would here be overruling Blue's *valid* refusal of consent at the time. And given the irreversibility of suicide, such acts might be justified even without such earlier consent. There is here an important asymmetry. If we frustrate Blue's attempt to kill herself, she could later try again, but if we allow her to kill herself, she could not later try to stay alive.

For an example of a different kind, suppose that, in

> *False Belief*, we could save Brown's life with a blood-transfusion. Brown refuses her consent, since she is a Jehovah's Witness who believes blood-tranfusions to be wrong.

For people to give valid consent, I have said, they must know the relevant facts. If Brown knew these facts, she would know that blood-transfusions are *not* wrong, and she could then have rationally consented to our saving her life in this way. But we might believe that, since Brown actually refuses her consent, it would be wrong for us to save her life in this way. When people refuse consent to some act because they have certain kinds of false belief, such as certain moral or religious beliefs, we can plausibly believe that this refusal should be regarded as valid.

In these remarks, I have assumed that present consent matters more than past consent, which matters more than retroactive endorsement. It is worth asking why these differences in timing have such significance.

If I cannot communicate with you, I might try to decide which of my possible acts would be most likely to fulfil your desires or preferences. As I have said, though our own preferences give us only derivative reasons, we can have non-derivative reasons to try to fulfil other people's preferences. In trying to do what would fulfil your preferences, I would have no reason to give priority to what you *now* prefer. Suppose that I have reasons to believe both that you would now want me to act in one of two ways, and that you would later change your mind, and would be glad if I had acted in the other way. I also have reasons to believe that, when you later changed your mind, you would know more of the relevant facts, so that your later preference would be better grounded. On these assumptions, I believe, I could rationally and justifiably give priority to fulfilling this later preference.

As one example of this kind, we can suppose that, as your doctor, I must decide whether to treat you in some way. Since you are unconscious, I cannot ask for your consent, and can only try to predict what you would prefer, and choose. This treatment would cause you some pain in the near future, but it would later save you from much greater pain. I have good reasons to believe that you would now prefer me not to treat you in this way, but that when you later learnt how bad that greater pain would be, you would change your mind. Given these facts, I could plausibly believe that I should fulfil your predictable later, better informed preference.

Suppose next that, in a different version of this case, you *are* conscious, so that I can ask for your consent to my proposed treatment. If you refuse consent, this fact might clearly morally outweigh my plausible prediction that you would later regret having made this decision. Though I have no reason to give your present *preferences* priority over your future preferences, I do have reason, when you are able to decide how I shall treat you, to give priority to what you now *decide*.

To explain this difference, we can first note a similar fact about our attitudes to our own and other people's beliefs. When I am trying to reach the truth about some question, and I take into account other people's beliefs, I would have no reason to give greater weight to other people's *present* beliefs. If I had some way of knowing what other people would later believe, I might have good reasons to give greater weight to these people's future beliefs, since these beliefs would be better grounded. I might also have good reasons to give greater weight to some of these people's past beliefs, which were freer from some distorting influence. I must, in contrast, give priority to *my present* beliefs. I can believe, for example, that some claim is false, though I did earlier believe, or shall later predictably believe, that this claim is true. But I cannot believe that some claim is false though I *now* believe that this claim is true. We can never base our decisions on the truth *rather* than on what *we now believe* to be true.

Similar claims apply to our decisions. We must give some priority to what we now decide, since these decisions are based on what we now believe to be true. And even when our beliefs have not changed, or we believe that they will not change, we must give priority to what we now decide, since we cannot make our decisions from some past or future

point of view. We have to live our lives from our own present point of view. These facts may explain why, when other people ought to act only with our consent, these people should also give priority to whether we *now* consent to their way of treating us.

28 Deontic Beliefs

The Consent Principle claims to describe only one of the ways in which our acts may be wrong. Acts may be wrong even though everyone could rationally consent to them.

Many such acts are wrong because some people do not, or would not, actually consent to them. That may be true, as I have said, of most kinds of direct interference with our bodies. Another much larger group of cases involve ownership. People do not always have a right to veto how we treat their property, since we could justifiably use or even destroy many kinds of property, despite the owner's refusal of consent, if that is our only way to save someone else from death or injury. But there are also many cases in which it would be wrong to use or destroy someone's property without this person's actual consent. If I do not have your consent, it may be wrong for me to live in your apartment, wear some of your clothes, and eat what is in your kitchen. In most cases, the Consent Principle would condemn such acts, since we could not have rationally consented in advance to other people's acting in such ways without our consent at the time. But if I had earlier been homeless, cold, and hungry, these facts might have given you sufficient reasons to consent in advance to my acting in these ways. The Consent Principle would not then condemn my acts. Despite this fact, it might be wrong for me to live in your apartment, wear your clothes, and eat what is in your kitchen, without your *actual* consent to these acts.

There might also be acts that are wrong even if everyone involved actually and rationally gives their valid consent. Many people have that view, for example, about *voluntary euthanasia*: killing someone, as this person asks us to do, for his or her own good. And some acts are wrong for reasons other than the ways in which they treat other people, so that the question of consent does not arise. That is true of cruelty to animals, for example, and some believe it to be true of suicide.

Since acts can be wrong in other ways, or for other reasons, what the Consent Principle implies may in part depend on which acts would be wrong for such other reasons. So when we apply this principle, we must sometimes appeal to our beliefs about which acts are wrong. These beliefs I shall call *deontic*, and the reasons that might be provided by some act's wrongness I shall call *deontic* reasons.

It might be objected that, if we apply the Consent Principle in a way that appeals to these beliefs, our moral reasoning would be circular, or question-begging. Such reasoning could not support our beliefs about which acts are wrong.

This objection is, in part, correct. It could not be true both that

> (M) some act would be wrong because someone could not rationally consent to it,

and that

> (N) this person could not rationally consent to this act because it would be wrong.

For some act to be wrong *because* someone could not rationally consent to it, this person must have decisive *non-deontic* reasons to refuse consent. But people often have such reasons. In *Earthquake*, for example, White has such a reason to refuse consent to my saving Grey's leg rather than White's life. White could not rationally consent to this act, not because it would be wrong, but because White's loss in dying would be so much greater than Grey's loss in losing a leg. When applied to such cases, and many other kinds of case, the Consent Principle supports and helps to justify some of our deontic beliefs.

As I have just said, however, we must sometimes apply the Consent Principle, in a way that appeals to our other deontic beliefs. Suppose that in a second version of *Earthquake*, which we can call

> *Means*, White and Grey are trapped in slowly collapsing wreckage. Though White's life is threatened, Grey is in no danger. I could save White's life, but only by using Grey's body as a shield, without Grey's consent, in some way that would destroy her leg.

Many of us would believe that, given Grey's refusal of consent, it would be wrong for me to save White's life in this way, by destroying Grey's leg. On this view, which we can here suppose to be true, it is wrong to act in any way that gravely injures someone, without this person's consent, as a means of benefiting someone else.

In applying the Consent Principle to this case, we can first set aside our assumption that this act would be wrong. If this act would not be wrong, this case would not be relevantly different from *Earthquake*. In both *Earthquake* and *Means*, either White will die or Grey will lose her leg. These cases would differ only in how the saving of White's life would be causally related to the loss of Grey's leg. Grey would have no strong reason to prefer to lose her leg in one of these ways. Neither, we can suppose, would be worse for her. In both cases, I believe, Grey could have rationally given her irreversible consent to my later saving White's life, even though Grey would then lose her leg. And in both cases, since White's loss would be so much greater than Grey's, White could not have rationally consented to my failing to save her life. On these assumptions, the Consent Principle would require me in *Means* to save White's life by destroying Grey's leg, since that is the only act to which both White and Grey could rationally consent.

Return now to our assumption that this act would be wrong. If the Consent Principle required this wrong act, that would be a strong objection to this principle. But this principle would not, I believe, require this act. If it would be wrong for me to save White's life by destroying Grey's leg, this act's wrongness would give White a sufficient reason to consent to my failing to act in this way. We all have sufficient reasons, I believe, to consent to someone's failing to benefit us, even when this benefit would be as great as the saving of our life, if this way of benefiting us would wrongly injure someone else.

Here is another way to defend this belief. We are discussing possible consent in the act-affecting sense. For White to be able to give or refuse such consent, we must suppose that I have given White the power to choose how I shall act. If White chose that I save her life by wrongly injuring Grey, she would be partly responsible for my wrong act. That would make it wrong for White to make this choice. And we always have sufficient reasons, I believe, not to make choices that would be morally

wrong. I am not claiming here that it would be irrational for White to make this choice. Perhaps White could rationally choose that I act wrongly, since this choice would save White's life, and that might give her sufficient reasons to make this choice. But White would also have sufficient reasons to choose instead not to be partly responsible for this wrong act. Since White could rationally consent to my failing to save her life by destroying Grey's leg, the Consent Principle would not mistakenly require this wrong act.

It might next be objected that, since *Grey* could rationally consent to my saving White's life in this way, the Consent Principle mistakenly permits this act even when, because Grey actually refuses consent, this act would be wrong. But this objection misunderstands the Consent Principle. This principle claims to describe only one of the facts that make acts wrong. So, when this principle does not condemn this way of saving White's life, it does not thereby *permit* this act, by implying that this act is morally permitted.

Similar remarks apply to other cases. We are discussing cases in which some act of ours would be wrong, not even in part because someone could not rationally consent to this act, but for other reasons. We can argue:

> The Consent Principle requires some act only when someone would not have sufficient reasons to consent to our failing to act in this way.
>
> (O) Whenever some act would be wrong for other reasons, this act's wrongness would give everyone a sufficient reason to consent to our failing to act in this way.

Therefore

> The Consent Principle could never require acts that are wrong for other reasons.

We can similarly argue that this principle could never condemn acts that are morally required for other reasons. If some act is required, all of

its alternatives would be wrong, and that would give everyone sufficient reasons to consent to this act.

On some views, premise (O) might be denied. Suppose that, in

> *Fire*, Black is trapped in burning wreckage, and will soon, if I do nothing, die a slow and painful death. I cannot save Black from this pain except by killing her now, before the increasing heat forces me to withdraw.

Suppose next that, knowing these facts, Black asks me to kill her. This act, I believe, would be morally justified. If that is true, Black could not rationally consent to my failing to benefit her, by giving her a swifter, painless death. On these assumptions, the Consent Principle requires me to kill Black, as she requests.

On one view, even in cases like *Fire*, such voluntary euthanasia is wrong. If it would be wrong for me to benefit Black by giving her this better death, would this act's wrongness give Black a sufficient reason to consent to my failing to act in this way? Some people might answer No. These people might agree that, in *Means*, White could rationally consent to my failing to save her life by destroying Grey's leg. But White's reason to give such consent is provided by the fact that I could save White's life only by wrongly injuring someone else. No such claim applies to *Fire*. If I killed Black at her request, I would not be wrongly injuring anyone else. These people might believe that, given this difference, the wrongness of my killing Black would *not* give Black a sufficient reason to consent to my failing to benefit her in this way. On these assumptions, premise (O) would here be false, and the Consent Principle would require an act that would be wrong.

This example does not, I believe, provide a strong objection to the Consent Principle. Few people would believe both that this act would be wrong and that its wrongness would not give Black a sufficient reason to consent to my failing to act in this way. And we could plausibly reject this view.

Consider next a different version of *Fire*. Suppose that, though Black knows that my killing her would be better for her, she refuses her consent. Some people might believe both that this act would be wrong without

Black's consent, and that Black could not have rationally consented in advance to my failing to give her, without her later consent, this swifter, better death. If these beliefs were both true, premise (O) would be false, since the Consent Principle would here require me to act wrongly. But I believe that, if it would be wrong for me to kill Black without her actual consent at the time, this act's wrongness would have similarly given Black sufficient reasons to consent in advance to my failing to act in this way.

For an example of a different kind, suppose that, in

> *Parents*, after some shipwreck, you and I each have a child whose life is in danger. I have a life-belt, which I could use to save either my child or yours.

Suppose next that, as most of us would believe, I ought to save my child. Could you rationally consent to my acting in this way?

On one view, the answer is No. If I gave you the power to choose how I would act, you ought to choose that I act wrongly, by saving your child. Though you would be partly responsible for my wrong act, your duty to protect your child would morally outweigh your reason not to choose that I act wrongly. Given this fact, and your other strong reasons to want me to save your child, you could not rationally consent to my failing to act in this way. On these assumptions, the Consent Principle would here require me to act wrongly, by saving your child rather than mine.

If we accept this view, and we have similar beliefs about other relevantly similar cases, we would have to revise the Consent Principle, so that it did not apply to this kind of case. According to

> CP4: It is wrong to treat people in any way to which they would not have sufficient reasons to consent, except when these people would not have such reasons because the case involves conflicting person-relative moral obligations.

Though this revision would restrict the scope of the Consent Principle, it would not make this principle less plausible. When we apply this principle, we appeal to a thought-experiment, by asking whether other people could rationally choose that we act in some way. We cannot usefully ask this question when it makes a moral difference whether it is we

or someone else who chooses how we shall act. In such cases, it might be wrong for us to do what it would be right for someone else to choose that we do. Our thought-experiment would here lead us to ignore this fact. We should not expect that, in such cases, the Consent Principle could help us to decide which acts are wrong. Since we can give this explanation of *why* this principle should not be applied to cases of this kind, such cases would not cast doubt on the moral idea that this principle expresses.

This revision may not, however, be needed. We can ask

> Q1: Could we have a duty to choose, or bring it about, that someone else acts wrongly?

On some moral views, the answer is sometimes Yes. One such view is the kind of moral nationalism that was widely accepted in Europe before and during the First World War. On this view, if your nation is at war with mine, it might be my patriotic duty to try to get you to act wrongly, by unpatriotically giving me the information with which my nation's army can defeat yours.

Kant's answer to Q1 would be No. And if we are right to accept this answer, *Parents* does not undermine the Kantian ideal. On such a view, we can have what are in one sense conflicting personal-relative obligations. It might be my duty to save my child, and your duty to save yours, though my doing my duty would make it impossible for you to do yours. But in *Parents* I could act in a way to which you could rationally consent. Since it would be wrong for me to save your child rather than mine, you could not have a duty to choose that I act in this way, and this act's wrongness would give you a sufficient reason to consent to my doing my duty, by saving my child.

For a different objection, suppose next that, in

> *Equal Claims*, I could save either your life or Grey's.

It may seem that, in this case, you could not rationally consent to my saving Grey's life rather than yours. You would have strong personal reasons not to give such consent. And since your death would be impartially as bad as Grey's, these personal reasons may seem to be decisive. Grey would have similar reasons not to consent to my saving

your life rather than Grey's. The Consent Principle may seem here to fail, by mistakenly implying that, whatever I do, I shall be acting wrongly, since I shall be treating someone in some way to which this person could not rationally consent. We can plausibly claim, however, that I ought to give both you and Grey an equal chance of being saved. And if it would be wrong for me not to give you both an equal chance, this fact would give you both sufficient reasons to consent to this act.

For another example, suppose that, in

> *High Price*, I sell you some product that, as only I know, you could have bought much more cheaply elsewhere.

Suppose next that, since you are not rich, you could not have rationally chosen to pay this higher price. The Consent Principle then implies that, in taking your money, I act wrongly. Some of us would believe that, since you freely consent to my taking your money, I do not act wrongly. But the Consent Principle is not obviously mistaken here. We could plausibly believe that, just as I ought to warn you if the product that I am selling is in some way defective, I ought to tell you that you could buy this product much more cheaply elsewhere.

My remarks about these cases do not prove that we could always justifiably follow the Consent Principle, thereby achieving Kant's ideal. Some people would reject these claims. And there might be other kinds of case in which there would be no possible act to which everyone could rationally consent. But, as these various cases show, the Consent Principle has implications that are often plausible, and never obviously mistaken. That makes it worth asking, of the most plausible views about both morality and rationality, which views are compatible with Kant's ideal.

29 Extreme Demands

There is, however, another objection to this principle. Suppose that, in

> *Self*, I am trapped with White in slowly collapsing wreckage. I could save either White's life or my leg.

On some views, this case is morally just like *Earthquake*. I ought to save White's life rather than my leg, since White's loss would be much greater than mine. Most of us would have a different view. On this view, though it would be wrong for me to save some other stranger's leg rather than White's life, I would be morally permitted to save *my* leg. We ought to save any stranger's life when that would cost us little. But the cost to me here would be too great.

What does the Consent Principle here imply? If White could choose how I would act, could White rationally choose that I save my leg rather than her life?

The answer may seem to be No. It may seem that White could not rationally consent to anyone's saving anyone's leg rather than White's life. But this view is too simple. We can have reasons to care, not only about *what* will be done, but also about *who* will be doing these things, and *why* they will be doing them.

To illustrate this point, it will help to lower the stakes. Suppose first that I could either save White from a week of pain, or save some other stranger from only one day of similar pain. There is no other relevant difference between White and this other stranger. On that assumption, I would have no reason to give less weight to White's well-being. And White could not rationally consent to my choosing, *for no reason*, to help the other stranger rather than saving White from her much greater burden. That choice would treat White as if she were inferior, or didn't even exist.

Suppose next that, rather than saving White from her week of pain, I could save myself from one day of pain. Though I would have no reason to care more about the well-being of one of two strangers, I do have reasons to care more about my own well-being. We all have reasons to be specially concerned about what happens to ourselves. Since everyone has such reasons, we could often rationally consent to other people's giving priority, for these reasons, to their own well-being. Though White could not rationally consent to my choosing, for no reason, to save some other stranger from a day of pain rather than saving White from her week of pain, White may have sufficient reasons to consent to my saving *myself* from this much smaller burden. *This* act would not treat White as if she were inferior, or didn't even exist.

In *Self*, however, the stakes are much higher. White may not have sufficient reasons to consent to my saving my leg rather than White's life.

Would it make a difference if, as most of us would believe, I would be morally permitted to save my leg rather than White's life? Perhaps not. There may be a difference here between permissibility and wrongness. If I could save White's life only by acting wrongly, as we have supposed to be true in *Means*, this act's wrongness, I have claimed, would give White a sufficient reason to consent to my failing to save her life. In *Self*, however, I could save White's life without acting wrongly. And even if I would be morally permitted to save my leg rather than White's life, this act's permissibility may not give White a sufficient reason to consent to my failing to save her life.

If this act's permissibility would *not* give White such a reason, White could not rationally consent to my failing to save her life, so the Consent Principle would require me to save White's life rather than my leg. This principle would here conflict with what most of us believe.

Though few people could save someone else's life only at the cost of a serious injury to themselves, there are many cases to which similar reasoning applies. Many of us could often either benefit ourselves or give some greater benefit to others. When the benefits to other people would be *much* greater, these people may not have sufficient reasons to consent to our failing to benefit them. Suppose that, in

> *Aid Agency,* I could either spend $200 on some evening's entertainment, or give this money to some efficient aid agency, such as Oxfam, which would use this money to save some poor person in a distant land from death, blindness, or some other great harm.

When applied to these two alternatives, the Consent Principle seems to imply that I ought to give this money to this aid agency. This poor person seems not to have sufficient reasons to consent to my failing to act in this way. Similar claims will apply to me tomorrow, and on every other day. And similar claims apply, on every day, to most readers of this book. Compared with the more than a billion people who now live on around $2 a day, most readers of this book are *very* rich.

It would be no objection to the Consent Principle if, for these reasons, this principle requires the rich to transfer much of their wealth or income to the poor. Now that the rich could so easily save so many of the poor from death or suffering, any plausible principle or moral theory makes similarly strong demands. And though the rich are legally entitled to all their property, they may be morally entitled to much less than that. Kant writes:

> Having the resources to practice such beneficence as depends on the goods of fortune is, for the most part, a result of certain human beings being favoured through . . . injustice.

And he is reported to have said:

> one can participate in the general injustice, even if one does no injustice . . . even acts of generosity are acts of duty and indebtedness, which arise from the rights of others.

The Consent Principle may, however, be *too* demanding. After thinking seriously about what justice requires, and considering the relevant arguments, we may have to admit that we rich people ought all to transfer to the poor a tenth of our wealth or income, or a fifth, or a third. But the Consent Principle might require much more than that.

If this principle is too demanding, it could be revised. We might claim

> CP5: It is wrong for us to treat people in any way to which they would not have sufficient reasons to consent, except when, to avoid such an act, we would have to bear too great a burden.

In applying this version of the Consent Principle, we would have to decide when such burdens would be too great. When we consider the moral problems raised by extreme global inequality, that is a very difficult question. One problem is whether and how we should assess the cumulative costs of many small gifts. But we could start by claiming that, in *Self*, I would be permitted to save my leg rather than White's life.

If the Consent Principle is too demanding, and must be weakened in this way, Kant's ideal of interpersonal relations may seem to be in principle impossible, since there would be some right acts to which

some people could not rationally consent. But these acts would be right only in the sense that they would be morally permitted. There might be no morally *required* acts to which some people could not rationally consent. So we might still be able to achieve Kant's ideal. It might still be possible for everyone to act only in ways to which everyone could rationally consent. And there might always be at least one such act that would be right. In *Self*, for example, I could save White's life rather than my leg, and this admirable act would be right. If the Consent Principle is too demanding, this would at most imply that, to achieve Kant's ideal, we would have to do more for each other than we are morally required to do. That would not be surprising.

We have, I conclude, strong reasons to accept some version of the Consent Principle. This principle may be too demanding, and there may be some other ways in which it should be revised. But at least in most cases, it is wrong to act in ways to which anyone could not rationally consent. When our acts would affect many people, and there is only one possible act to which everyone could rationally consent, this fact gives us a strong reason to act in this way, and may be enough to explain why such acts are morally required. And on some plausible assumptions, the Consent Principle could never go astray, by requiring acts that are wrong for other reasons, or condemning acts that are required.

The Consent Principle cannot, however, be what Kant was trying to find: the supreme principle of morality. Some acts are wrong even though everyone could rationally consent to them. The Consent Principle states one of the ideas that are expressed in Kant's Formula of Humanity. Since we need at least one other principle, we can now turn to another part this formula.

9

Merely as a Means

30 The Mere Means Principle

Using people, it is often claimed, is wrong. But this claim needs to be qualified. If we are climbing together, I might use you as a ladder, by standing on your shoulders. And I might use you as a dictionary, by asking you what some word means, or use you as a witness to my signing of my will. Such ways of using people are not wrong. What is wrong, Kant claims, is *merely* using people. As others say, 'You were just using me'.

According to what we can call

> *the Mere Means Principle*: It is wrong to treat anyone merely as a means.

How can we use people without *merely* using them? In explaining this distinction, we can first compare how two scientists might treat the animals in their laboratories. One scientist, we can suppose, does her experiments in the ways that are most effective, regardless of the pain she causes her animals. This scientist treats her animals merely as a means. Another scientist does her experiments only in ways that cause her animals no pain, though she knows these methods to be less effective. This scientist, like the first, treats her animals as a means. But she does not treat them *merely* as a means, since her use of them is restricted by her concern for their well-being.

Similar claims apply to our treatment of each other. According to one rough definition,

> we treat someone *as a means* when we make any use of this person's abilities, activities, or body to help us to achieve some aim.

This definition needs to be qualified in certain ways. We should sometimes distinguish, for example, between doing something to someone as a means of achieving some aim, and treating this *person* as a means. Suppose that, to find out whether I have a broken rib, my doctor presses all over my chest, saying 'Tell me where it hurts'. My doctor is using my body, and hurting me, as a means of getting this information, but she isn't treating *me* as a means. To cover such cases, we might suggest that we do not treat someone as a means when our aim is to benefit this person, and we act with this person's consent.

According to another rough definition,

> we treat someone *merely* as a means if we both treat this person as a means, and regard this person as a mere instrument or tool: someone whose well-being and moral claims we ignore, and whom we would treat in whatever ways would best achieve our aims.

Kamm rejects this second definition. She objects that, if this were the sense in which, on Kant's principle, we must never treat people merely as a means, this principle would be too weak, and too easy to follow. On this definition, for example, if some slave-owner gave even slight weight to the well-being of his slaves, by letting them rest in the hottest part of the day, he would not be treating his slaves merely as a means. But this man surely treated his slaves in ways that Kant's principle condemns.

This objection shows, I believe, not that we ought to revise this definition, but that we ought to revise Kant's principle. For a similar example, consider Kant's claim that

> (A) it is wrong for the rich to give nothing to the poor.

Suppose that some rich man gives to the poor, in his whole life, a total of one dollar and 3 cents. Since this man gives something to the poor, (A) does not imply that he acts wrongly. As this example shows, (A) is

too weak, since this man's failure to give more is wrong. The rich act wrongly, we should claim, if they give *too little* to the poor. This kind of wrongness is a matter of degree.

So is the wrongness, we might claim, of treating people merely as a means. On a stronger form of Kant's principle, which we can call

> *the Second Mere Means Principle*: It is wrong to treat anyone merely as a means, or to come close to doing that.

We *come close* to treating someone merely as a means when we both treat this person as a means and give too little weight to this person's well-being or moral claims. That is how my imagined slave-owner treated his slaves, even though he let them rest in the hottest part of the day. So this revised principle condemns this man's acts.

We can next claim that

> (B) we do *not* treat someone merely as a means, nor are we even close to doing that, if either
>
> > (1) our treatment of this person is governed or guided in sufficiently important ways by some relevant moral belief or concern,
>
> or
>
> > (2) we do or would relevantly choose to bear some great burden for this person's sake.

For some moral belief to be *relevant* in the sense intended in (1), this belief must require direct concern for the well-being or moral claims of the person whom we are treating in some way. Suppose that some other slave-owner never whips his slaves because he believes that such acts would be wrong. But what would make such acts wrong, he believes, is not the fact that he would be inflicting pain on his slaves, but the fact that he would be giving himself sadistic pleasure. If that is why this man never whips his slaves, this fact would not count against the charge that he treats his slaves merely as a means. Another example is Kant's view that cruelty to animals is wrong because it dulls our sympathy, making us more likely to be cruel to other people. If it is

only this moral belief that leads some scientist to avoid causing her laboratory animals any pain, she would be treating these animals merely as a means.

Since relevance and importance are both matters of degree, it is often unclear whether (1) is true. Some other slave-owner might refrain from whipping his slaves because he cares about their well-being. But this concern, though relevant, would not govern this man's acts in a sufficiently important way. In a case that is less clear, when my mother travelled on a Chinese river in the 1930s, her boat was held up by bandits, whose moral principles permitted them to take, from ordinary people, only half their property. These bandits let my mother choose whether they would take her engagement ring or her wedding ring. If these people treated my mother as a means, they did not treat her *merely* as a means. Were they *close* to doing that? I am inclined to answer No. But this may be a borderline case, in which this question has no definite answer.

For condition (2) to be met, it is not enough that we would be prepared to bear some great burden for someone's sake. This fact may not be sufficiently relevant to the acts that we are considering. Consider some man who loves his wife, and who, in some disaster, would give up his life to save hers. It may still be true that, in much of this man's ordinary domestic life, he treats his wife merely as a means.

Whether we are treating someone *as a means* depends only on what we are intentionally doing. Whether we are treating someone *merely* as a means depends also, I believe, on our underlying attitudes or policies. And that is in part a matter of what we would have done, if the facts had been different. Return to our scientists who both use laboratory animals in their research. Suppose that, in one experiment, both these scientists use the most effective method, which causes their animals no pain. Though these scientists are acting in the same way, the first scientist would still be treating her animals merely as a means, since it would still be true that she *would* have used the most effective method even if that would have caused her animals great pain. And the second scientist would *not* be treating her animals merely as a means, because she would not have acted in that other way. Consider next these claims:

> He treats her merely as a means.
>
> On this occasion, in acting as he did, he treated her merely as a means.

The first claim is more natural, and it is often clearer whether such claims are true.

It is wrong, Kant claims, to treat any rational being merely as a means. On a similar but wider view, it is wrong to treat any sentient or conscious being merely as a means. These views rightly imply that it is wrong to *regard* any rational or sentient being as a mere tool, whom or which we could treat as we please. But Kant's claim seems also to imply that, in treating anyone merely as a means, we would be *acting* wrongly.

That may not be true. Consider some gangster who, unlike my mother's principled bandits, regards most other people as a mere means, and who would injure them whenever that would benefit him. When this man buys a cup of coffee, he treats the coffee seller just as he would treat a vending machine. He would steal from the coffee seller if that was worth the trouble, just as he would smash the machine. But though this gangster treats the coffee seller merely as a means, what is wrong is only his attitude to this person. In buying his cup of coffee, he does not act wrongly.

Consider next some Egoist, who treats others in whatever way he believes would be best for him. Kant claims

> he who intends to make a lying promise . . . wants to make use of another human being merely as a means.

We could similarly claim that, when this Egoist *keeps* some promise to someone whose help he will later need, he wants to make use of this other human being, and treats him merely as a means. Suppose next that this Egoist saves some child from drowning, at a great risk to himself, but that his only aim is to be rewarded. Since this man treats these other people merely as a means, Kant's principle implies that, in keeping his promise and saving this child's life, this man acts wrongly. That is clearly false.

To avoid such conclusions, we might claim that

> (3) we do not treat someone merely as a means if, as we know, our acts will not harm this person.

But suppose that, in

> *Mutual Benefit*, *Green* marries *Gold*, a 90-year old billionaire, to whom Green gives various services, and in other ways treats well. Green's sole aim, as Gold knows, is to inherit some of Gold's wealth. Though Gold would prefer genuine affection from Green, he accepts a mutually advantageous arrangement on Green's egoistic terms.

Suppose next that Green regards Gold as a mere tool, whom she would treat in whatever way would best achieve her aims. Green's first plan was to forge Gold's will and then murder him, and she changed her plan to marrying Gold, and treating him well, only because that seemed a safer way to get some of Gold's wealth. According to (3), since Green knows that her acts will not harm Gold, she is not treating Gold merely as a means. That claim is implausible. Though Green knows that her acts will not harm Gold, this fact makes no difference to her decisions. She would have murdered Gold if that had seemed a safer plan. We should admit, I believe, that Green treats Gold merely as a means.

If we cannot appeal to (3), Kant's view implies that Green acts wrongly. Perhaps we should accept that conclusion. But when my Egoist keeps his promises, or risks his life to save some drowning child, we should not claim that these acts are wrong. Our claim should be only that, given this man's self-interested motives, his acts do not have what Kant calls *moral worth*.

To avoid condemning such acts, we might again revise Kant's view. According to

> *the Third Mere Means Principle*: It is wrong to treat anyone merely as a means, or to come close to doing that, if our act will also be likely to harm this person.

In moving to this principle, we would be giving up the view that, if we treat someone merely as a means, or we are close to doing that, these facts are enough to make our act wrong.

I have discussed two ways in which, on Kant's view, we ought to treat all rational beings, or persons. We ought to follow the Consent Principle, by treating everyone only in ways to which they could rationally consent. And it is wrong to treat anyone merely as as a means. On our latest version of this second claim, such acts are wrong only if they are also likely to harm this person.

We can next connect these parts of Kant's view. We do not treat someone merely as a means, nor are we even close to doing that, if our treatment of this person is governed or guided in sufficiently important ways by some relevant moral belief or principle. Kant's own example is the Consent Principle. We treat people as ends, Kant claims, and not merely as a means, if we deliberately treat these people only in ways to which they could rationally consent.

Return now to

> *Lifeboat*: I am stranded on one rock, and five people are stranded on another. Before the rising tide drowns all of us, you could use a lifeboat to save either me or the five.

Consider also

> *Tunnel*: A driverless, runaway train is headed for a tunnel, in which it would kill the same five people. As a bystander, you could save these people's lives by switching the points on the track, thereby redirecting this train on to another track and through another tunnel. Unfortunately, as you know, I am in this other tunnel.

> *Bridge*: The train is headed for the five, but there is no other track and tunnel. I am on a bridge above the track. Your only way to save the five would be to open, by remote control, the trap-door on which I am standing, so that I would fall in front of the train, thereby triggering its automatic brake.

In all three cases, if you save the five, I would die. But my death would be differently causally related to your saving of the five. In *Lifeboat*, you would let me die because, in the time available, you could not save both me and the five. In *Tunnel*, you would save the five by redirecting the train with the foreseen side-effect of thereby killing me. In *Bridge*, you would kill me as a means of saving the five. I and the five, we should suppose, are all of about the same age, none of us is responsible for the threats to our lives, nor are there any other morally relevant differences between us.

It might be claimed that, in *Bridge*, you would not really be *killing* me as a means of saving the five. You would be merely using my body as a means of stopping the train, and you would be delighted if I survived. On this view, we kill someone as a means only when this person's death is an essential part of what achieves our aim. That might have been true, for example, of some medieval king's second son, who wanted to be the legitimate or rightful heir to his father's throne. Only his elder brother's death would achieve that aim. In a wider sense, however, we kill or injure someone as a means when we act in some way that involves and foreseeably kills or injures this person, as a means of achieving some aim. That is how I shall use the phrase 'kill or injure as a means'.

Most people would believe that, in *Lifeboat*, you either may or ought to save the five. Some people would believe that, in both *Tunnel* and *Bridge*, it would be wrong for you to save the five. On this view, we have a duty not to kill which outweighs, or has priority over, our duty to save people's lives. Many other people would believe that, though our duty not to kill usually has such priority, that is not true in cases like *Tunnel*. On these people's view, it is not wrong to redirect some unintended threatening process—such as some flood, avalanche, or runaway train—so that it kills fewer people. Of those who hold this view, most would believe that you *would* be acting wrongly if, in *Bridge*, you killed me as a *means* of stopping the train and saving the five. There are also some people who reject these distinctions, believing that in all these kinds of case we ought to save as many lives as possible. My aim here is not to resolve this disagreement, but only to ask what is implied by the Kantian principles that we have been considering.

In *Lifeboat*, I have claimed, I could rationally consent to your saving the five rather than me. If the choice were mine, I would have sufficient reasons to save my own life, but I would also have sufficient reasons to save the five rather than myself. Since I could also rationally consent to your saving the five, the Consent Principle would not condemn this act.

Similar claims apply to *Tunnel*. As before, if the choice were mine, I would have sufficient reasons to save either myself or the five. It would make no relevant difference that I would here be saving the five by redirecting the train so that it would kill me instead. This way of dying, we can suppose, would be no worse for me. Since I could rationally save the five by redirecting the train, I could also rationally consent to *your* acting in this way. So the Consent Principle would not condemn this act.

Similar claims apply to *Bridge*, in which you could save the five only by killing me. If the choice were mine, I would have sufficient reasons to jump in front of the train, so that it would kill me rather than the five. And compared to killing myself as a side-effect of saving the five, in *Tunnel*, it would be no worse for me, in *Bridge*, if I killed myself as a means of saving the five. Since I could rationally kill myself as a means of saving the five, I could also rationally consent to your treating me in this way.

It might be objected that I could not rationally consent to your killing me as a means, because this act would be wrong. But if I consented to this act, it would not be wrong. So even if this act would be wrong without my consent, that would not give me any reason to refuse consent.

Suppose next that, as I know, you accept the Consent Principle, and you always act upon it, so that this principle governs your acts. If I had the time, I might then think:

> According to this principle, it is wrong to treat people in any way to which they could not rationally consent.
>
> I could rationally consent to your killing me as a means of saving the five.

Therefore

> Even if I would not in fact consent, the Consent Principle would not condemn this act.

> We do not treat people merely as a means if our treatment of them is governed by the Consent Principle.

Therefore

> Since your treatment of me would be governed by the Consent Principle, you would neither be treating me merely as a means, nor be close to doing that, so no version of the Mere Means Principle would condemn your act.

This argument, I believe, is sound. It might be wrong for you to kill me, without my consent, as a means of saving the five. But that is not implied by these Kantian principles.

31 As a Means and Merely as a Means

It may seem that, in making these claims, I must be misunderstanding or misapplying the Mere Means Principle. On one widely accepted view, which I shall call

> *the Standard View*, if we harm people, without their consent, as a means of achieving some aim, we thereby treat these people merely as a means, in a way that makes our act wrong.

This view involves, I believe, three mistakes. When we harm people as a means, we may not be treating these *people* as a means. Even if we *are* treating these people as a means, we may not be treating them *merely* as a means. And even if we *are* treating them merely as a means, we may not be acting wrongly.

Suppose first that, in

> *Self-Defence*, when Brown attacks me with a knife, trying to kill me, I save myself by kicking Brown in a way that predictably breaks his leg.

Though I am *harming* Brown as a means of stopping him from killing me, I am not treating *Brown* as a means. Just as we do not *use* falling rain when we wear raincoats to protect ourselves from being drenched,

we do not *use* the people who attack us when we protect ourselves from their attack. We can add that, though I ought to treat *Brown himself* as an end and not merely as a means, I ought to *harm* Brown *merely* as a means and not even in part as an end, or for the sake of harming Brown.

It might be objected that, since harming someone is a way of treating this person, harming someone as a means must be a way of treating this person as a means. But this objection overlooks the difference between *doing* something to someone as a means and using *this person*. As I have said, when my doctor hurts me to find out whether my rib is broken, she isn't thereby using *me*. She isn't treating *me* as a means, I suggested, because she is hurting me for my own good and with my consent. Though I might be benefiting Brown by preventing him from committing murder, that is not the best way to explain why, in harming Brown as a means, I would not be using Brown. We might instead suggest that, since I am merely protecting myself from Brown's attack, my aims would be more easily achieved if Brown wasn't even there. If I was using Brown, I *would* want him to be there.

Turn next to the cases in which, when we harm people as a means, we *do* also treat these *people* as a means. On the Standard View, if we impose harm on someone as a means of achieving some aim, that is enough to make it true that we are treating this person *merely* as a means. To test this view, consider

> *Third Earthquake*: You and your child are trapped in slowly collapsing wreckage, which threatens both your lives. You cannot save your child's life except by using *Black's* body as a shield, without her consent, in a way that would crush one of her toes. If you also caused Black to lose another toe, you would save your own life.

Suppose you believe that it would be wrong for you to save your life in this way. Only the saving of a child's life, you believe, could justify imposing such an injury on someone else. Acting on this belief, you save your child's life by causing Black to lose only one toe. Since your act harms Black, without her consent, as a means of achieving your aim, the Standard View implies that you are treating Black merely as a means.

But that is not true. If you were treating Black merely as a means, you would save your own life as well as your child's, by causing Black to lose two toes. We are not treating someone merely as a means if we are letting ourselves die rather than imposing a small injury on this person.

The Standard View might be revised. It might be suggested that, though you are not treating Black merely as a means, that is because you are limiting the harm that you impose on Black, in a way that is worse for you, or less effectively achieves your aims. No such claim would apply to your act, in *Bridge*, if you killed me as a means of saving the five. You would not be limiting the harm that you imposed on me. And you would have acted in the very same way even if you had regarded me as a mere means. That may seem enough to justify the charge that, in acting in this way in *Bridge*, you would be treating me merely as a means. On this suggestion,

(C) we treat someone merely as a means if

(1) we harm this person, without his or her consent, as a means of achieving some aim,

unless

(2) we limit the harm that we impose, in some way that would or might be significantly worse for us, or make our act significantly less effective in achieving our aims.

This view is also, I believe, mistaken. We have supposed that, in *Third Earthquake*, you decide not to save your life by causing Black to lose a second toe. Suppose next that, just before you act, the situation changes, since the collapsing wreckage now threatens only your child's life. When you save your child's life by causing Black to lose one toe, you are not now limiting the harm that you impose on Black, so (C) implies that you are treating Black merely as a means. That is an indefensible conclusion. Rather than causing Black to lose a second toe, you would have let yourself die. That is enough to make it true that you are not treating Black merely as a means. It is irrelevant that you cannot now act in this way.

For another example, suppose that I am a soldier in some just war, fighting my way with my platoon through some occupied city. Before attacking the enemy soldiers in any building, I risk my death from

sniper fire so that I can shout to these people, giving them a chance to surrender. If these people refuse my offer, and I kill or injure them as a means of capturing some building, (C) rightly allows that I am not treating these people merely as a means, since I have risked my life for their sake. Suppose next that the enemy soldiers in some building have already been given a chance to surrender, and have refused this offer. According to (C), if I kill or injure these people, I am treating them merely as a means. That is not true. I would have risked my life to give these people a chance to surrender. It is irrelevant that, on this occasion, I do not act in this way, because these people have already been given this chance. My attitude to all enemy soldiers is the same, and I treat none of them merely as a means.

Similar claims apply to *Bridge*. Suppose that you use remote control to cause me to fall onto the track, so that my body would stop the runaway train. Your aim is to ensure that the five will be saved. You also try, however, to save my life by running to the track, so that you can jump in front of the train, thereby stopping it before it reaches me. If your attempt succeeds, you would not be treating me merely as a means, since you would be killing yourself for my sake. It would make no relevant difference, I believe, if you failed to reach the track in time. Nor would it make such a difference if, though you would have sacrificed your life to avoid killing me, this was never possible. In both versions of *Bridge*, your act may be wrong. And if it is, what makes it wrong may be the fact that you would be *killing* me *as a means* of saving the five. But you would not be treating *me* as a *mere* means.

I have rejected the standard account of what is involved in treating people as a mere means. Some writers give other accounts. For example, O'Neill writes:

> if we coerce or deceive others . . . we do indeed use others, treating them as mere props or tools in our own projects . . . a maxim of deception or coercion treats another as mere means . . .

Korsgaard similarly writes:

> Coercion and deception are the two ways of using others as mere means.

But suppose that, in a variant of *Self-Defence*, I stop Brown from killing me by threatening to shoot him, or by falsely telling him that the police will soon arrive. Though I would be coercing or deceiving Brown, I may not be treating Brown as a mere means. I may be coercing or deceiving Brown because these are the only ways in which, without harming Brown, I could stop him from killing me. Suppose next that, in

> *Desperate Plight*, you and I are in some diving bell which is caught on the ocean's floor. Though we cannot hope to be rescued in less than ten hours, we have enough oxygen to keep two people alive for only six or seven hours. So, as I know, unless one of us dies soon, we shall both die. I start acting in some way that will kill me and thereby save your life. When you try to stop me, I coerce you or deceive you so that your attempt fails.

Though I am coercing or deceiving you, I am not treating you as a mere means. As before, we are not treating someone as a mere means if we are sacrificing our life for this person's sake.

When O'Neill explains her claim that deception and coercion treat others as a mere means, she writes

> To treat something as a mere means is to treat it in ways that are appropriate to things.

Deception and coercion are not, however, appropriate ways of treating things, since neither is even possible.

On Kant's view, Korsgaard also writes,

> Any attempt to control the actions and reactions of another by any means except an appeal to reason treats her as a mere means . . .

This claim implies that whenever people in positions of authority tell us to do something—such as to show them our train ticket, or fill out a customs declaration, or fasten our safety-belts—they are treating us as a mere means. That is not true. Korsgaard also writes that, on Kant's

view, we treat others as a mere means whenever 'we do something that only works because most other people don't do it'. But when poor people feed themselves with the scraps that others throw away, they do not treat these other people as a mere means.

Suppose next that, in

> *Bad Samaritan*, while driving across some desert, I see you lying injured by the road, needing help. I ignore you, and drive on.

According to some writers, Kant would claim that I am here treating you merely as a means. That claim would be false. In ignoring you, I am not using you in any way, so I cannot be merely using you.

These writers might reply that, when Kant uses the phrase 'merely as a means'—or, more accurately, its German equivalent—Kant does not use this phrase in its ordinary sense. Kant often uses words in special senses. When I drive past you, ignoring your need for help, it might be true that, in Kant's special intended sense, I am treating you merely as a means. O'Neill and Korsgaard might similarly claim that all deception and coercion does, in Kant's special sense, treat people merely as a means.

We are sometimes justified in using words in something other than their ordinary senses. For example, it can be worth stretching the sense of 'painful', so that it applies to unpleasant sensations, such as nausea. By using 'painful' in this wider sense, we avoid the need to keep writing 'painful or unpleasant', and the distinction that we are ignoring seldom matters. Some unpleasant sensations are much worse to have than some pains. It is often a mistake, however, to use words in special senses. We may then make claims that are misleading and only seem to be important. For example, Rawls suggests that, if we accept his Contractualist moral theory, we should use 'right' to mean: in accordance with the principles that would be chosen by his imagined contractors. That would make it trivial to claim that acting in accordance with these principles is right. Rawls also suggests that we could call these principles 'true' in the sense that they would be chosen

by these contractors. That would make it trivial to claim that these chosen principles are true.

If we believe that Kant uses 'merely as a means' in some special sense, we ought not to say that, on Kant's view, we must never treat people merely as a means. If that is what we say, our hearers may take us to be claiming that, on Kant's view, we must never treat people merely as a means. To avoid being misunderstood, we should use some other phrase. We might say that, on Kant's view, we must never treat people in certain ways, which we shall call treating people *shmerely as a means*. We could then explain what we use this new phrase to mean.

The phrase 'merely as a means' has, I believe, an ordinary sense that is both fairly clear, and morally significant. Though Kant may sometimes use this phrase in a special sense, he also uses it, I believe, in the ordinary sense. It is not misleading to say that, according to Kant's Formula of Humanity, we must never treat people merely as a means. And this is the version of Kant's formula that is most worth discussing.

On my rough definition of this ordinary sense, we treat someone merely as a means if we both use this person in some way and regard her as a mere tool, someone whose well-being and moral claims we ignore, and whom we would treat in whatever way would best achieve our aims. We do *not* treat someone merely as a means, nor are we even close to doing that, if either (1) our treatment of this person is governed in a sufficiently important way by some relevant moral belief, or (2) we do or would relevantly choose to bear some great burden for this person's sake.

When people give other definitions, they are often trying to make Kant's claim cover a wider range of acts. That can sometimes be done, I have suggested, not by using 'merely as a means' in some special sense, but by revising Kant's claim so that it also condemns acts that are *close* to treating people merely as a means. And rather than stretching Kant's claim so that it covers other kinds of act, we should sometimes make other, similar claims. When Bad Samaritans ignore someone who needs urgent help, they do not treat this person as a mere means. But they do treat this person as a *mere thing*, something that has no importance, like a stone or heap of rags lying by the road. That, we could claim,

is just as bad. And there are ways of treating people that are worse than treating them as a mere means. Though Hitler treated the Slavs in his conquered Eastern territories as a mere means, that is not how he treated the Jews.

32 Harming as a Means

We can now return to the question of whether, as Kant claims, it is wrong not only to *regard* people merely as means, but also to *act* in ways that treat them merely as a means.

Kant's claim, as we have seen, is too strong. When my gangster buys his cup of coffee, he treats the coffee seller merely as a means, but though this man's attitude is wrong he is not acting wrongly. Nor does my Egoist act wrongly when he risks his life to save a drowning child, though he is using this child as a mere means of getting some reward.

To meet such objections, as I have said, we can revise Kant's claim. According to

> *the Third Mere Means Principle*: It is wrong to act in any way that treats anyone merely as a means, or comes close to doing that, if our act will also be likely to harm this person.

But we ought, I believe, to reject this principle. Let us again compare

> *Lifeboat*, in which you could save either me or the five,
>
> *Tunnel*, in which you could redirect a runaway train so that it kills me rather than the five,

and

> *Bridge*, in which you could save the five only by killing me.

According to one view, in all three cases, you ought to save the five. It makes no difference whether, in saving the five, you would be killing me. When people's lives are threatened, we ought to do whatever would save the most lives.

According to a second view, you ought to save the five only in *Lifeboat*. We have a duty not to kill which outweighs our duty to save people's lives. On this view, it would be wrong for you to save the five in both *Tunnel* and *Bridge*, since these ways of saving the five would both kill me. As before, it makes no difference whether you would be killing me as a means.

According to a third view, you ought to save the five in *Lifeboat*, and you would be at least permitted to save the five in *Tunnel*, but it would be wrong for you to save the five in *Bridge*. This, I believe, is the most widely held of these three views. On this view, it *does* make a difference whether you would be killing me as a means.

If we accept this third view, we might appeal to

> *the Harmful Means Principle*: It is wrong to impose harm on someone as a means of achieving some aim, unless
>
> (1) there is no better way to achieve this aim,
>
> and
>
> (2) given the goodness of this aim, the harm we impose is not disproportionate, or too great.

This principle does not tell us which harms would be too great. We would have to use our judgment here. On one view, there is an upper limit on the amount of harm that we could justifiably impose on someone as a means. According to Thomson, for example, it would be wrong to kill or seriously injure one innocent person, however many other people's lives we could thereby save. Most of us would accept a less extreme view. We would believe it to be right to kill one innocent person if that were the only way in which we could prevent some nuclear explosion that would kill as many as a million other people. But we may believe it to be wrong to kill one person as a means of saving only five, or only fifty other people. There would be cases in between in which this moral question would have no clear or determinate answer.

On what I have called the Standard View, if we harm someone, without this person's consent, as a means of achieving some aim, we thereby treat this person merely as a means. As I have argued, that may not be true. When I break Brown's leg to stop him from murdering me, I am *harming* Brown as a means of defending myself. But I am not treating *Brown himself* as a means, so I cannot be treating Brown merely as a means.

Return next to cases in which, if we impose harm on someone as a means, we may also be treating this person as a means. When we ask whether such an act would be wrong, we have two questions:

> Q1: Might the wrongness of this act partly depend on whether we would be harming this person as a means of achieving some aim?
>
> Q2: Might the wrongness of this act partly depend on whether we would also be treating this person *merely* as a means?

When we compare cases like *Bridge* and *Tunnel*, we may decide that the answer to Q1 is Yes. We may believe that, though you could justifiably redirect the runaway train so that it would kill me rather than the five, it would be wrong for you to save the five *by* killing me. I have *not* been arguing against this view.

The answer to Q2, I believe, is always or nearly always No. If you killed me in *Bridge* without my consent, you might not be treating me merely as a means, or be close to doing that. Your treatment of me might be governed in a sufficiently important way by some relevant moral principle, such as Kant's Consent Principle. And it might be true that, if you had been closer to the train, you would have saved the five by killing yourself rather than me. But these facts would not, I believe, affect whether your act would be wrong. If it would be wrong for you to kill me as a means of saving the five, this act would be wrong *whether or not* you would also be treating me merely as a means. Even if you were

not treating me merely as a means, and were not even close to doing that, these facts would not justify your act.

Turn next to cases in which we *could* justifiably impose harm on someone as a means. In *Third Earthquake*, you cannot save your child's life except by crushing Black's toe, without Black's consent. This act, I believe, would be justified. If someone crushed my toe to save their child's life, I would not (I hope) complain. Though some people would believe this act to be wrong, these people would accept that there are some lesser harms that we could justifiably impose on someone, if that was our only way to save someone else's life. On Thomson's view, for example, we could permissibly save someone's life by bruising someone else's leg, causing this other person 'a mild, short-lasting pain'. So we can suppose that, in

> *Fourth Earthquake*, my gangster cannot save his child's life except by bruising Black's leg, without her consent, causing her a mild, short-lasting pain.

This gangster regards Black as a mere means. He would kill or gravely injure Black if that would help him to achieve any of his aims. So if this gangster saved his child by bruising Black's leg, he would both be imposing harm on Black and be treating Black merely as a means. According to Kant's Formula of Humanity, which includes the Mere Means Principle, it is wrong to act in any way that treats people merely as a means. According to the Third Mere Means Principle, it is wrong to impose harm on people in any way that also treats them merely as a means. These principles both imply that, if my gangster saved his child's life by bruising Black's leg, he would be acting wrongly.

That is an unacceptable conclusion. Though this gangster has the wrong attitude to Black, he could justifiably save his child's life by imposing this small harm on Black. This child has a moral claim to be saved; and her claim is not undermined, or overridden, by the wrongness of her father's attitude to Black. Similar claims apply to other cases. If you would be morally permitted to save your child in *Third*

Earthquake by causing Black to lose one toe, my gangster would be morally permitted to save his child in the same way.

It has been widely believed that, to explain the wrongness of harming some people as a means of benefiting others, we could appeal to Kant's claim that we must never treat people merely as a means. This belief, I have argued, is mistaken. If it would be wrong to impose certain harms on people *as a means* of achieving certain good aims, these acts would be wrong even if we were *not* treating these people *merely* as a means. And if it would *not* be wrong to impose certain lesser harms on people as a means of achieving such aims, these acts would not be wrong even if we *were* treating these people merely as a means.

Kant's claim contains an important truth. It is wrong to *regard* anyone merely as a means. But the wrongness of our *acts* never or hardly ever depends on whether we are treating people merely as a means.

10

Respect and Value

33 Respect for Persons

In another comment on his Formula of Humanity, Kant writes

> every rational being . . . must always be regarded as an end . . . and is an object of respect.

This requirement to respect all persons is one of Kant's greatest contributions to our moral thinking. But it does not tell how we ought to act.

Wood suggests that

> (A) we must always treat people in ways that express respect for them.

We can treat people rightly, however, without *expressing* our respect for them. Wood suggests that, whenever we treat people rightly, our acts could be taken to express respect for these people. But on this suggestion (A) would tell us only that we must always treat people rightly. (A) would not help us to decide which acts are right, since we could not decide whether some act would express respect for people except by deciding whether this act would be right.

Some writers suggest that

> (B) it is wrong to treat people in ways that are incompatible with respect for them.

Some wrong acts are clearly incompatible with respect for persons. Kant's examples are: disgraceful or humiliating punishments, ridicule, defamation, and acts that display arrogance or contempt. But Kant's formula is intended to cover all wrong acts, and most wrong acts do not treat people in such disrespectful ways.

All wrong acts, some writers suggest, are in a wider sense incompatible with respect for persons. On this suggestion, (B) would not be a useful claim. As before, to decide whether some act would be in this wider sense incompatible with respect for persons, we would first have to decide whether this act would be wrong. If this act would *not* be wrong, it would be compatible with respect for persons. As both Kant and Sidgwick warn, moral philosophers often make claims that seem to give us 'valuable information' but really tell us only that acts are wrong if they are wrong.

Kant also claims that

> (C) we must always respect *humanity*, or the 'rational nature' that makes us persons.

Wood calls (C) 'the most useful formulation' of Kant's supreme principle of morality. Though (C) cannot directly solve all moral problems, this principle provides, Wood claims, 'the correct basis for deciding moral questions'. To support this claim, Wood points out that in his last and longest book about morality, Kant often makes remarks that seem to appeal to (C).

Kant's remarks do not, I believe, show (C) to be a useful principle. As Wood himself concedes, Kant's appeals to (C) are 'usually both brief and casual'. Such remarks add little to Kant's view. For example, Kant writes that our duty to develop our talents 'is bound up with the end of humanity in our own person'. Kant makes other claims that Wood rightly rejects. It would be wrong, Kant claims, for any of us to give ourselves sexual pleasure, or to hasten our deaths to avoid suffering, because such acts debase or defile humanity. And when he condemns telling some lie even 'to achieve some really good end', Kant writes that any liar 'violates the dignity of humanity in his own person', so that he becomes a 'mere deceptive appearance of a human being', who has 'even

less worth than if he were a mere thing'. These are not the claims that make Kant the greatest moral philosopher since the ancient Greeks.

Wood suggests that, in making these claims, Kant misapplies (C). We can reject Kant's views about sex, suicide, and lying, Wood writes, 'because we justifiably believe that we know more about what respect for humanity requires in these matters'. It is 'an advantage' of this principle 'that both sides in profound moral disagreements can use it to articulate what they regard as their strongest arguments'.

This assessment seems to me mistaken. When Kant claims that certain acts would violate or debase humanity, and we reject these claims, neither Kant nor we are giving our strongest arguments. Nor would (C) help us to decide, in difficult cases, which acts would be wrong.

34 Two Kinds of Value

When Kant explains the sense in which we must always treat rational beings as ends, he claims that such beings have *dignity*, by which he means a kind of supreme value. This claim raises one of the deepest questions in ethics: that of how what is good is related to what is right, or to what we ought morally to do. Kant also claims that, rather than following the ancient Greeks by first asking which ends are good and then drawing conclusions about which acts are right, we ought to reverse this procedure. Rawls calls it a central feature of Kant's moral theory that 'the right' is, in this way, 'prior to the good'. But Wood in contrast claims that, though Kant's Formula of Humanity 'takes the form of a rule or commandment, what it basically asserts is the existence of a substantive value'. And Herman suggests that Kant's 'fundamental theoretical concept' is 'the Good', and that 'Kant's ethics is best understood as an ethics of value'.

Before we consider Kant's claims about value, it will help to draw some more distinctions. Many things are good or bad in what I have called *reason-implying* senses. Such things have certain kinds of properties or features that would, in some situations, give us or others reasons to respond to these things in certain ways.

Some of these good things have a kind of value that, as Scanlon and others say, is *to be promoted*. Two examples are happiness and the relief or prevention of suffering. When things have this kind of value, it is really these things, not their value, that we have reasons to promote.

What we can promote are events, in the wide sense of 'event' that also covers acts and states of affairs. Events can be good or bad either as an end or as a means to some end. On some views, acts can be good or bad only as a means. We ought, I believe, to reject such views. We act well, for example, if we bring up our children well, or we act as good friends or lovers, or we engage with some success in various other worthwhile activities, or we act rightly and treat people with respect. Such things might be worth doing, not merely as a means to happiness or other good ends, but partly or wholly for their own sake. So we should include acts among the events that might be good or bad as ends.

On what seems to me the best view about the goodness of events, which I shall call

> *Actualism*: Possible acts and other events would be good as ends when they have intrinsic properties or features that give us reasons to want them to be actual, by being done or occurring, and to make them actual if we can. Possible acts and other events would be good as a means when our making them actual would be an effective way of achieving some end.

Similar claims apply to events that would be bad as ends, or bad as a means to some end. Events may be good as ends either for particular people or in the impartial-reason-implying sense, or both. As well as having reasons to try to cause or prevent good or bad events, we have reasons to have various other attitudes towards them, such as hope, gladness, fear, and regret. These are all attitudes towards the possibility or fact that such events are actual or real, being a part of the way things go.

Since Actualism applies to all possible acts and all of their possible effects, this view covers everything whose goodness is directly relevant to any decision about what we should do. We have a reason to act in some way if and only if, or just when, this act would be in some way good either

as an end, or as a means to some good end. Actualism does not, however, claim to cover the goodness of things that are not acts or other events.

According to some writers, this view can be widened to cover the goodness of some persisting things, such as people and works of art. Such things are claimed to be good when their nature gives us reasons to want them to exist, or continue to exist, and reasons to make that happen if we can. Moore even writes:

> when we assert that a thing is good, what we mean is that its existence or reality is good.

But these claims are mistakes. Something's existence can be good though this thing itself is not good, and vice versa. There are many bad people, for example, whose continued existence would be good as an end. When some good person is dying a slow and painful death, the continued existence of this person may be bad as an end. And there would be nothing good in the continued existence of good works of art if no one could ever see them.

According to what Scanlon calls *teleological* theories, it is only acts and other events that have *intrinsic* value in the sense of being in themselves good. Scanlon rightly rejects this claim. There are other things that can be in themselves good, such as people, books, and arguments. Since these things are not events, we cannot want them to happen, or make them happen. But we can respond to them in other ways. We can have reasons to read good books, be convinced by good arguments, and try to become more like good people.

We can now turn to a kind of value which, as Scanlon and others say, is to be *respected* rather than promoted. As before, when things have such value, it is really these things, not their value, that we have reasons to respect. Though people are the best example of what can be claimed to have such value, we can start with some other examples. These can be things that are claimed to have symbolic, historical, or associational value, such as our nation's flag, the oldest living tree, icons and other religious paintings, and the bodies of dead people.

Understanding something's value, Scanlon writes, is in part 'a matter of knowing *how* to value it — knowing what kinds of actions and attitudes

are called for'. Many of these acts and attitudes can be loosely called ways of respecting or honouring this thing. We might respect our nation's flag, the oldest tree, and some religious painting by refusing to use these things as a dishcloth, firewood, and the target in a game of darts. To respond appropriately to the value of many such things, we ought to protect them, so that they continue to exist. But that is not always true. We can respond appropriately to the value of dead people's bodies, not by trying to preserve them as the ancient Egyptians did, but by destroying them in some respectful way, such as burning them bedecked with flowers on some funeral pyre, rather than throwing them onto some rubbish dump.

The value of such things is quite different from the goodness of good ends, or good people. It is not a kind of *goodness*. Though some dead people's bodies would be good as *cadavers*, for use in teaching anatomy or surgery, and some other bodies would be good as corpses in some horror film, these are not the kind of value that all dead people's bodies can be claimed to have. And some religious paintings are not good. Though this kind of value is not a kind of goodness, and is not a value that is to be promoted, when we could respond to the value of such things by treating them in respectful ways, these *acts* would be good as ends, having the kind of value that is to be promoted.

We can turn next to claims about the value of human life. Appreciating this value, Scanlon writes,

> is primarily a matter of seeing human lives as something to be respected, where this involves seeing reasons not to destroy them, reasons to protect them, and reasons to want them to go well.

To see that we have such reasons, however, we don't need to see human lives as having a kind of value that is to be respected *rather* than promoted. When people's lives go well, that is both good for these people and impersonally good, in the reason-implying senses. Such happy and well-lived lives are good as ends. We have reasons to protect the living of such good lives, and to help these people in other ways to make their lives go well.

On some views, human life has a different kind of value. Suppose that you have begun to die a slow, painful, and undignified death, and you

have nothing important left to do. You may have strong reasons to kill yourself, and other people may have strong reasons to help you to act in this way. Of those who appeal to the value of human life, some would believe that this act would be wrong. These people might agree that it would be both better for you, and impersonally better, if you died an earlier, natural death. But you ought not to kill yourself, these people believe, and other people ought not to help you, since such acts would fail to respect the value of human life. On this view, respecting the value of someone's life is not the same as, and may conflict with, doing what would both be best for this person and be what this person chooses.

Scanlon rejects this view. We have reasons not to end someone's life, he writes, only 'as long as the person whose life it is has reason to go on living or wants to live'. Scanlon here denies that a person's life has the kind of value that we ought to respect in ways that conflict with this person's well-being and autonomy. This, I believe, is the right view about the value of human life. To defend the claim that suicide and assisting suicide would be, in such cases, wrong, we would need some other argument.

It is not human life but the people who *live* these lives who should be claimed to have the kind of value that should be respected rather than promoted. We should respect this value, Scanlon claims, by treating people only in ways that could be justified to them. Kant similarly claims that, to respect people, we should treat them only in ways to which they could rationally consent.

35 Kantian Dignity

We can next consider Kant's claims about value. While making these claims, Kant distinguishes three kinds of end. What Kant calls *ends-to-be-produced* are the aims or outcomes that we could try to achieve or bring about. These are ends in the ordinary sense, as in the claim that the relief of suffering is a good end. Kant contrasts such ends with what he calls *existent* or already existing ends, of which his main examples are rational beings, or people. Kant's third kind of end he calls *ends-in-themselves*. Such things have what Kant calls *dignity*, which he defines as absolute, unconditional, and incomparable value or worth. Such value is supreme, or unsurpassed, in the sense that nothing else has greater value.

According to some writers, Kant believes that such supreme value is had only by some existent ends, such as rational beings, whose value is of the kind that is to be respected rather than promoted. But there are several ends-to-be-produced which Kant claims to have supreme value, and to be ends that we ought to try to promote, or achieve.

One such end is having a *good will*. Our will is good, Kant claims, when we do our duty because it is our duty, and not with some other aim, such as avoiding punishment. Our having a good will can be taken to be either a mental state or disposition, or an activity which consists in good willing. Regarded in either way, having a good will is something that, on Kant's view, we ought to try to achieve. In Kant's own words, 'the true vocation of reason must be to produce a will that is good'.

Another end-to-be-produced with supreme goodness is what Kant calls the *Realm of Ends*. This is the possible state of affairs, or *possible world*, that we together would produce if everyone had good wills and always acted rightly.

A third such end is what Kant calls the *Highest* or *Greatest Good*. This possible world is the Realm of Ends with the further feature that everyone would have all of the happiness that their virtue would make them deserve. Kant claims that 'we ought to try to promote' this end, and that 'reason . . . commands us to contribute everything possible to its production'.

There may be a fourth such end. Kant calls rational beings 'something whose existence in itself has absolute worth'. And he writes that, if there were no rational beings, the Universe would be 'a mere waste, in vain, without a final purpose'. These remarks suggest that, on Kant's view, the continued existence of rational beings is another end-to-be-produced with supreme value.

We can now return to Kant's claim that rational beings or people are ends-in-themselves, who have dignity, or supreme value. As I have said, people are not ends-to-be-produced. And their value is of a different kind. On Kant's view, as Wood and Herman claim, 'even the worst human beings have dignity', and a person whose will is good 'is of no greater value' than someone with an ordinary or a bad will. This part of Kant's view is, I believe, a profound truth. But the value of the morally worst people is not a kind of goodness. Hitler and Stalin were not good.

People have dignity or value in the quite different sense that, given their nature as rational beings, they must always be treated in certain helpful or respectful ways. A similar claim applies, I believe, to all sentient beings. Even the lowliest worm, if it can feel pain, has a kind of dignity, in an extended Kantian sense. A worm cannot be in itself good, but its nature makes it a being on which it would be wrong to inflict pointless pain.

I have been ignoring one complication. Kant sometimes uses 'humanity' to refer to rationality, or what he also calls 'rational nature'. So, when Kant claims that humanity is an end-in-itself with dignity, or supreme value, he might mean that rationality has such value. And though the value of rational beings is not a kind of goodness, their being rational might be claimed to be good. Herman writes that, in Kant's ethics, 'The domain of "the good" is rational activity and agency', and that Kant 'grounds morality' on 'rationality as a value'. Wood even calls Kant's claim about rationality's value 'the most fundamental proposition in Kant's entire ethical theory'.

On Kant's view, like having a good will, rationality is in part an end-to-be-produced, or promoted. We ought to use our rationality, and we can try to become more rational by developing our rational abilities. Kant calls *dignity* a value that is 'infinitely far above' a lower kind of value, which he calls *price*. Among the things that have mere price Kant includes pleasure and the absence of pain. So, if Kant meant to claim that rationality or rational activity had dignity, Kant's view would imply that rationality has infinitely greater value than the relief of pain. Cardinal Newman claims that, though both sin and pain are bad, sin is infinitely worse, so that, if all mankind suffered extremest agony, that would be less bad than if one venial sin were committed. Though this view is horrific, we can understand why it has been held, since we can see how sin might seem infinitely worse than pain. If rationality or rational activity had dignity in the sense of infinite value, and preventing pain had only finite value, Kant's view would have implications that would be even harder to accept. On this view, for example, we ought to increase our ability to play chess, or to solve crossword puzzles, rather than saving any number of other people from any amount of pain. That conclusion would be insane.

It might be objected that, even on this view, we ought to save these other people from pain, since that would help them to act rationally. But

we might be saving these people from pain during surgical operations, by making them unconscious. That would not help them to act rationally.

It might next be claimed that rationality's value is of the kind that is to be respected rather than promoted. That is not Kant's view, since Kant often claims that we ought to try to develop and use our rational abilities. And this revised version of Kant's view would face a similar objection. We respect the value of persons, not by adding new people to the world, but by following various other moral requirements, such as the requirement not to kill or injure people. If rationality had similar value, as Hill points out, there would be similar requirements not to damage or impair people's rational abilities. And if rationality's value was infinitely far above all price, it would be wrong to 'trade' or 'sacrifice' any rational ability for the sake of anything with mere price, such as relief from pain. So it would be wrong for us to damage our ability to play chess or solve crossword puzzles, even if that would be how we could save any number of people from any amount of pain. That conclusion would also be insane.

Kant's view does not, I believe, have such implications. When Kant claims that humanity has dignity, he is seldom referring, I believe, to rationality. Kant distinguishes between (1) our capacity for acting morally and having a good will, and (2) our other rational capacities and abilities. We can call (2) our *non-moral rationality*. Just after defining dignity as a kind of absolute and incomparable value, Kant writes:

> morality, and humanity insofar as it is capable of morality, is that which alone has dignity.

The word 'humanity' cannot here refer to non-moral rationality. In many other passages, Kant distinguishes between ourselves and what he calls 'the humanity in our person'. These uses of 'humanity' mostly refer, I believe, not to our rationality, but either to our capacity for acting morally and having a good will, or to ourselves as what Kant calls *noumenal* beings. Though some of Kant's remarks suggest that non-moral rationality is an end-in-itself, with supreme value, he is not, I believe, committed to this view. Kant is 'the least exact of the great thinkers', and his uses of 'humanity' are shifting and vague. Kant does condemn some vices, such as gluttony and drunkenness, on the ground

that such vices interfere with our rational activities or abilities. But Kant's main claims do not imply that it would be wrong for us to eat too much, or to make ourselves drunk, even if these were the only ways of saving any number of people from any amount of pain.

In his claims about value, Herman writes, Kant provides 'a radical critique of traditional conceptions'. On Kant's view, 'past moral philosophy . . . mistakes the nature of the good'.

Kant does not, I believe, provide such a critique. If Kant claimed that nothing has the kind of value that is to be promoted, he would be rejecting many earlier views. But as we have seen, Kant claims that such value is had by our having good wills, and by the Realm of Ends, and by Kant's Greatest Good, the possible state of affairs or world in which everyone would be virtuous and happy. On Kant's view, these are all ends-to-be-produced, which we ought to promote as much as we can. In his claims about which things have such value, Kant also follows earlier philosophers, many of whom claim that virtue and happiness are the two things that are good as ends.

Kant may not accept one widely held view about value, since he often ignores the reason-implying senses in which things can be non-morally good or bad. He claims for example, that the principle of prudence, or of doing what would promote our own happiness, is a merely *hypothetical imperative*, which applies to us only because we want to be happy. Kant here ignores our non-moral reasons to want to be happy. In his account of practical reason, Kant describes morality and instrumental rationality, with little but a wasteland in between. Kant's ignoring of non-moral goodness, which I discuss in Appendix G, is not, however, a critique.

There is another widely held view that Kant may not accept. On this view, to be valuable is always to be in some way good. When Kant claims that all rational beings have the kind of value that he calls dignity, he does not mean that all rational beings are good. As I have said, Kant means that all rational beings have a kind of value that is to be respected, since these beings ought to be treated only in certain ways. This value is a kind of *status*, or what Herman calls 'moral standing'. Such value is ignored by many traditional views.

Kant, I believe, is right to claim that even the morally worst people have the same moral status as anyone else. And by calling this status *dignity* or *supreme value*, Kant expresses this claim in a helpfully persuasive way. But for the idea of moral status to be theoretically useful, it must draw some distinction, by singling out, among the members of some wider group, those who meet some further condition. In Roman law, to give one analogy, only those human beings who were not slaves had full legal status, and counted as persons. In democracies, only those persons who are adults have the status of being entitled to vote, and in many countries only those persons who are citizens have the status of being entitled to certain benefits. On Kant's view, in contrast, *all* rational beings or persons ought to be treated only in certain ways. We add little if we say that all rational beings or persons have the moral status of being entities who ought to be treated only in these ways.

Kant's claims about value are also, in one way, misleading. As I have said, when Kant claims that all rational beings have dignity, or supreme value, he does not mean that all such beings are good. But Kant claims that such supreme value is also had by morality, good wills, the possible worlds which are the Realm of Ends, and the Greatest Good. The value of these things, on Kant's view, *is* a kind of goodness. So, in his claims about value, Kant fails to distinguish between being supremely good and having a kind of moral status that is compatible with being, like Hitler and Stalin, very bad. It is easy, however, to add this distinction to Kant's view.

36 The Right and the Good

The Highest or Greatest Good, Kant claims, would be a world in which everyone was both wholly virtuous, or morally good, and had all of the happiness that their virtue would make them deserve. Kant also writes:

> Everyone ought to strive to promote the Greatest Good.
>
> The moral law commands me to make the greatest possible good in a world the final object of all my conduct.

According to what we can call this

> *Formula of the Greatest Good*: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

This ideal world would be hard to achieve. So, in applying this formula, we should compare unideal but more achievable states of the world, and ask how we could get as close as possible to Kant's ideal.

It would be best, Kant claims, if everyone's degree of happiness was *in proportion* to their degree of virtue, or worthiness to be happy. That would be true in the ideal world in which we would all be wholly virtuous and happy. Some writers suggest that, of the worlds that are not ideal, the best would be those in which this *proportionality condition* would be met. But this seems unlikely to be Kant's view. Everyone's happiness might be in proportion to their virtue if no one was either virtuous or happy, or if everyone was both vicious and miserable. These worlds would clearly be much worse than worlds in which everyone had great virtue and great happiness, but some people had slightly less or slightly more happiness than they deserved. So we can assume that, on Kant's view, it would always be better if there was more virtue, and more deserved happiness, even if the proportionality condition would be less well met.

Kant claims, implausibly, that no one can affect how virtuous other people are. On this assumption, we can promote virtue only by increasing our own virtue. We can best do that by trying to have good wills, and doing whatever else we ought to do. We can best promote deserved happiness by trying to give happiness to people who are less happy than they deserve. It is often claimed that we cannot act in this way, since we cannot know how much happiness people deserve. We do not, however, need *knowledge*. It would be enough to have rational beliefs about which people are more likely to deserve more happiness. As Kant assumes, we often have such beliefs. We could act on these beliefs by trying to make these people happier. So Kant's Formula of the Greatest Good gives us an aim that we could try to achieve.

We can next draw some more distinctions, and introduce some of Kant's other claims. Moral theories are in the widest sense

> *Act Consequentialist* if they claim that everyone ought always to do, or try to do, whatever would best achieve one or more common aims.

According to one such theory, *Hedonistic Act Utilitarianism* or

> *HAU*: Everyone ought always to produce, or try to produce, the greatest possible amount of happiness minus suffering.

These theories are *person-neutral* in the sense that they give the same common aims to everyone. According to most moral theories, and most people's moral beliefs, there are some common aims that everyone ought to try to achieve, such as the aim that people be saved from starving. But each of us ought also to try to achieve many *person-relative* moral aims. On such views, for example, rather than having the common aims that promises be kept and children be cared for, each of us ought to try to keep our own promises, and to care for our own children. A third group of views do not give us any common moral aims. That is true, for example, of the view that our only duties are to obey the Ten Commandments.

Some moral theories are wholly or partly *value-based*, in the sense that they appeal to claims about what is good or bad, in some significant, substantive sense. According to what we can call *Value-based Act Consequentialism*, or

> *VAC*: Everyone ought always to do, or try to do, whatever would make things go best.

On this version of HAU, for example, everyone ought to produce, or try to produce, the greatest net sum of happiness because that is how we could make things go best.

As well as making claims about what is good and what we ought morally to do, some moral theories make claims about how the concept *good* is related to the moral version of the concept *ought*. According to some theories, the concept *good* is fundamental, and can be used

to define this version of the concept *ought*. According to some other theories, it is the concept *ought* that is fundamental, and can be used to define the concept *good*. According to a third group of theories, neither of these concepts can be defined in terms of the other. The best theories, I believe, are of this third kind. Because these are the only theories that use *good* and *ought* in senses that are independent, these are the only theories that can make true substantive claims about the relations between what is good and what we ought morally to do.

As one example of the first kind of theory, we can take Moore's *Principia Ethica*. Moore claims that, when we say that

> we *ought* to do something, or that this act is *right*, we mean that this act would do the most good, by making things go best.

We can call this the *goodness-promoting* sense of 'ought'. Moore also claims

> M1: Everyone ought always to do what would make things go best.

This claim may seem to be a version of Value-based Act Consequentialism. But if Moore is using 'ought' in his goodness-promoting sense, M1 is a concealed tautology, one of whose open forms would be

> M2: Everyone would always do what would make things go best if everyone always did what would make things go best.

Everyone could accept this claim, whatever their moral beliefs. Moore's *Principia* does not put forward a substantive moral view.

Kant's view is the opposite of Moore's, since Kant claims that we should define *good* in terms of *ought*. In Kant's words,

> the concepts of *good* and *evil* must not be determined before the moral law . . . but only after it . . . and by means of it.

Surprisingly, Kant also claimed:

> All imperatives are expressed by an 'ought'. . . and say that . . . some act would be good.

Kant may here seem to be doing just what he claims that we must not do, by defining *ought* in terms of *good*. Kant similarly calls certain acts 'practically necessary, that is, good'. But these remarks do not use 'good' in any of its ordinary senses. In these ordinary senses, for example, some act may be good, though some other act would be even better. In these and other passages, Kant does not distinguish between some act's being good and this act's being practically necessary, or what we ought to do. And it is these latter words that better express what Kant has in mind. So I suggest that, when Kant calls some act 'good', he means that this act is what we ought to do. Kant would then be following his requirement that *good* be defined in terms of *ought*, since he would be using 'good' in an *ought-based* sense.

When Kant calls some end or outcome 'good' or 'best', he seems often to be using a similar ought-based sense. For example, when Kant claims

K1: Good wills are supremely good,

he seems in part to mean

K2: Everyone ought to try to have a good will.

But Kant may also mean that we ought to try to have such wills *because* such wills are supremely good. This use of 'good' would not be ought-based. In this respect Kant's moral theory may be, as Herman claims, an ethics of value. But Kant would not be doing what he claims that we must not do, by deriving the content of the moral law from his beliefs about what is good. From the claim that good wills are supremely good we may be able to derive K2. But we cannot draw any other conclusions about what we ought to do.

The ancient Greeks, Kant claims, did make this mistake, since they tried to derive the moral law from their beliefs about the *Summum Bonum*, or the *Greatest Good*. As we have seen, however, Kant himself describes an ideal world which he calls the Highest or Greatest Good, and he claims that everyone ought always to try to produce this world. Is Kant here making what he calls the 'fundamental error' of the ancient Greeks? Is

he deriving his beliefs about what we ought to do from his beliefs about the Greatest Good?

It may seem so. As we have seen, Kant claims

> K3: Everyone ought always to strive to promote the Greatest Good.

This may seem to be another version of Value-based Act Consequentialism. Kant may seem to be claiming that everyone ought always to try to produce the world that would be the best, or be the greatest good. And he makes other such remarks, as when he writes, of every human being, 'his duty at each instant is to do all the good in his power'.

This is not, I believe, the best way to interpret K3. Kant, I suggest, uses the phrase 'the Greatest Good' in an ought-based sense, to mean 'what everyone ought always to strive to promote'. If this is what Kant means, K3 could be restated as

> K4: Everyone ought always to strive to promote the world that everyone ought always to strive to promote.

This claim may seem to be a mere tautology, which everyone could accept. But that is not so. K4 implies that we should accept some version of Act Consequentialism, since K4 implies that there is some world that everyone ought always to strive to promote. Many people would reject that claim.

K4 does not, however, imply a *value-based* version of Act Consequentialism. And when Kant claims K3, he may also be using 'the Greatest Good' to refer to the possible world that he elsewhere claims to *be* the Greatest Good. K3 could then be more fully stated as

> K5: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

This is the clearest statement of this part of Kant's view, and this claim does not even use the words 'good' or 'best'. So Kant's version of Act Consequentialism is *not* significantly value-based.

37 Promoting the Good

Nor is Kant's view clearly *Act* Consequentialist. Kant's Formula of the Greatest Good might be claimed to be the only principle we need, because we ought always to try directly to promote Kant's ideal world. But that is not Kant's view. Kant claims that we ought to follow certain other formulas, such as his Formulas of Humanity and of Universal Law. So we can next ask how Kant's claims about the Greatest Good are related to his other formulas.

We can assume, Kant writes, that

> the laws of morality lead by their fulfilment to the highest end.

He also writes:

> the strictest observance of the moral laws is to be thought of as the cause of the ushering in of the Greatest Good (as end).

In these and other passages, Kant assumes

> K6: It is by following the moral law, as described by Kant's other formulas, that everyone could best promote the Greatest Good.

If everyone followed the moral law, and had good wills, everyone would thereby promote one element in Kant's ideal world, universal virtue, since such universal virtue would *consist* in everyone's following the moral law and having good wills. But this is not all that Kant means. When Kant claims that, if everyone followed the moral law, this would *lead to* or be the *cause of the ushering in* of the Greatest Good, Kant must be referring to the other element in this ideal world, universal deserved happiness. So Kant seems to assume

> K7: It is by following the moral law that everyone could best give everyone the happiness that their virtue would make them deserve.

Though everyone's following the moral law would make the world much closer to Kant's ideal, this would not be enough, Kant claims, fully to achieve this aim, since we would not be able to give everyone all of

the happiness that they would deserve. Some good people, for example, would die young. But we can hope that our souls are immortal, and that after our deaths God will give everyone the rest of the happiness that they deserve.

We may doubt that Kant could have assumed K7. Kant seems to have believed that we ought to follow certain strict rules, such as rules forbidding lying, stealing, and breaking promises. It may seem unlikely that Kant could have believed that following such rules would most effectively promote deserved happiness.

That is *not*, however, unlikely. It was widely assumed, when and before Kant lived, that

> (A) it is by following the rules of common sense morality, rather than by trying directly to promote everyone's happiness, that everyone could best promote everyone's happiness.

This assumption is also fairly plausible, as Sidgwick later argued. In trying to predict which acts would produce most happiness, people would make serious mistakes. And they would often deceive themselves in their own favour. It is easy to believe, for example, that our need for the property that we could steal is greater than the owner's need. If everyone was always trying to maximize happiness, that would also undermine or weaken various valuable social practices or institutions, such as the practice of trust-involving promises. And it would be in several ways bad if everyone had the motives of those who always try to maximize happiness. To be able always to act in this way, most of us would have to lose too many of the motives—such as strong love for particular people—on which much of our happiness depends.

We can next draw some distinctions that many earlier thinkers did not draw. I shall now use 'Consequentialist' to refer only to value-based views, and I shall use 'best' as short for 'best or expectably-best'. If we suppose that everyone will try to follow some set of rules, some possible rules would be

> *optimific* in the sense that, if these are the rules that everyone tries to follow, things would go best.

For the reasons just given, Sidgwick believed that the rules of common sense morality are fairly close to being optimific. According to one version of *Rule Consequentialism*, or

> *RC*: Everyone ought always to try to follow the optimific rules.

According to one version of *Act Consequentialism*, or

> *AC*: Everyone ought always to try do what would make things go best.

Of the people who accept either of these views, most now assume that these views conflict, so that we must choose between them. These people believe that

> (B) in some cases, breaking some optimific rule would be likely or certain to make things go best.

As an Act Consequentialist, Sidgwick claims that, in such cases, we ought to break this optimific rule. According to most Rule Consequentialists, we ought instead to follow the optimific rules even when, by acting in this way, we would be likely or even certain to make things go worse.

There have been some people, however, who reject (B). These people believe that

> (C) it is by trying to follow the optimific rules that everyone would always be most likely to make things go best.

Moore came close to accepting (C). In trying to do the most good, Moore claims, we ought always to try to follow certain optimific common sense rules. If (C) were true, these two forms of Consequentialism would not conflict but coincide, and we could accept them both. According to what we can call *Act-and-Rule Consequentialism*, or

> *ARC*: Everyone ought always to try to follow the optimific rules, since that is how everyone would be most likely to do what would make things go best.

In asking whether (C) is true, so that these forms of Consequentialism coincide, we must appeal to some view about how we ought to assess the effects of our acts. According to what we can call

> *the Marginalist View*: To decide how much good some act would do, we should ask what *difference* this act would make. The good that some act would do is the amount by which, if this act were done, things would go better than they would have gone if this act had not been done.

When we consider some kinds of case, this view can seem implausible. One example are cases in which some good result would be fully achieved if some number of people act in some way. If *more* than this number of people act in this way, the Marginalist View may imply that none of these people does any good. Suppose that, in

> *Rescue*, a hundred miners are trapped underground, with flood-waters rising. These miners' lives will all be saved if four people join some rescue mission.

To make the causal relationships clear, we can suppose that, if four people stand on some platform, these people's weight will together be enough to raise each miner to the surface. On the Marginalist View, if five people join this mission, none of these people will save anyone's life. It is true of each of these five people that, if this person hadn't joined this mission, and stood on this platform, that would have made no difference, since the other four people would have saved all of the hundred miners' lives. According to Marginalists, none of these people does any good.

That conclusion may seem absurd. If none of these people saves anyone's life, how did a hundred lives get saved? Some writers claim that, to avoid such absurd conclusions, we should appeal to the effects of what people *together* do. According to one such view, which we can call

> *the Share of the Total View*: When some group of people together produce some good effect, the good that each person does is this person's share of the total good.

This view implies that, if five people join our rescue mission, thereby together saving a hundred lives, each person should be counted as saving twenty lives. It is irrelevant that, if any of these five people had not joined this mission, that would have made no difference. On this

view, in deciding which of our possible acts would do the most good, we should ignore the effects of each act when considered on its own.

When Hume discusses our obligations not to steal and to respect other property rights, he asserts a similar but vaguer view. Justice and fidelity, Hume claims, 'are absolutely necessary to the well-being of mankind'. But the benefits of justice are 'not the consequence of every single act', since any particular just act, when 'considered in itself', may have effects that are 'extremely hurtful'. The benefits of justice arise only 'from *the whole scheme*' or 'the observance of the general rule'. Hume therefore claims that, to produce these benefits, we must follow strict rules, making no exceptions even when breaking some rule would when 'considered in itself' have good effects. Such rules must be strict, or inflexible, because it is 'impossible to separate the good from the ill'.

On Hume's view, which we can call

> *the Whole Scheme View*: To decide how much good some act would do, we should not ask how much difference this act *by itself* would make. Each of our acts would do the most good if this act is one of a set of acts that would *together* do the most good.

If Act Consequentialists reject the Marginalist View and accept the Whole Scheme View, they might accept Hume's claim that we ought to follow certain strict rules, such as 'Never steal', since they might believe that this is how each of our acts would do the most good. These Act Consequentialists would then also be *Rule* Consequentialists. If the Whole Scheme View were true, so would be the claim that

> (C) it is by trying to follow the optimific rules that everyone would be most likely to make things go best.

On these assumptions, these two forms of Consequentialism would not conflict but coincide.

When Kant defends another strict rule, 'Never lie', he makes similar claims. In a notorious article, Kant condemns lying even to a would-be murderer who asks where his intended victim is. It is often assumed that, in claiming that we must never lie, Kant states a view that could

not possibly be Act Consequentialist. That is not so. Kant writes that, in telling a lie,

> I bring it about, as far as I can, that statements . . . in general are not believed, and so too that all rights which are based on contracts come to nothing and lose their force, and this is a wrong inflicted upon humanity in general.

And he writes

> Thus a lie . . . always harms another, even if not another individual, nevertheless humanity generally, inasmuch as it makes the source of right unusable.

In these passages Kant condemns all lies by appealing to the harm that these acts bring about. As before, these claims might be made by those Act Consequentialists who reject the Marginalist View and accept the Whole Scheme View. Kant may have believed, like Hume, that each of our acts would do most good if we always followed certain strict rules.

Return next to Kant's claim that everyone's deserved happiness would be best promoted by 'the strictest observance of the moral laws'. Kant often makes such claims. For example, he writes:

> to promote the happiness of others is an end, the means to which I can furnish in no other way than through my own perfection . . .

What Kant calls 'our own perfection' chiefly consists in our having good wills and acting rightly. So Kant here claims that acting rightly is the only way—or, as he may mean, the best way—to promote the happiness of others. Kant also claimed:

> If there is to be a Greatest Good, then happiness and the worthiness thereof must be combined. Now in what does this worthiness consist? In the practical agreement of our actions with the idea of universal happiness. If we conduct ourselves in such a way that, if everyone else so conducted themselves, the greatest happiness would arise, then we have so conducted ourselves as to be worthy of happiness.

Kant here claims that, to be virtuous and act rightly, we must act in the ways which are such that, if everyone acted in these ways, that would produce universal happiness. This claim states one version of a Consequentialist theory: *Hedonistic Rule Utilitarianism*. If the Whole Scheme View and (C) were true, Kant's claim would also state a version of Hedonistic Act Utilitarianism, since these views would coincide.

These claims, however, have only historical importance, since we ought to reject both the Whole Scheme View and (C). Suppose again that, in

> *Rescue*, a hundred miners are trapped underground, with flood-waters rising. These miners will all be saved if four people join some rescue mission. I know that four other people have already joined this mission. I could either join this mission as well, or go elsewhere and save the life of some other single person.

On the Whole Scheme View, I ought to join this mission, since my act will then be one of a set of acts that will together do the most good, by saving a hundred people. That is clearly the wrong conclusion. I ought to save the single person, since one more person's life would then be saved. At least in most cases, we ought to accept the Marginalist View. When we ask which is the act that would do the most good, we ought to ask what *difference* this act would make. Since we ought to accept the Marginalist View, we could not be Act-and-Rule Consequentialists. Consequentialists have to choose between these forms of their view.

According to what I have called Kant's

> Formula of the Greatest Good: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

As I have argued, Kant seems to assume

> K6: It is by following the moral law, as described by Kant's other formulas, that everyone could best promote this ideal world.

On these assumptions, Kant's moral theory has the unity or harmony that Kant claims to be one of the goals of pure reason. Kant's Formula of the Greatest Good describes a single ultimate end or aim that everyone ought always to try to achieve, and Kant's other formulas describe the moral law whose being followed by everyone would best achieve this aim.

In deciding whether we ought to accept these claims, we would have two questions:

> Q1: Ought we always to strive to promote a world of universal virtue and deserved happiness?
>
> Q2: Is it by following Kant's other formulas that we can best promote this ideal world?

We cannot yet try to answer Q2, since we have not yet considered what is implied by Kant's other main formula, his Formula of Universal Law.

Though we might try to answer Q1, I shall not do that. I shall, however, discuss one of Kant's assumptions about his ideal world. It is sometimes said that Kant's claims about the Greatest Good add nothing to the rest of his moral theory. Kant claims elsewhere that we have two ends that are also duties, our own virtue and the happiness of others. But in describing his ideal world, Kant adds that happiness is good only when it is *deserved*. On Kant's view, it would be bad if people had more happiness, or less suffering, than they deserve. These claims about desert cannot be plausibly derived from, or claimed to be supported by, Kant's other formulas. Nor does Kant try to support these claims in this way. He simply asserts these claims, or takes them to be obvious, as when he writes:

> Reason does not approve happiness . . . except insofar as it is united with worthiness to be happy, that is, with moral conduct.

Kant's claims about desert are, I believe, false. And as I shall now argue, Kant came close to seeing that.

11

Free Will and Desert

38 The Freedom that Morality Requires

According to *determinists*, all events are causally inevitable, so that, whenever we act in some way, it would have been causally impossible for us to have acted differently. Kant claims that, if determinism were true, morality would be undermined, since we wouldn't have the kind of freedom that morality requires. And Kant believes that, in one way, determinism is true. But determinism is not, he claims, the whole truth. Kant distinguishes between the spatio-temporal *phenomenal* world, or reality as it appears to us to be, and the world of *noumena*, or things-in-themselves, which is reality as it really is. In this noumenal world, Kant argues, there is neither space nor time. It is conceivable that, as well as being phenomenal beings in the spatio-temporal world, we are also noumenal beings in this other world. Though our acts are partly events which occur in time in the spatio-temporal world, these acts might have undetermined origins in the timeless noumenal world. That, Kant claims, would give us the freedom that morality requires.

Kant also argues that we have such freedom. Kant's argument can be stated as follows:

> (A) Our acts cannot be wrong unless we ought to have acted differently.
>
> (B) 'Ought' implies 'can'. We ought to have acted differently only if we could have acted differently.

Therefore

(C) Our acts cannot be wrong unless we could have acted differently.

(D) If our acts were merely events in the spatio-temporal world, these acts would be causally determined, so it would never be true that we could have acted differently.

Therefore

(E) If our acts were merely such events, none of our acts could be wrong, so morality would be an illusion.

(F) Morality is not an illusion. We ought to act in certain ways, and some of our acts are wrong.

Therefore

(G) Our acts are not merely events in the spatio-temporal world.

In considering this argument, we might first object that, if (E) is true, we could not know that (F) was true unless we knew that (G) was true. If morality is an illusion unless our acts are not merely events in the spatio-temporal world, and we don't know whether our acts are merely such events, how could we know that morality is not an illusion? But there might be ways in which, without first knowing that (G) was true, we could rationally believe that morality is not an illusion. This belief might, for example, be implied by some set of religious beliefs that we could rationally accept, and claim to know, as revealed truths.

We should also accept Kant's argument for (C). As Kant assumes, 'ought' implies 'can'. If we could not possibly act in some way—such as saving someone's life by running faster than a cheetah—it cannot be true that we *ought* to act in this way. For some act of ours to be wrong, because we ought to have acted differently, it must be true that we *could* have acted differently. There are, however, conflicting views about the sense in which this must be true. These are conflicting views about the kind of freedom that morality requires.

Suppose that, while I am standing in some field during a thunderstorm, a bolt of lightning narrowly misses me. If I say that I could have been killed, I might be using 'could' in a *categorical* sense. I might mean that, even with conditions just as they actually were, it would have been causally possible for this bolt of lightning to have hit me. If we assume determinism, that is not true, since it was causally inevitable that this lightning struck the ground just where it did. I may instead be using 'could' in a different, *hypothetical* or *iffy* sense. When I say that I could have been killed, I may mean only that, *if* conditions had been in some way slightly different — if, for example, I had been standing a few yards to the West — I would have been killed. Even if we assume determinism, that claim could be true.

We ought to have acted differently, Kant assumes, only if we could have done so in the categorical sense. It must be true that, even given our actual state of mind, it would have been causally possible for us to have chosen to act differently, and to have done so. If it was causally inevitable that we chose and acted as we did, it would not be relevantly true that we could have acted differently. On this view, as (E) claims, determinism is *incompatible* with the kind of freedom that morality requires.

As many writers argue, however, we ought to reject this *incompatibilist* view. Return to the case in which I say, 'You ought to have helped that blind man cross the street', and you say, 'I couldn't have done that'. If I ask 'Why not?', it would not be enough for you to reply, 'Because I didn't want to'. Perhaps you could not have acted differently, in the relevant sense, if you were in the grip of some irresistible desire, or were insane. But most of us are not in these or other such ways unfree. In most cases, for it to be relevantly true we *could* have acted differently, it need only be true that

> (H) we *would* have acted differently *if* we had wanted to, and had chosen to do so.

We can call this the *hypothetical, motivational sense* of 'could'. This sense of 'could' is compatible with determinism. You could have helped the blind man cross the street in the sense that you would have done so *if* you had chosen to do so. It is irrelevant whether, given your actual desires and other mental states, it was causally inevitable that you did not choose to act in this way.

Someone might now object:

> If all of our decisions, choices, and acts are causally inevitable, we would have acted differently only if we had miraculously defied, or broken, the laws of nature. It is pointless to ask whether we ought to have acted in some way that would have required such a miracle.

Such questions, however, can be well worth asking. What we do often depends on our beliefs about what we ought to do. And if we come to believe that some act of ours was wrong, or irrational, because we ought to have acted differently, this belief may lead us to try to change ourselves, or our situation, so that we do not act wrongly, or irrationally, in this kind of way again. These changes in us or our situation may affect what we later do. It does not matter that, for us to have acted differently in the *past*, we would have had to perform some miracle. If we come to believe that we ought to have acted differently, this change in our beliefs may cause it to be true that in similar cases, without any miracle, we *do* in the *future* act differently. That is enough to make it worth asking whether we ought to have acted differently.

Kant calls this *compatibilist* view 'a wretched subterfuge'. On this view, he claims, we would have only the 'freedom of a turnspit': a mechanical device that, when wound up, turns all by itself. But Kant's objections to compatibilism seem to depend in part on his failure to draw another distinction.

According to *fatalism*, it is inevitable that we shall later act in certain ways, *whatever* we decide to do. All of our different possible decisions would merely be different ways in which we would end up doing the same things. On this view, there is no point in our trying to make good decisions, since that would make no difference to what we later do. Since it is clear that most of our acts *do* depend on our decisions, fatalism is believable only when it is restricted to certain particular acts. According to the Ancient Greek myth, for example, Oedipus was fated, whatever he decided, to kill his father and marry his mother. For this to be true, some Greek god would have had to be ready to intervene,

to ensure that Oedipus's decisions would not have prevented his later acting in these two ways.

Determinism is a quite different view. On this view, what we shall later do will depend on what we decide to do. Though our decisions will be causally inevitable, we often don't know in advance, and could not possibly always know, what we shall later decide to do. And if we make better decisions, and act upon them, things will be likely to go better. These facts are enough to give us reasons to try to make good decisions. If we believed that there was no point in trying to make good decisions, we would be mistakenly slipping back into fatalism, by assuming that our decisions would make no difference to what happens.

Kant sometimes makes this mistake, as when he writes:

> unless we think of our will as free this imperative is impossible and absurd and what is left for us is only to await and observe what sort of decisions God will effect in us by means of natural causes, but not what we can and ought to do of ourselves, as authors.

These remarks imply that, if determinism is true, there would be no point in our trying to decide what we ought to do. We would have to be *passive*, waiting to see what sort of decisions we shall be caused to make. That is not so. Even if determinism is true, we can be *active*, by trying to make and to act upon good decisions. If we are in some burning building, for example, we might try to decide how we can escape. If we merely wait and see what decision we shall later be caused to make, we shall be likely to make a worse decision, and be more likely to die.

Kant also suggests a different, compatibilist view. He writes:

> the practical concept of freedom has nothing to do with the speculative concept . . . For I can be quite indifferent as to the origin of my state in which I am now to act, I ask only what I now have to do, and then freedom is a necessary practical proposition.

Kant seems here to see that, when we are deciding what to do, we can ignore the speculative or theoretical question of whether determinism is

true. If we don't yet know what we shall decide, we are free in the sense that nothing will stop us from acting in certain ways, *if* we decide to do so. For practical purposes, this compatibilist kind of freedom is all we need. It is irrelevant whether, given our actual state of mind, some other decision would have been causally impossible.

Though Kant sometimes suggests that, for practical purposes, the freedom that we need is compatible with determinism, his dominant view is clearly incompatibilist. Kant even claims that noumenal causeless freedom is the keystone of his entire philosophy. He would not have made that claim if he had accepted this compatibilist view.

According to the argument that we have been discussing, more briefly stated:

> (A) to (E): If our acts were merely events in time, these acts would be causally determined, and morality would be an illusion, since we would not have the kind of freedom that morality requires.
>
> (F) Morality is not an illusion.

Therefore

> (G) Our acts are not merely events in time.

We ought, I have claimed, to reject the reasoning that is summed up in (A) to (E). For some act of ours to be wrong, because we ought to have acted differently, it must be true that we *could* have acted differently. But the relevant sense of 'could' is the hypothetical, motivational sense. And this sense of 'could' is compatible with determinism. Even if our acts are causally determined, we could have the kind of freedom that morality requires.

39 Why We Cannot Deserve to Suffer

There is, however, another kind of compatibilism that Kant rightly rejects. Some of Kant's claims suggest this argument:

(I) For it to be true that some act of ours was wrong, we must be morally responsible for this wrong act in some way that could make us deserve to suffer.

(J) If our acts were merely events in time, we could never be responsible for these acts in this suffering-deserving way.

Therefore

(E) If our acts were merely events in time, none of our acts could be wrong, so morality would be an illusion.

(F) Morality is not an illusion.

Therefore

(G) Our acts are not merely events in time.

Premise (I) may seem plausible. There are some people whom no one believes to be morally responsible for their acts in some way that could make them deserve to suffer. That is true, for example, of young children, and some people who are insane. As well as believing that these people are not in this way responsible for their acts, we may believe that, for this reason, they cannot act wrongly.

There is a better way to explain why these people cannot act wrongly. Young children and these insane people cannot have or act upon beliefs about which acts are wrong. But ordinary sane adults can have and act on such beliefs. That is enough to justify our belief that most people are moral agents, whose acts can be right or wrong. So we should reject Kant's assumption that, for us to be moral agents, we must be responsible for our acts in some way that could make us deserve to suffer. We can coherently believe both that our acts can be right or wrong, and that no one could deserve to suffer.

According to premise (J), if our acts were merely events in time, we could not be responsible for our acts in this suffering-deserving way. This part of Kant's view is, I believe, a profound truth. We can be morally responsible in several other ways, or senses, but no one could ever be responsible, I believe, in any way that could make them deserve to suffer. Nor, I believe, could anyone deserve to be less happy.

Of Kant's reasons for assuming (J), one is his belief that

> (K) if our acts were merely events in time, these acts would be causally determined,

and that

> (L) if our acts were causally determined, we could never be responsible for these acts in some way that could make us deserve to suffer.

The kind of freedom that morality requires is, I have claimed, compatible with determinism. We could have acted differently, in the relevant sense, when nothing stopped us from acting differently except our desires or other motives. As Kant assumes, however, this kind of freedom is *not* enough to justify the belief that we can deserve to suffer for what we did. Kant here rightly rejects what we can call *compatibilism about desert*.

Of the other people who reject this view, some would reject Kant's claim that, if our acts were merely events in time, these acts would all be causally determined. Most physicists now believe that determinism is not true, since some events that involve sub-atomic particles are partly uncaused, or random. Such claims may not apply to our decisions to act, and to other mental events. Most neuroscientists believe that mental events consist in, or causally depend upon, physical events in our brains which *are* fully causally determined, because these events occur on too large a scale to be affected by random events at the level of sub-atomic particles. But some people reject this view, believing that some of our decisions are not fully causally determined. Of those who have this belief, some appeal to randomness at the sub-atomic level. Others are *interactionist dualists*, who believe that mental events do not either consist in, or fully causally depend upon, physical events in our brains.

To justify the belief that we can deserve to suffer, it is not enough to defend the claim that our decisions to act in certain ways are not fully caused. If that is all we claim about any such decision, this would be, in Kant's phrase,

> tantamount to handing it over to blind chance.

On this view, we would have the freedom not of a turnspit, whose movement is causally inevitable, but of a sub-atomic particle, whose movement is random. We could not deserve to suffer when and because some of the matter in our brains moved or changed in certain random ways. Nor would it help if, as some dualists claim, our decisions are non-physical events that are partly random.

Many people have claimed that, though *most* events must be either fully caused or partly random, that may not be true of our decisions and acts. These people try to describe some third possibility. Some of these people appeal to our rationality. When we claim that someone acted *for some reason*, these people suggest, we are not claiming that this person's act was fully caused, nor are we claiming that this act was partly random. Our ability to act for reasons may thus seem to provide a third alternative.

When someone acts for some reason, however, we can ask *why* this person acted for this reason. In some cases, the answer is given by some further reason. My reason for telling some lie, for example, may have been to conceal my identity, and my reason for concealing my identity may have been to avoid being accused of some crime. But we shall soon reach the beginning of any such chain of motivating reasons. My ultimate reason for telling my lie may have been to avoid being punished for my crime. When we reach someone's ultimate reason for acting in some way, we can ask why this person acted for this reason, rather than acting in some other way for some other reason. If I had a self-interested reason to try to avoid being punished, and a moral reason not to tell this lie, why did one of these reasons weigh more heavily with me, so that I chose to act as I did? *This* event did not occur for some further motivating reason. So the suggested third alternative here disappears. This event was either fully caused or partly random. And there is always such an event at the start of any chain of motivating reasons. Since our decisions to act as we do all involve such events, there is no coherent third alternative.

To avoid this argument, some people claim that acts can be caused by *agents* in a way that does not involve any *event*. Such believers in *agent-causation* partly accept Kant's view that, if our acts were merely events in time, we could not have any kind of freedom that could make it true that we can deserve to suffer because of what we did. But these

writers believe that, as agents, we are fully part of the spatio-temporal world, so they cannot intelligibly claim that the causing of acts by agents are *not* events.

Kant makes some other relevant claims. To be responsible for our acts, Kant assumes, we must be responsible for our own character. In his words:

> The human being must make or have made *himself* into whatever he is . . . in a moral sense, good or evil. Either condition must be an effect of his free choice . . .

And Kant writes of

> a man's character, which he himself creates,

and of

> a person who is his own originator.

Aristotle similarly writes:

> thus it was open at the beginning to the unjust and the self-indulgent man not to become like that, and so they are voluntarily as they are: but when they have become so, it is no longer possible for them not to be so.

But Aristotle does not ask what could have happened 'at the beginning', when someone chose to make himself unjust or self-indulgent. Kant asks that question, and rightly claims that, if we are merely beings in the spatio-temporal world, we cannot have freely created our own character, thereby freely choosing to be either good or evil.

With the claims just quoted, and some other similar claims, Kant suggests another argument for his belief that our acts are not merely events in time. This argument is, in part:

> (M) What we decide to do depends on our character, and on certain other facts about what we are like, or *how we are*.

Therefore

(N) To be responsible for our acts in some way that could make us deserve to suffer, we must be responsible for being in the relevant ways how we are.

(O) If our acts were merely events in time, we could not be responsible for being how we are unless we acted *earlier* in ways that made us how we are.

(P) To have been responsible for these earlier acts, we must have been responsible for how we *then* were, by having acted even earlier in ways that made us how we then were.

To have been responsible for these earlier acts, we must have been responsible for how we then were, by having acted even earlier in ways that made us how we then were.

To have been responsible for these earlier acts *etc*. . . . *and so on to infinity*.

(Q) We could *not* have been responsible for such an infinite series of character-forming acts.

Therefore

(J) If our acts are merely events in time, we cannot have chosen our own character, or be responsible for our acts in any way that could make us deserve to suffer.

This part of Kant's argument is valid, and has, I believe, true premises. So we ought to accept (J).

Kant's argument continues:

(R) We *are* responsible for our acts in a way that can make us deserve to suffer.

Therefore

> (S) Our acts are not merely events in time. We are responsible for our acts because, in the timeless noumenal world, we freely choose to give ourselves our character, and to act as we do.

When other writers try to describe some third alternative to some act's being fully caused, or partly random, it is a decisive objection to such claims that they are incomprehensible. Compared with such claims, Kant's appeal to our noumenal timeless freedom is in one way easier to defend. We should not expect, Kant claims, to understand this noumenal timeless world. All we can expect to understand is the spatio-temporal phenomenal world. In Kant's words, though such noumenal freedom is incomprehensible, we can at least 'comprehend its incomprehensibility'.

This is not, I believe, a sufficient defence of Kant's view. We can vaguely understand how some part of reality might be timeless. And we can make some sense of the idea that all the features of the spatio-temporal world may, in some non-temporal way, depend on something that vaguely resembles a decision. Such claims may make some sense when applied to God. But some of Kant's claims about our timeless freedom are not even vaguely intelligible. On Kant's view, for example, though everything that happens in the spatio-temporal world is fully causally determined, everything that happens is also in part jointly brought about by a vast number of free and separate decisions, made timelessly, by all of the rational beings who ever live. It is inconceivable that so many free decisions, some of them good and others bad, could all select and bring about parts of the same single wholly determined sequence of events which is the entire history of the spatio-temporal world. And since these decisions would in part determine which rational beings ever exist, these beings must somehow bring it about that they themselves exist. It is not enough to say that we can at least understand why such claims are incomprehensible. We can understand that such claims could not possibly be true.

According to the argument that we are now discussing:

> (J) If our acts were merely events in time, we could never deserve to suffer.
>
> (R) We can deserve to suffer.

Therefore

> (S) Our acts are not merely events in time.

We ought, I have claimed, to reject this argument's conclusion. Our acts *are* merely events in time. Since this argument is valid, and we ought to reject its conclusion, we must reject one of its premises.

Some people would reject (J). There are people who believe that, though our wrong acts are merely events in time, and are causally inevitable, we could deserve to be sent by God to suffer in Hell. On such views, to deserve to suffer, we don't have to have any kind of contra-causal freedom, or to be in any way responsible for our own character, or for being as we are.

Of those who make such claims, some admit that they cannot understand how such claims could be true. God's justice, these people claim, is incomprehensible. Compared with Kant's claim that we should not expect to understand the timeless noumenal world, it is less plausible to claim that we should not expect to understand how we could deserve to suffer. We have no reason to expect such moral truths to be incomprehensible.

Rather than rejecting (J), we ought, I believe, to reject (R). Kant rightly claims that

> (J) if our acts were merely events in time, we could not deserve to suffer.

We can add

> (T) Our acts *are* merely events in time.

Therefore

> (U) We cannot deserve to suffer.

Kant, I have said, came close to seeing the truth of (U). Kant believed that

> (V) we could not deserve to suffer if our acts were either all causally inevitable, or were subject to blind chance, and we were not responsible for our own character.

These things *would* be true, Kant believed, if our acts were merely events in time. If Kant had lost his belief in our noumenal freedom, and come to believe that all our acts *are* merely events in time, he might have continued to believe (V), and drawn the conclusion that we cannot deserve to suffer. But I cannot claim to know that Kant would have drawn this conclusion. Kant might instead have ceased to believe (V), concluding that we *can* deserve to suffer even if our acts *are* causally inevitable, or are subject to blind chance, and we are not responsible for being as we are. I can merely hope that Kant would have continued to believe (V), and would have therefore seen that we cannot deserve to suffer.

Of those who believe that we can deserve to suffer, some would give this counter-argument:

> (W) God makes some people suffer in Hell.
>
> (X) God is just.

Therefore

> (R) We can deserve to suffer.

But we don't, I believe, know that (W) is true. If we believe in a just God, we must accept either

> (Y) God acts justly in making wrongdoers suffer in Hell, though it is unintelligible how such acts can be just,

or

> (Z) God does not make anyone suffer in Hell.

Of these two claims, we would have more reason, I believe, to accept (Z). If God does not make anyone suffer in Hell, it may be surprising that so many people have believed that God *does* act in this way. But we can understand how these people might have come to have this false

belief, and we cannot understand how a just God could make anyone suffer in Hell.

We can deserve many things, such as gratitude, praise, and the kind of blame that is merely moral dispraise. But no one could ever deserve to suffer. For similar reasons, I believe, no one could deserve to be less happy. When people treat us or others wrongly, we can justifiably be indignant. And we can have reasons to want these people to understand the wrongness of their acts, even though that would make them feel very badly about what they have done. But these reasons are like our reasons to want people to grieve when those whom they love have died. We cannot justifiably have ill will towards these wrong-doers, wishing things to go badly for them. Nor can we justifiably cease to have good will towards them, by ceasing to wish things to go well for them. We could at most be justified in ceasing to like these people, and trying, in morally acceptable ways, to have nothing to do with them.

If Kant had seen that no one could deserve to suffer, or to be less happy, his ideal would still have been a world in which we were all virtuous and happy. But he would have changed his view about less than ideal worlds, since he would have ceased to believe that it would be bad if some people suffered less, or were happier, than they deserved.

Though Kant makes various other claims about his ideal world, these are not the most valuable parts of Kant's moral theory. Many other writers claim that the two greatest goods are virtue and happiness. And Kant says little to defend his assumption that, if we follow his other formulas, we shall be doing what will best promote his ideal world. What is most valuable are some of the parts of Kant's theory that are not in these ways Consequentialist. We have considered Kant's Formula of Humanity, and his related claims that to treat people as ends, we must treat them only in ways to which they could rationally consent, and must never treat them merely as a means. We can now turn to Kant's other main statement of his supreme principle: the Formula of Universal Law. Though many people have discussed this formula, none, I believe, has fully seen what Herman calls the 'untapped theoretical power and fertility of this alternative to Consequentialist reasoning'.

PART THREE

THEORIES

This page intentionally left blank

12

Universal laws

40 The Impossibility Formula

Whether our acts are right or wrong, Kant claims, depends on our *maxims*, by which Kant usually means our policies and their underlying aims. Some of Kant's examples are: 'Increase my wealth by every safe means', 'Let no insult pass unavenged', 'Make lying promises when that would benefit me', 'Give no help to those who are in need', and 'the maxim of self-love, or one's own happiness'.

According to one of Kant's versions of his Formula of Universal Law, which we can call

> *the Impossibility Formula*: It is wrong to act on any maxim that could not be a universal law.

This formula needs to be explained. In one passage, Kant refers to a maxim's being 'a universal permissive law'. This may suggest that Kant means

> (A) It is wrong to act on any maxim if we could not all be permitted to act upon it.

But Kant never appeals to (A). And as I explain in a note, (A) would not be a useful claim.

Some writers suggest that Kant means

> (B) It is wrong to act on any maxim that we could not all *accept*, in the sense of deciding to act upon it.

On this suggestion, Kant's formula would be unreliable. If (B) condemned acting on any maxim that it would be inconceivable, or logically impossible, for all of us to accept, this formula would fail to condemn most wrong acts. We can easily conceive or imagine worlds in which everyone accepts bad maxims, such as the maxim 'Deceive and coerce other people whenever that would benefit me'. Such worlds might be *causally* impossible, because there are some good people who would be psychologically unable to accept these bad maxims. But there are also some bad people who would be psychologically unable to accept some good maxims. So if (B) appealed to such causal impossibility, this formula would mistakenly condemn acting on these good maxims. We might appeal to some other kind of impossibility. But as these remarks suggest, (B) is implausible. We have no reason to believe that whether maxims are good or bad, and whether it is wrong to act upon them, depends on whether everyone could accept them.

Some writers suggest that Kant means

> (C) It is wrong to act on some maxim if it would be impossible for everyone to act upon it.

The word 'everyone' here refers only to the people to whom some maxim applies. The maxim 'Care for my children', for example, applies only to parents.

This formula would also be unreliable, since (C) condemns many morally required or permissible acts. There are many good maxims on which some people could not act, because they do not have the opportunity or ability to act in these ways. Some parents cannot care for their children, because they are in prison, or are mentally ill. But caring for our children is not wrong. To avoid this objection, (C) might condemn acting on any maxim that could not be acted on by everyone who has both the opportunity and the ability to act upon it. But no maxim would fail this test. And (C) is also implausible. We have no since we have no reason to believe that whether maxims are good or bad, and whether it would be *wrong* to act upon them, depends on whether everyone *could* act upon them.

Some writers suggest that Kant means

> (D) It is wrong to act on some maxim if it would be impossible for everyone who could act upon it to act *successfully*, in the sense that they would achieve their aims.

This formula would be no better. There are many maxims on which it would be permissible or good to act, though we could not all successfully act upon them. Some examples are: 'Become a doctor or a lawyer', 'Adopt an orphan', 'Give more to charity than the average person gives', and 'Be the last person to use any fire-escape, or to leave any sinking ship'. If we all tried to achieve these aims, some of us would fail. (D) is also implausible. We have no reason to believe that, if we could not all successfully act on some maxim, it would be wrong for anyone to act upon it. It is not wrong to make attempts some of which we know will fail.

We have been trying to understand Kant's claim that it is wrong to act on maxims that could not be universal laws. (A) to (D) are the most straightforward ways to interpret this claim. But as well as being either unhelpful or both unreliable and implausible, these are not claims to which, when Kant applies his formula, he himself appeals. Though Kant's *stated* Impossibility Formula is

> (E) It is wrong to act on any maxim that could not be a universal law,

Kant's *actual* formula is

> (F) It is wrong to act on any maxim of which it is true that, if everyone accepted and acted on this maxim, or everyone believed that it was permissible to act upon it, that would make it impossible for anyone successfully to act upon it.

Could this formula help us to decide which acts are wrong?

Consider first the maxim 'Kill or injure other people when that would benefit me'. As Herman points out, if we all accepted and acted on this maxim, that would not make it impossible for any such act to

succeed. So (F) does not condemn such acts. Nor does (F) condemn self-interested coercion. If we all tried to coerce other people whenever that would benefit ourselves, some of these acts would succeed.

Turn next to lying. Herman writes that (F)

> seems adequate for maxims of deception . . . Universal deception would be held by Kant to make speech and thus deception impossible.

Korsgaard similarly writes:

> lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive . . .

But no one acts on the maxim 'Always lie'. Many liars act on the maxim 'Lie when that would benefit me'. Kant's formula condemns such acts only if, in a world of self-interested liars, it would be impossible for any such lie to succeed. That would not be impossible. Even in such a world, it would often be in our interests to tell others the truth. And when it would be in our interests to deceive someone, there would often be no point in lying, since this person would not believe our lie. So, even if we were all self-interested liars, many of our statements would be true. Most of us would know this fact. And since we could not always tell which statements by others were lies, some lies would be believed, and would achieve the liar's aim.

To explain why theft is wrong, Kant writes:

> Were it to be a general rule to take away his belongings from everyone, *mine* and *thine* would be altogether at an end. For anything I might take from another, a third party would take from me.

As before, however, no one acts on the maxim 'Always steal'. Many thieves act on the maxim 'Steal when that would benefit me'. If this maxim were universally accepted and acted upon, that would not produce a world in which such acts would never succeed. There would

still be property, which would not always be successfully protected. Thieves would sometimes achieve their aims.

When Kant discusses the maxim 'Let no insult pass unavenged', he claims that, if this maxim were universal, it would be 'inconsistent with itself', and would not 'harmonize with itself'. But if everyone acted on this maxim, that would not make it true that no one could succeed. It might even be true that every insult was avenged, so that *everyone* would succeed.

Kant's actual formula, we have found, fails to condemn many of the acts that are most clearly wrong. This formula does not condemn self-interested killing, injuring, coercing, lying, and stealing.

These failures may suggest that Kant's formula condemns nothing. But we have still to consider Kant's best example: that of someone who makes a lying promise so that he can borrow money that he does not intend to repay. This man acts on the maxim 'Make lying promises when that would benefit me'. Kant claims that, if everyone accepted this maxim, and believed that lying promises are permissible, that would make it impossible for any such promise to succeed. In his words:

> the universality of a law that everyone . . . could promise whatever he pleases with the intention of not keeping it would make the promise . . . impossible, since no one would believe what was promised him but would laugh at all such expressions as vain pretenses.

In assessing this claim, as Rawls suggests, we should ask what would be true after some period that was long enough for everyone's acceptance of the lying promiser's maxim to have its full effects. Kant seems right to claim that, in such a world, no one would be able to benefit themselves by making any lying promise. Not only would such promises not be believed; the social practice of morally motivated, trust-involving promises would have ceased to exist. Kant's formula therefore condemns such lying promises. And most of these acts, we can assume, are wrong.

Now that we have found one kind of wrong act that Kant's formula condemns, we can ask whether this formula is plausible. Kant's formula is, in part:

> (G) It is wrong to act on any maxim of which it is true that, if everyone believed such acts to be permissible, that would make it impossible for any such act to succeed.

This claim condemns those acts whose success depends on other people's refraining from such acts, because they believe such acts to be wrong. And (G) may seem to condemn these acts for a good reason. Lying promisers act wrongly, we might suggest, because if everyone believed such acts to be permissible, that would undermine a valuable social practice.

Kant's claims are not restricted, however, to *valuable* social practices. The soldiers in Hitler's armies, for example, were required to swear oaths of unconditional obedience. Kant condemns lying promises with the claim that, if everyone believed that lying promises were permissible, the practice of making promises would be a 'vain pretense', or sham. Some of these German soldiers rightly believed that it was morally permissible for them, despite having sworn this oath, to disobey all immoral commands. We could similarly claim that if all these soldiers had believed such disobedience to be permissible, the practice of swearing oaths of unconditional obedience would have been a vain pretense or sham. Kant's remarks seem to imply that such disobedience would be wrong. But as Kant himself claims, everyone ought to disobey immoral commands.

For another test of (G), we can suppose that, during the Second World War, some non-Jewish German civilian knows that German Jews are being rounded up and killed. This person successfully acts on the maxim 'Tell lies to the police when that would save some Jewish person's life'. Suppose next that, if everyone had been known to believe that such lies were permissible, that would have made it impossible for anyone to save people's lives in this way. German policemen would have been required to search every building, ignoring anyone's claims that this building contained no Jews. On these assumptions, (G) would have condemned this person's life-saving acts.

Kant might have accepted this conclusion, given his claim that it would be wrong to lie even to a would-be murderer who asks where his intended victim is. But such life-saving lies would be clearly justified. And when applied to this example, (G) is implausible. It would be no objection to this way of saving people's lives that, if everyone believed such acts to be permissible, that would make them impossible.

This imagined case is like Kant's case of a lying promiser. Kant's promiser achieves his aim because there are many people who can be trusted not to make lying promises, given their belief that such promises are wrong. Kant claims that, if everyone was known to believe that such promises are not wrong, that would have made it impossible for anyone to act successfully on this lying promiser's maxim. If that is true, Kant's formula implies that this person's lying promises are wrong. Similar claims apply to my example. My German civilian achieves her aim because there are many people who can be trusted not to lie to the police, given their belief that such lies are wrong. I have supposed that, if everyone was known to believe that such lies are not wrong, that would have made it impossible for anyone to act successfully on this person's life-saving maxim. If that is true, Kant's formula mistakenly implies that this person's life-saving lies were wrong. The important difference between these acts is in what they are intended to achieve; and this difference is ignored by (G).

As this and other such cases show, (G) is unacceptable. As well as failing to condemn nearly all of the acts that are most clearly wrong, (G) condemns some acts that are clearly right. And though (G) correctly condemns lying promises, it condemns these acts for a bad reason.

Kant's formula is also, in part,

> (H) It is wrong to act on any maxim whose being universally accepted and acted upon would make it impossible for anyone successfully to act upon it.

This formula, some writers claim, condemns acting on several good maxims, such as 'Refuse to accept bribes' and 'Give generously to the poor'. If these maxims were universally acted upon, that would soon make it impossible for anyone to act successfully on these maxims, since

no one would offer any bribes, and there would cease to be any poor people. So Kant's formula mistakenly implies that it would be wrong both to refuse bribes and to give generously to the poor.

Korsgaard partly answers this objection. When people act on the maxim of giving to the poor, their aim, Korsgaard suggests, is to abolish poverty. If all rich people acted on these people's maxim, that might abolish poverty, thereby making it impossible for anyone later to act on this maxim. But (H) would not mistakenly condemn these people's acts, because by giving to the poor these people would *achieve* their aim.

These claims do not apply, however, to some rich people. When these people act on the maxim 'Give generously to the poor', their aim is not to abolish poverty but to be admired for their generosity. If all rich people acted on this maxim, their acts might abolish poverty, thereby making it impossible for any of these people to act on their maxim in a way that would achieve their aim. (H) would then mistakenly condemn these people's acts. When these people give large sums to the poor, their acts have no moral worth, but they are not acting wrongly.

Consider next those men who accepted codes of honour, like the code that led the Russian poet Pushkin to fight his fatal duel in the snow. Suppose that Pushkin had accepted the maxim 'Fight duels to show my courage, but always shoot into the sky'. If all these men had accepted and acted on this maxim, the practice of duelling would have become farcical, and would not have survived. That would have made it impossible for Pushkin to act on his maxim in a way that would achieve his aim, so (H) would have condemned Pushkin's acting on this maxim. (H) may seem to give the right answer here, since duelling is wrong. But (H) would *not* have condemned acting on the maxim 'Fight duels to show my courage, and always shoot to kill'. And acting on this second maxim would have been much worse. As this comparison suggests, (H) would have condemned Pushkin's act for a bad reason. It would have been no objection to Pushkin's maxim that, if this maxim were universally accepted, the practice of duelling would not survive. As before, Kant's formula mistakenly ignores the question of whether some social practice is good, and ought to be supported.

For another example, consider the maxim, 'Have no children, so as to have more time and energy to work for the future of humanity'. If

everyone acted on this maxim, that would make it impossible for anyone successfully to act upon it, since humanity would have no future. So (H) mistakenly condemns such acts.

O'Neill proposes a weaker version of (H). Kant's formula, O'Neill suggests, could become

> (I) It is wrong to act on any maxim whose being successfully acted on by some people would prevent some other people from successfully acting on it.

This formula condemns deception and coercion, O'Neill claims, since those who deceive or coerce others thereby 'guarantee that their victims cannot act on the maxims they act on'. But this claim is false. Of those who have been deceived or coerced, most can deceive or coerce other people. O'Neill also claims that, while we are deceiving or coercing people, we 'undercut their agency', thereby preventing them 'for at least some time' from acting successfully in the same way as us. But this claim is also false. Two people can simultaneously deceive each other. And there can be mutual simultaneous coercion. Two wrestlers might simultaneously use force to keep each other on the ground. And I might coerce you by making one credible threat, while you are coercing me by making another. That is how hostile nations with nuclear weapons might deter each other from using these weapons.

O'Neill could reply that, to show that (I) condemns deception and coercion, it is enough to claim that *some* deceivers and coercers prevent *some* of their victims from deceiving or coercing others. This weaker claim is true. O'Neill similarly claims that, if we acted on maxims of 'severe injury', some of us would disable some of our victims, thereby preventing these people from severely injuring others. So (I) condemns some wrong acts. But (I) condemns these acts for a bad reason. What is wrong with deceiving, coercing, and severely injuring others isn't that, by acting in these ways, we prevent some other people from successfully doing the same.

(I), moreover, mistakenly condemns many good or morally permissible acts. There are many good or permissible maxims of which it is true that, if some people successfully acted on them, that would prevent some other people from doing the same. As O'Neill points out,

(I) implies that we act wrongly if we play competitive games with the aim of winning. Though some English schoolboys were told to accept this view, it seems too severe. And there would be nothing wrong with acting on the maxim 'Become a doctor', even if, by applying and being admitted to some medical school, we prevented someone else from being admitted to any medical school. Or consider the maxims 'Discover what killed all the dinosaurs', 'When travelling with others, always carry the heaviest load', and 'Find someone with whom I can happily live my life'. It is not wrong to try to make some discovery, or to carry the heaviest load, even though, if we succeed, we shall make it impossible for some other people to do these things. Nor is it wrong to live happily with the only person with whom someone else could have happily lived.

Korsgaard proposes another version of Kant's Impossibility Formula. What this formula forbids, she suggests, are acts whose success 'depends upon their being exceptional'. This test, she adds, 'reveals unfairness'. But that is not, I believe, true. And this version of Kant's formula also mistakenly condemns many permissible acts. Some poor people get their food by searching through the rubbish that others throw away. That method must be exceptional, but is not wrong, or unfair. It was not wrong for romantic poets to give themselves the experience of being the only human being in some wilderness. Nor is it wrong, or unfair, to use tennis courts when they are least crowded, pay the debts on our credit cards before interest is charged, buy only second-hand books, or give surprise parties.

Though there are other ways in which we might interpret or revise Kant's Impossibility Formula, these possibilities are not worth considering. Of the interpretations and revisions that we have considered, none contains a good idea. There is no useful sense in which we could claim it to be wrong to act on maxims that could not even *be* universal laws.

41 The Law of Nature and Moral Belief Formulas

Kant proposes another, better formula. According to Kant's main statement of his

> *Formula of Universal Law*: It is wrong to act on maxims that we could not *will* to be universal laws.

Kant remarks that, when maxims fail this test, we have unstrict duties not to act upon them. Such duties are *unstrict* in the sense that we are sometimes morally permitted to act on such maxims. We should ignore this remark, as Kant often does. Kant claims that our *strict* duties can be derived from his Impossibility Formula. As we have seen, that is not true. So we should ask whether Kant's Formula of Universal Law can do better, by correctly implying that some kinds of act are always wrong. As Herman points out, it would not be enough if Kant's formula implied that, though it would be wrong to have a *policy* of killing others for our own convenience, such acts are *sometimes* permitted.

When we apply Kant's formula, we suppose or imagine that we have the power to *will*, or choose, that certain things be true. We are doing a *thought-experiment*, which involves comparing different possible states of the world, or what we can call different *possible worlds*. Like the thought-experiments of some scientists, our thoughts about these possible worlds may lead us to conclusions which also apply to the actual world.

When Kant asks whether we could will it to be true that some maxim is a universal law, he sometimes asks whether we could *consistently* will this to be true. He asks, for example, whether our will would *conflict* with itself, or would *contradict* itself. In other passages, Kant seem to ask what we could *rationally* will, or choose. Kant's formula is more likely to succeed if we use 'could will' in this second, wider sense. On some views, this would make no difference, since our choices fail to be rational only when they are inconsistent, or conflict with each other. But as I have argued, for our choices to be rational, we must also respond well to reasons or apparent reasons. We could not rationally choose or will it to be true that some maxim is a universal law if we are aware of facts that give us clearly decisive reasons not to make this choice.

In willing that some maxim be a universal law, what would we be willing? Kant sometimes claims that, when we apply his formula, we should ask whether we could will that our maxim be a 'universal law of

nature', in the sense that everyone would accept and act on this maxim. On this version of Kant's formula, which we can call

> *the Law of Nature Formula*: It is wrong for us to act on some maxim unless we could rationally will it to be true that everyone accepts this maxim, and acts upon it when they can.

As before, the word 'everyone' refers only to the people who might act on some maxim. The maxim 'Give up smoking', for example, applies only to smokers.

In some other passages, Kant appeals to what we can call

> *the Permissibility Formula*: It is wrong for us to act on some maxim unless we could rationally will it to be true that everyone is morally permitted to act on this maxim.

When Kant applies this formula, he assumes that, if everyone were permitted to act on some maxim, at least some people would be more likely to act upon it. This effect would be produced, not by these people's *being* permitted to act on this maxim, but by their *believing* that such acts are permitted. So Kant must also be appealing to what we can call

> *the Moral Belief Formula*: It is wrong for us to act on some maxim unless we could rationally will it to be true that everyone believes that such acts are morally permitted.

Given their similarity, it is not worth using both these formulas. And unlike the Permissibility Formula, as I explain in a note, the Moral Belief Formula can be plausibly used on its own. So we can ignore the Permissibility Formula.

Kant remarks that he is proposing, not a 'new principle', but only a more precise statement of the principle that 'common human reason . . . has always before its eyes'. This remark understates Kant's originality. But Kant's Law of Nature and Moral Belief Formulas develop the ideas that are expressed in two familiar questions: 'What if everyone did that?' and 'What if everyone thought like you?'

When we apply these formulas, we must appeal to some beliefs about rationality and reasons. We might appeal to what Kant himself believed. But that would be difficult, since Kant did not clearly state these beliefs. And we are asking whether Kant's formulas can help us to decide which acts are wrong, and help to explain why these acts are wrong. In asking these questions, we should try to appeal to true beliefs about rationality and reasons. We should therefore appeal to our own beliefs, since we are then appealing to what we believe to be the truest or best view. Though we know that we might be mistaken, we cannot appeal to what *is* true rather than to what we believe to be true.

There are, however, some beliefs to which we should not appeal. First, we should not appeal to our beliefs about which acts are wrong. I am calling these our *deontic beliefs*. Nor should we appeal to the *deontic reasons* that an act's wrongness might provide. When we apply Kant's Law of Nature Formula, it would be pointless to claim both that

> (1) it is wrong to act on a certain maxim because we could not rationally will it to be true that everyone acts on this maxim,

and that

> (2) we could not rationally will it to be true that everyone acts on this maxim because such acts are wrong.

If we combined these claims, that would be like pulling on our boot laces in an attempt to hold ourselves in mid air. To vary the metaphor, we would be going round in a circle, getting nowhere. Kant does not make this mistake. When Kant claims that we could not rationally will it to be true that everyone acts on some bad maxim, he never appeals to his beliefs that such acts are wrong and that we could not rationally will it to be true that everyone acts wrongly. Kant knew that, if he appealed to such beliefs, his Law of Nature Formula would achieve nothing, since this formula could not then help us to reach true beliefs about which acts are wrong, nor could it support these beliefs.

Similar remarks apply to Kant's Moral Belief Formula. It would be pointless to claim both that

> (3) it is wrong to act on a certain maxim because we could not rationally will it to be true that everyone believes such acts to be permitted,

and that

> (4) we could not rationally will it to be true that everyone believes such acts to be permitted because such acts are wrong.

When we ask whether we could rationally will that everyone *believes* some kind of act to be wrong, we should not appeal to our beliefs about whether such acts *are* wrong. As before, when Kant applies this formula, he follows this *Deontic Beliefs Restriction*, making no appeal to such beliefs.

There is another belief to which we should not appeal. Many wrong acts benefit the agent in ways that impose much greater burdens on others. On some views, such acts are irrational, since we are rationally required to give great weight to everyone else's well-being. If we accept such a view, we should ignore it when we apply Kant's formulas. The main idea behind Kant's Law of Nature Formula is that, even if wrong-doers could rationally act on certain bad maxims, they could not rationally will it to be true that *everyone* acts on their maxims. When we apply this idea, it would be irrelevant to claim that, because these people are rationally required to give great weight to other people's well-being, they could not even rationally will it to be true that *they themselves* act on their maxims.

As before, Kant does not make such claims. When Kant discusses a rich and self-reliant man who has the maxim of not helping others who are in need, Kant does not appeal to the belief that this man is rationally required to give such help. As Rawls and Herman suggest, when we apply Kant's formulas to people who act on such maxims, we should suppose that these people's maxims and acts are both rational. We can add that, if we combine Kant's formulas with less controversial and more widely accepted assumptions about rationality and reasons, these formulas would, if they succeed, achieve more.

42 The Agent's Maxim

Whether some act is wrong, Kant's formulas assume, depends on the *agent's* maxim. Of the maxims that Kant discusses, most involve some *policy*, which could be acted on in several cases. Two maxims may be different, though they involve the same policy, because they involve different underlying motives or aims. Two merchants, for example, may both act on the policy 'Never cheat my customers'. But these merchants act on different maxims if one of them never cheats his customers because he believes this to be his duty, while the other's motive is to preserve his reputation and his profits.

Kant's appeal to the agent's maxim raises various problems. Let us call some maxim

> *universal* when everyone both acts on this maxim whenever they can, and believes such acts to be permitted.

Suppose that I wrongly steal some wallet from some woman dressed in white who is eating strawberries while reading the last page of Spinoza's *Ethics*. My maxim is to act in precisely this way, whenever I can. I could rationally will it to be true that this maxim is universal, because it would be most unlikely that anyone else would ever be able to act in precisely this way, so this maxim's being universal would be most unlikely to make any difference. Since I could rationally will this maxim to be universal, Kant's formulas mistakenly permit my act. Similar claims apply to other highly specific maxims. When wrong-doers act on such maxims, they could rationally will that their maxims be universal, because they would know that other such acts would be rare, and would therefore make little difference. Kant's formulas would mistakenly permit these wrong acts. We can call this the *Rarity Objection*.

This objection can be partly answered. Just as it is a factual question what someone believes, or wants, or intends, it is a factual question on which maxim someone is acting. And real people seldom act on such highly specific maxims. When we describe someone's maxim, as O'Neill and others claim, we should not include any details whose absence

would have made no difference to this person's decision to do whatever he is doing. In a realistic version of my example, I would have stolen from my victim even if if she had been dressed in red, or had been eating blueberries, or had been reading the first page of *Right Ho Jeeves!* My real maxim would be something like 'Steal when that would benefit me'. This may *not* be a maxim that I could rationally will to be universal. Kant's formulas would then correctly imply that my act is wrong.

These remarks do not fully answer the Rarity Objection. Even if actual wrong-doers never acted on such highly specific maxims, we can imagine such people. Kant's formulas ought to be able to condemn these imagined people's acts. And as we shall see, this objection applies to some actual cases.

Kant's appeal to the agent's maxim raises other, more serious problems. Consider some man who often acts on

> *the Egoistic Maxim*: Do whatever would be best for me.

This man, we can plausibly assume, could not rationally will it to be true either that everyone always acts on this maxim, or that everyone believes that all such acts are morally permitted. Most Egoists could not rationally choose to live in a world of Egoists, since that would be much worse for them than a world in which everyone accepts various moral maxims. Since this Egoist could not rationally will that his maxim be universal, Kant's formulas imply that, whenever he acts on his maxim, his act is wrong. This man acts wrongly not only when he steals and lies, but also when, for self-interested reasons, he pays his debts, keeps his promises, and saves a drowning child, because he hopes to get some reward. These are unacceptable conclusions. When this Egoist acts in these ways, his acts have no moral worth. But these acts are not wrong.

It might be claimed that, when this man acts in any of these ways, *what* he is doing is not wrong, but *his doing* of it is. Kant suggests a similar distinction when he claims that, to fulfil some *duties of virtue*, we must not only act rightly, but also act with the right motive. On Kant's view, Rawls claims, even if we do not kill ourselves, we may have failed to fulfil our duty not to kill ourselves. To fulfil this duty, we must

refrain from killing ourselves for the right reason. Kant similarly claims that to fulfil a duty of gratitude, we must feel grateful.

These distinctions cannot answer this objection to Kant's formulas. My Egoist may never fulfil his duties of virtue, since he may never have the right motive. As Kant claims, however, we also have many *duties of justice*, which we can fulfil by doing what is morally required, whatever our motive. One example is our duty to pay our debts. Kant's prudent merchant would do his duty if he acted on the maxim 'Pay my debts', even if this merchant's only motive was to preserve his reputation and his profits. Kant's formula gives the right answer here, since this merchant would be acting on a maxim that he could rationally will to be universal. But when my Egoist pays his debts, he is acting on his Egoistic maxim, which he could *not* rationally will to be universal. So Kant's formulas mistakenly imply that, when this man pays his debts, he is *not* doing his duty, but is acting wrongly.

Return now to the drowning child. Suppose that, because this child has fallen into some fastly flowing river near some deep waterfall, any attempt to save this child would be too risky to be anyone's duty. If some good person saved this child, despite these risks, this person would be heroically acting beyond the call of duty. My Egoist decides to take these risks, since he could then hope to get a greater reward. Acting on his maxim, he dives into the river. On the suggestion we are now considering, if this man saves this child's life at this great risk to his own life, what he is doing is not wrong, but his doing of it is. That is clearly false. This man is not failing to fulfil any duty, or acting wrongly in any sense.

Turn next to prudent acts which affect no one else. When this Egoist takes some medicine, or puts on warmer clothing, he may be acting on his maxim 'Do whatever would be best for me'. Since this man could not will that this maxim be universal, Kant's formulas again mistakenly imply that he is acting wrongly. Nor could we claim that, though *what* he is doing is not wrong, his *doing* of it is. There is no sense in which, when this man puts on warmer clothing, his acting in this way is wrong.

Some writers suggest that we should not apply Kant's formulas to maxims that are as general as 'Do whatever would be best for me'. But Kant often discusses this Egoistic maxim, which he calls 'the maxim of

self-love, or one's own happiness'. And if we claimed that such maxims are too general, we would be ignoring many people's actual maxims. Kant discusses the maxim 'Make a lying promise when that would benefit me'. There are other, similar maxims, such maxims of stealing, cheating, or breaking the law whenever that would be best for ourselves. Since these maxims all involve the same more general policy, they are unnecessary clutter, and could all be replaced by the single maxim 'Do whatever would be best for me'. When many actual people act on this Egoistic maxim, or policy, it may be simply false to claim that these people also accept, and are acting upon, on any other, less general policy.

For examples of a different kind, we can turn to conscientious people who have false moral beliefs. One example could be Kant himself during the period in which, as some of his remarks suggest and we can here suppose, Kant accepted the maxim 'Never lie'. This maxim is condemned by Kant's formulas. Kant could not have rationally willed it to be true that no one ever tells a lie, not even to a would-be murderer who asks where his intended victim is. Nor could he have rationally willed it to be true that everyone believes these life-saving lies to be wrong. So Kant's formulas imply that, whenever Kant acted on this maxim by telling anyone the truth, his act was wrong. He acted wrongly even when he told someone the correct time of day. That is clearly false. Similar claims would apply to people who accept the maxims 'Never steal' and 'Never break the law'. These people could not rationally will it to be true that no one ever steals or breaks the law, not even when such acts were the only way to save some innocent person's life. So Kant's formulas imply that, whenever these people act on these maxims, by returning someone's property or keeping some law, they act wrongly. These implications are also clearly false.

Our problem can be redescribed as follows. Some maxims are *wholly bad*, or *wholly good*, in the sense that it is always wrong, or always right, to act upon them. Two examples are the maxims 'Torture others for my own amusement' and 'Prevent pointless suffering'. When applied to such maxims, Kant's formulas succeed. But many maxims are

> *morally mixed* in the sense that, if we always acted on these maxims, some of our acts would be wrong, but other acts would be permissible or even morally required.

Two examples are the Egoistic maxim and Kant's maxim 'Never lie'. In proposing his formulas, Kant overlooks such mixed maxims. Kant's formulas assume that acting on some maxim is either always wrong, or never wrong. When applied to mixed maxims, Kant's formulas fail, since these formulas condemn some acts that are clearly permissible or morally required. When my Egoist prudently pays his debts, and Kant tells most people the truth, they are not acting wrongly, as Kant's formulas mistakenly imply. We can call this the *Mixed Maxims Objection*.

After considering this and other objections to Kant's Formula of Universal Law, in either its law of nature or moral belief versions, some writers conclude that we cannot use Kant's formula to help us to decide which acts are wrong. Wood claims that, when used as such a criterion, Kant's formula is 'radically defective' and 'pretty worthless'. Herman claims that, despite a 'sad history of attempts . . . no one has been able to make it work'. O'Neill suggests that, in some cases, Kant's formula may 'give either unacceptable guidance or none at all'. Hill doubts whether, when used on its own, Kant's formula can provide 'even a loose and partial action guide'.

Because these people believe that Kant's formula cannot provide a criterion of wrongness, some of them suggest that Kant was not trying to provide such a criterion. Kant's formula, Herman suggests, may be intended only to show that there is a 'deliberative presumption' against acting in certain ways for certain reasons. O'Neill suggests that Kant's formula may be intended to provide a test, not of which acts are wrong, but only of which acts have moral worth.

Kant, I believe, had more ambitious aims. Our acts are in one sense right or wrong when, in Kant's words, these acts *conform with duty* or are *contrary to duty*. This is the sense of 'right' and 'wrong' with which

Kant's formula is concerned. While discussing or applying his formula, Kant writes:

> to inform myself in the shortest and yet infallible way . . . whether a lying promise is in conformity with duty, I ask myself: would I indeed be content that my maxim . . . should hold as a universal law?
>
> someone feels sick of life . . . but [asks] himself whether it would not be contrary to his duty to himself to take his own life.
>
> he still has enough conscience to ask himself, is it not forbidden and contrary to duty?
>
> he asks himself whether his maxim of neglecting his natural gifts . . . is consistent with what one calls duty.

Kant also claims that his formula

> determines quite precisely what is to be done . . . with respect to all duty in general,

and that

> common human reason, with this compass in hand, knows very well how to distinguish in every case what is good and what is evil, what conforms with duty or is contrary to duty.

These last claims are overstatements. But so, I believe, are the claims that, as a criterion of wrongness, Kant's formula is worthless, and cannot be made to work. Kant's formula *can* be made to work. When revised in some wholly Kantian ways, this formula is, I shall argue, remarkably successful.

In asking how we should revise our two versions of Kant's formula, we can first restate the Mixed Maxims Objection. To judge whether some act is wrong, we must know all of the facts that are, or might be, *morally relevant*. It is not enough to know, for example, that some man moved one of his fingers, or that, in moving this finger, this man pulled the trigger of some gun, or that he thereby killed someone. We must know

some other facts, such as whether this man was intending to kill this other person, and, if so, whether he was acting in self-defence, and, if so, whether he was defending himself while attacking someone else.

Of the maxims that Kant discusses, as I have said, most involve some *policy* which could be acted on in several cases. Kant's formula assumes that, to judge whether someone's act is wrong, it is enough to know on which policy this person is acting. That is sometimes true. It would be enough to know that someone is acting on the policy 'Torture others for my own amusement'. But in many other cases Kant's assumption fails. If all we know is that my Egoist is acting on the policy 'Do whatever would be best for me', we cannot possibly decide whether this man is acting wrongly. We don't know whether this man is killing someone, saving someone's life, stealing, paying some debt, or putting on warmer clothing. And if all we know is that Kant has acted on the policy 'Never lie', we don't know whether Kant has told some would-be murderer where his intended victim is, or has merely told someone the correct time of day. As these examples show, if all we know is the policy on which someone is acting, we often don't know all of what might be the morally relevant facts.

There is another problem. When we ask whether some act is wrong, or contrary to duty, Kant's formula often makes the answer depend on morally *irrelevant* facts. When my Egoist risks his life to save some drowning child, it is irrelevant that he is acting on the policy of doing whatever would be best for himself. When Kant told someone the correct time, it was irrelevant that he was acting on the policy 'Never lie'. These facts at most give us reasons to believe that in some *other* cases this Egoist would, and Kant might act, wrongly.

For Kant's formula to succeed, it would have to be true that there are no maxims or policies on which it would be sometimes but not always wrong to act. That is obviously false. So Kant's formula should not appeal to the agent's maxim, in the sense of 'maxim' that can refer to policies.

Some writers suggest that, rather than appealing to the agent's actual maxim, Kant's formula should appeal to the possible maxims on which

the agent might have acted. In its law of nature version, Kant's formula might then become

> LN2: We act wrongly unless what we are doing is something that we could have done while acting on some maxim on which we could rationally will everyone to act.

This formula avoids the Mixed Maxims Objection. When my Egoist saves the drowning child, and Kant tells most people the truth, they might have been acting on maxims on which they could rationally will everyone to act. But if we appeal to LN2, we lose our partial answer to the Rarity Objection. Return to the case in which I wrongly steal from a white-dress-wearing strawberry-eating woman. What I am doing is something that I could have done while acting on a maxim of stealing from white-dress-wearing strawberry-eating women, whenever I can. I could rationally will it to be true that everyone acts on this maxim, since such acts would at most be very rare. So LN2 mistakenly permits my act. Similar claims apply to other cases. When people act wrongly, there is always some possible maxim on which these people *might* have been acting which they could have rationally willed to be universal. So LN2 fails to condemn all wrong acts.

To avoid this objection, we can revise Kant's formulas in a simpler way. Kant's Law of Nature Formula can become

> LN3: We act wrongly unless we are doing something that we could rationally will everyone to do, in similar circumstances, if they can.

Kant's Moral Belief Formula can become

> MB2: We act wrongly unless we could rationally will it to be true that everyone believes such acts to be morally permitted.

These formulas avoid the Mixed Maxims Objection. When my Egoist saves the drowning child, and Kant tells someone the correct time, they could rationally will it to be true both that everyone acts in these ways, and that everyone believes such acts to be permitted. So these formulas do not mistakenly condemn these acts.

These revised formulas also avoid the Rarity Objection. When we apply these formulas to someone's act, we must describe this person's act in the morally relevant way. Suppose that, being a whimsical kleptomaniac, I really *am* acting on the maxim of stealing from white-dress-wearing strawberry-eating women, whenever I can. This maxim does not provide the morally relevant description of my act. It is irrelevant that I am stealing from someone who is a woman, and who is wearing white and eating strawberries. The relevant facts may be that I am stealing from someone who is no richer than me, merely for my own amusement. In applying these revised formulas, we should ask whether I could rationally will it to be true that everyone acts in this way, and that everyone believes such acts to be permitted. If the answer is No, as we can plausibly claim, these revised formulas would rightly condemn my act.

In many cases, to give the morally relevant description of some act, it is enough to describe what the agent is, or would be, *intentionally doing*. We must describe this person's immediate aims, or what this person is directly trying to achieve. We should also describe the effects which this person believes that his or her acts might have. What people *intentionally do* is not the same as what they *intend*. To give Sidgwick's example, if some Russian revolutionary in the late nineteenth century blows up the train on which the Czar is travelling, this man may be intending only to kill the Czar. But what this man is intentionally doing is blowing up this train knowing that, as well as killing the Czar, he will kill many other people.

When we describe people's acts, we are usually describing what these people are intentionally doing. It is sometimes unclear what is the morally relevant description of some act. It may be unclear, for example, how much we ought to include in our list of some act's foreseeable effects, or what we ought to describe as separate acts or as parts of a single complex act. And to decide whether some act is wrong, we sometimes need to know not only *what* someone is intentionally doing, but also *why* this person does what he or she is doing. To illustrate both these points, we can suppose that some sadist saves someone's life so that he can then kill this person in a more painful way. It may not be enough to claim that what this sadist is intentionally doing is saving someone's life.

When it is unclear whether some fact is morally relevant, it often does no harm to include this fact in our description of some act. But when we apply certain moral principles to some act, it can be important not to include morally irrelevant facts. To apply both LN3 and MB2, as I have said, we must give the right description of what people are doing. Similar claims apply to some other moral principles, such as principles about the wrongness of lying, stealing, and breaking promises. It is sometimes unclear which acts should be regarded as being of these kinds. But we need not answer these questions here. My main claim is that, in many cases, the agent's maxim does *not* give us the morally relevant description of some act.

On my proposed revisions of Kant's formulas, we no longer use Kant's concept of a maxim. It might be suggested that we could use the word 'maxim' in a narrower sense, which does not cover the policy on which someone is acting, but refers only to what this person is doing. Kant sometimes uses 'maxim' in this way, as when he discusses the maxim 'Kill myself to avoid suffering'. This maxim is not a policy, since we could act on it only once. But this narrower sense of 'maxim' would add nothing to the morally relevant descriptions of people's acts.

We can now add one more objection to Kant's use of the concept of a maxim. When people act, there is often *no* policy on which these people are acting. If we used the word 'maxim' to refer only to policies, we would have to admit that there are many *maximless* acts. To be able to cover such acts, Kant's formulas must often use the word 'maxim' to refer, not to some policy, but to what someone is doing, on the morally relevant description of this person's act. Since Kant's formulas must often be applied directly to people's acts, it is hard to see why these formulas should ever refer to people's policies *rather* than their acts.

It might be objected that, if we revise Kant's formulas by dropping the concept of a maxim, we are no longer discussing Kant's view. This claim is true, but no objection. We are asking whether Kant's formulas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise these formulas in ways that are clearly needed, we are developing a Kantian moral theory. And Kant's use of the concept of a maxim is not, I believe, a valuable part of Kant's own

theory. In ceasing to use this concept, we are not losing anything worth keeping.

Some people might question that last claim. Kant's appeal to the agent's maxim, O'Neill writes, is not 'a detachable or dispensable part of Kant's theory', since this feature of Kant's view enables us to claim that, when some wrong-doer wills that his bad maxim be universal, there is a contradiction in this person's will. We can thereby argue that wrong-doing involves 'failures to have coherent intentions'. But as Kant points out, wrong-doers do not in fact will that their maxims be universal, so 'there is really no contradiction' in these people's wills.

O'Neill also suggests that, by appealing to the agent's maxim, Kant answers the question of what are the morally relevant descriptions of people's acts. But as we have seen and O'Neill elsewhere claims, that is not so. If all we know is that my Egoist has acted on his maxim, we cannot possibly decide whether this man's act was wrong.

It may next be objected that, if we revise Kant's formulas so that they do not refer to maxims, we lose another valuable part of Kant's view. Kant defines a maxim as a subjective *principle* of action, and he asks whether we could will this principle to be a universal *law*. Our revisions of Kant's formulas do not refer to principles or laws. But MB2 could be restated as

> MB3: We act wrongly unless we could rationally will it to be true that everyone accepts some moral principle that permits such acts.

This revision keeps Kant's concern with principles and moral laws.

Return now to O'Neill's suggestion that, by applying Kant's formula to the agent's maxim, we can at least decide whether some act has moral worth. This suggestion has some plausibility, since an act's moral worth may depend on the agent's motive or underlying aim, which may be included in this person's maxim. When applied to my Egoist, O'Neill's suggestion rightly implies that this man's acts never have moral worth. As this man's maxim reveals, he never acts in some way because he believes this act to be his duty, nor does he act for any other moral motive.

When we turn to some other maxims, however, O'Neill's suggestion fails. Suppose that, when acting on his maxim 'Never lie', Kant tells

someone the truth, at what he knows to be some great cost to himself, because he believes correctly that he has a duty to tell this person the truth. If Kant is doing his duty, at such a cost, and his motive is to do his duty, that is more than enough to give his act moral worth. It would be irrelevant that Kant is acting on a maxim that he could not rationally will to be universal. Similar claims apply whenever people do their duty, because they truly believe their act to be their duty. It is irrelevant whether these people are acting on some maxim that they could not rationally will to be universal. Like an act's wrongness, an act's moral worth does not depend on the agent's maxim, in the sense of the policy on which this person acts.

We ought, I conclude, to revise Kant's formulas so that they do not refer to such maxims. After learning from the works of great philosophers, we should try to make some more progress. By standing on the shoulders of giants, we may be able to see further than they could.

13

What if Everyone Did That?

43 Each-We Dilemmas

Though I have claimed that we ought to revise Kant's formulas, I shall go on discussing Kant's own formulas. It is worth showing that we have other reasons to revise these formulas, and many of my claims would also apply to our revised versions.

When we apply Kant's Law of Nature Formula, we ask whether we could rationally will it to be true that everyone acts on some maxim. To answer this question, we must know what the alternative would be. We might be able rationally to will that everyone acts on some bad maxim, such as 'Pay less than my fair share', if the alternative would be that everyone *except us* acts in this way. Another alternative might be that everyone continues to do whatever they are now doing. But Kant's formula would then mistakenly permit us to act on many bad maxims. If many people are already acting on some bad maxim, it would often make too little difference if this maxim were acted on by everyone. On the best version of Kant's formula, which seems to be what Kant has in mind, we should ask whether we could rationally will it to be true that some maxim is acted on by everyone rather than by *no one*.

We also need to know on which *other* maxim everyone would act. We could rationally will it to be true that everyone acts on some bad maxim, if the alternative would be that everyone acted on some other even worse maxim. So we should ask whether there is some other maxim

that is better, in the sense that we have stronger reasons to will it to be true that everyone acts upon it.

Kant's Law of Nature Formula works best when it is applied to maxims or acts of which three things are true:

> it would be possible for many people to act on this maxim, or in this way,
>
> whatever the number of people who act in this way, the effects of each act would be similar,
>
> these effects would be roughly equally distributed between different people.

In discussing such cases, I shall use 'we' to refer to all of the people in some group; and I shall use 'he' and 'himself' in the senses that also apply to women. We are often members of some group of whom it is true that

> if *each* rather than none of us does what would be in a certain way *better, we* would be doing what would be, in this same way, *worse*.

We can call such cases *each-we dilemmas*.

It will be enough to consider cases in which each person's act would benefit one or more people. One large class of each-we dilemmas are the *self-benefiting* dilemmas that are often regrettably called *prisoner's dilemmas*. In such cases, we are members of some group of whom it is true that

> (1) each of us could either benefit himself or give some greater benefit to others,
>
> (2) these greater benefits would be roughly equally distributed between all these people,

and

> (3) what each person does would have no significant effects on what the other people do.

In these cases, if each of us benefits himself, each of us is doing what is certain to be better for himself, whatever the other people do. But if all rather than none of us act in this way, *we* are doing what is certain to be worse for all of us. None of us will get the greater benefits. These cases are *each-we* dilemmas in the sense that

> if *each* rather than none of us does what would be *better* for himself, *we* shall be doing what would be *worse* for each of us.

Put the other way around,

> if *we* do what would be *better* for each, *each* would be doing what would be *worse* for himself.

It would also be bad for us if most of us act in these ways, and worse for us if more of us do. These claims are not about what are misleadingly called *repeated prisoner's dilemmas*, which are much less important, as I explain in a note.

Though each-we dilemmas are often overlooked, they are very common. More exactly, there are few such cases that involve only two people, or only a few people; but there are many cases that involve many people.

Many such cases can be called *contributor's dilemmas*. These involve *public goods*: outcomes that benefit even those people who do not help to produce them. Some examples are clean air, national defence, and law and order. In many of these cases, if everyone contributed to such public goods, that would be better for everyone than if no one did. But it would be better for each person if he himself did not contribute. He would avoid the costs to himself, and he would be no less likely to receive the greater benefits from others. In many of these cases, the public good is that we avoid outcomes that would be bad for everyone, and the contributions that are needed are not financial, but some form of self-restraint.

There are countless actual cases of this kind. In *fisherman's dilemmas*, for example, if each fisherman uses larger nets, he will catch more fish, whatever the other fishermen do. But if all the fishermen use larger nets, the fish stocks will decline, so that, before long, they will all catch fewer fish. It would still be true, however, that it would be better for each fisherman if he uses larger nets, and that if they do they will all catch

even fewer fish. Some other cases involve the many acts that together cause pollution, congestion, deforestation, over-grazing, soil-erosion, droughts, and overpopulation.

These cases are often overlooked because, in many such cases, there are some people to whom these claims do not apply. There may, for example, be some fishermen who are so skilful that, even when there is overfishing, these people still catch as many fish. When that is true, however, the other fishermen would still face an each-we dilemma. In my description of these cases 'everyone' means 'all the members of some group'. Claims (1) to (3) can apply to some group of people even though there are some people in the same community who, though acting in similar ways, are not members of this group.

Many each-we dilemmas do not involve choices between benefiting ourselves or giving greater benefits to others. Such cases can arise whenever people have different and partly conflicting aims. It can be true that, if each rather than none of us does what will best achieve our own aim, everyone's aims will be worse achieved. Some of these may be morally required aims. According to common sense morality, which we can call *M*, we have special obligations to give certain benefits to those people to whom we are related in certain ways. These are people such as our children, parents, pupils, patients, clients, colleagues, customers, or those whom we represent. We can call these our *M-related people*. If we ought to give some kinds of priority to the well-being of these people, we can face each-we dilemmas. In *parent's dilemmas*, for example, each of us can either benefit our own children, or give greater benefits to the children of others. If each rather than none of us gives priority to benefiting our own children, that will be worse for all our children. Many such dilemmas ride on the back of self-benefiting dilemmas. When poor fishermen all catch fewer fish, for example, that may be worse not only for them but also for their malnourished children, who would be even worse fed.

Each-we dilemmas raise both practical and theoretical problems. In some cases, the practical problem has been at least partly solved. Some solutions are *political*, involving changes in our situation. In the case of many public goods, for example, failures to contribute have been made to be either impossible, or worse for each person, by taxation

that is either unavoidable, or enforced by penalties for non-payment. In many other cases, however, political solutions cannot be achieved, or are too costly. In some of these cases, we have achieved solutions that are *psychological*, in the sense that, without any change in our situation, all or most of us choose to give the greater benefits to others. Such solutions often depend on our having and acting upon certain moral beliefs. We may contribute to some public goods, despite the costs to ourselves, because we believe that we ought to contribute.

Of these *moral* solutions to each-we dilemmas, two are especially relevant here. We might be Act Consequentialists, who believe that we ought always to give the greater benefits to others, since we shall thereby do more good. If we all acted on this moral belief, we would all contribute to such public goods. But these solutions are seldom achieved, since there are few people who are both Act Consequentialists and often act on their moral beliefs.

There are also Kantian solutions. If no one contributed to such public goods, that would be much worse for all of us than if everyone contributed. We could not rationally will it to be true that everyone rather than no one acts on the maxim 'Don't contribute'. So, if we were all conscientious Kantians who always acted on Kant's Law of Nature Formula, we would all contribute to these public goods.

When we have achieved some moral solution to some contributor's dilemma, common sense morality requires everyone to go on contributing. In such cases, there are often some *free riders*: people who benefit from these public goods, without making any contribution. Each free rider benefits himself in a way that imposes a greater total burden on others. Common sense morality condemns such acts as unfair. And these are some of the cases in which we can best think and say 'What if everyone did that?'

In *unsolved* each-we dilemmas, things are in one way different. When no one is contributing to some merely possible public good, no one is free-riding, or failing to do their fair share. But Kant's Law of Nature Formula still implies that, in failing to contribute, everyone acts wrongly. These are the cases for which this formula might have been especially designed. If everyone is failing to contribute, we could not say to each other, 'What if everyone did that?' Everyone *is* doing

that. But we can ask our question in another way. Compared with a world in which everyone contributes, so that everyone gets these public goods, we could not rationally will it to be true that no one contributes, so that no one gets these goods. So Kant's formula requires us all to contribute.

When applied to such cases, Kant's formula conflicts with, and may lead us to revise, some widely held and at least partly mistaken moral beliefs. In unsolved each-we dilemmas, most of us believe that we are either permitted or required to give the smaller benefits to ourselves, or to some of our M-related people, rather than giving the greater benefits to others. According to Kant's Law of Nature Formula, such acts are wrong. None of us could rationally will it to be true that all rather than none of us continue to act in these ways, since that would be worse for all of us, or worse for all of our M-related people.

As well as conflicting with some widely held beliefs, Kant's formula challenges these beliefs in an especially forceful way. Though Act Consequentialists would also claim that everyone ought to give the greater benefits to others, the Kantian argument for this conclusion is harder to reject. In unsolved each-we dilemmas, each of us is trying to benefit ourselves, or our children, parents, pupils, patients, or other M-related people. When judged at the *individual* level, each of us succeeds, since each of us *is* doing what is better for himself, or for his children, parents, pupils, patients, etc. But *we* are doing what is *worse* for all these people. *We* are failing, or doing worse, even in our own terms, since we are making it true that everyone's morally required aims will be worse achieved. In these cases, in acting on common sense moral principles, we are acting in ways that are *directly collectively self-defeating*. If we were Rational Egoists, that would be no objection to our view, since this form of Egoism is a theory about *individual* rationality and reasons. But moral principles or theories are intended to answer questions about what *all* of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level.

Kant comes close to giving such an argument. When Kant discusses the limits on our duty to benefit others, he writes,

> a maxim of promoting the happiness of others with a sacrifice of one's own happiness . . . would conflict with itself if it were made into a universal law.

Kant must mean 'with a *greater* sacrifice of one's own happiness'. His point must be that, if everyone promoted the happiness of others at a greater cost to their own happiness, everyone would lose more happiness than they gained. If the effects of such acts would be roughly equally distributed between different people, that would be true. This would be how this maxim would 'conflict with itself'. A similar point applies to a maxim of promoting one's own happiness at a greater cost to the happiness of others. On similar assumptions, if this maxim were a universal law, it would also conflict with itself. There would be only one maxim that could be made universal without conflicting with itself, or being collectively self-defeating. This would be the maxim of doing whatever would, on the whole, best promote everyone's happiness.

Kant's formula has even greater value when it is applied to one kind of unsolved each-we dilemma. In many cases,

> (4) each of us could benefit ourselves or our M-related people in ways that would impose a greater total sum of burdens on others. But these burdens would be spread over very many people. So each act would impose burdens on each of these other people that would be trivial, and would often be imperceptible.

These claims are true in most of the contributor's dilemmas mentioned above. When we know that our acts would impose only such trivial or imperceptible burdens on each of many other people, our ordinary concern for others would not be aroused. Even if we were conscientious Act Consequentialists, we would be likely to ignore such effects. But when many of us act in these ways, the combined effects may be very great and very bad. One example is the way in which, by using fossil fuels, we are recklessly and selfishly overheating the Earth's atmosphere. In such cases, Kant's Law of Nature Formula can act like a moral magnifying glass, getting us to see what we are doing. We could not

rationally will it to be true that we together inflict such damage on ourselves, our children, and our children's children.

44 The Threshold Objection

We can now turn to some cases in which Kant's formulas do less well. According to Kant's

> *Law of Nature Formula*: It is wrong to act on some maxim unless we could rationally will it to be true that everyone acts upon it.

In some cases, however, whether some act is wrong depends on how many people act in this way. When that is true, Kant's formula may fail, by condemning acts that are right, or permitting acts that are wrong.

In discussing such cases, it will be enough to consider acts whose rightness depends in part on their predictable effects. There are many maxims of which it is true that

> (5) if too many people acted on this maxim, these people's acts would have bad effects, but when fewer people act on this maxim the effects are neutral or good.

It may then be true that

> (6) though such acts would be wrong if too many people acted on this maxim, when fewer people act on this maxim such acts are permissible, and may even be morally required.

In such cases,

> (7) most of us could not rationally will it to be true that everyone acts on these maxims.

Kant's formula may mistakenly condemn such acts when they are permissible or even morally required.

One example is the maxim 'Have no children, so as to devote my life to philosophy'. If Kant acted on this maxim, he did not act wrongly. But

he could not have rationally willed it to be true that everyone acts on this maxim, so Kant's formula seems to imply that Kant's deliberate failure to have children would have been wrong. Consider next the maxims: 'Consume food without producing any', 'Become a dentist', and 'Live in Iceland, to absorb the spirit of the Nordic Sagas'. It is not wrong, in the world as it is, to act on these maxims. But since we could not rationally will it to be true that everyone acts on these maxims, Kant's formula seems to imply that such acts are wrong. Other examples are: 'Don't take the first slice', 'Don't speak until others have spoken', and 'When you meet another car on a narrow road, stop and wait until the other car has passed'. We could not rationally will it to be true that everyone acts on these maxims. In such a world, most cakes would never get eaten, most conversations would never get started, and many people's journeys would never end. But acting on these maxims is not, in the actual world, wrong.

Since this problem is raised by acts that are wrong only if the number of such acts is above some rough threshold, we can call this the *Threshold Objection*.

Pogge suggests that, to answer this objection to Kant's view, we should turn from Kant's Law of Nature Formula to his Moral Belief Formula. Though we could not rationally will it to be true that everyone *acts* on such maxims, we *could* rationally will it to be true that everyone believes such acts to be morally permitted. Even if everyone had these beliefs, there is no danger that too many people would choose to act in these ways. Most people already believe that they are permitted to act on the maxims that I have just mentioned. But enough people are having children and producing food. Nor are there too many dentists or inhabitants of Iceland, or too many polite people who always let other people eat, speak, or go first. Since we could rationally will it to be true that everyone believes such acts to be permitted, Kant's Moral Belief Formula permits these acts.

These claims are not, I believe, a sufficient answer to this objection. If none of us had children, we would be ending human history. If none of us produced food, we would be ending history more brutally, by letting ourselves and our children starve to death. These are not merely consequences that we could not rationally will. If we all acted in

these ways, we would be acting wrongly. Nor could we rationally will it to be true that everyone falsely believes that these acts would not be wrong. It is not enough to say that, even if we all had these false beliefs, there is no danger that too many of us would act in these ways. We always have some reason to want ourselves and others not to have false moral beliefs, and these are not cases in which we have any contrary reason.

Pogge suggests another answer to this objection. Many maxims are *conditional*, in the sense that we intend to act in some way only when our acts would have certain effects. Such maxims would not apply when our acts would not have these intended effects, or would have certain other, bad effects. Our maxims may be implicitly conditional in such ways even if we have not had conscious thoughts about these conditions. It is enough that, if these conditions were not met, we would not act on these maxims, and would not have changed our mind.

Of the actual maxims that Kant's Law of Nature Formula may seem mistakenly to condemn, most are at least implicitly conditional. If we intend to produce no food, that intention would not apply if we were starving. Our maxim is something like 'Produce no food as long as enough other people are producing food.' We could rationally will it to be true that everyone acts on this maxim, so Kant's formula does not imply that, in failing to produce food, we are acting wrongly.

We can also assume that, of those who accept the maxim 'Become a dentist', most intend to act on this maxim only if they could thereby earn a living. Perhaps we could rationally will it to be true that everyone accepts this conditional maxim, since we would know that, in the case of most people, this maxim's condition would not be met. But Kant's Law of Nature Formula would here make our moral reasoning take a rather strange form. And we have some reason *not* to will it to be true that everyone accepts this maxim. That would be to will a world whose entire population wanted to become dentists, so that most people had the disappointment of an unfulfilled ambition because there was no room for them in the dental profession. It would be more plausible to follow Pogge's first suggestion, by turning to Kant's Moral Belief

Formula. Anyone is permitted to act on this conditional maxim, we might claim, because everyone could rationally will it to be true that everyone believes such acts to be permitted. That is a better way to explain why, in a world with teeth to be filled, becoming a dentist is not wrong.

We have not yet fully answered the Threshold Objection. Though most people's maxims take such conditional forms, there are some exceptions. Kant may have believed that, since most other people could be relied upon to have children, it was permissible for him to abstain. But of those who choose to have no children, some act on maxims that are unconditional. And moral principles ought to apply successfully to cases that are merely imaginary, when it is clear enough what such cases would involve. We can imagine fanatical, unconditional maxims whose universal acceptance would lead us all to become childless underemployed Icelandic dentists who starved themselves to death. Since we could not rationally will it to be true that everyone acts on these unconditional maxims, or believes such acts to be permitted, Kant's formulas mistakenly condemn our acting on these maxims even when we know that, because few people are acting on these maxims, our acts will have good effects.

This is not, however, a new objection. Like the Egoist's maxim 'Do whatever would be best for me' and Kant's maxim 'Never lie', these are *mixed maxims*, on which it would be sometimes but not always wrong to act. To answer this objection, I have claimed, we should make Kant's formulas apply, not to maxims in the sense that can refer to policies, but to the morally relevant description of what people are doing. On our revised version of Kant's Law of Nature Formula,

> LN3: We act wrongly unless we are doing something that we could rationally will everyone to do, in similar circumstances, if they can.

Suppose that, in acting on these unconditional maxims, we would be having no children, or producing no food, in circumstances in which we knew that there were not too many people who were acting in these

ways. We could rationally will it to be true that everyone acts in these ways, in similar circumstances, if they can. In such a world, there would not be too many people who acted in these ways. So LN3 would not mistakenly imply that these acts would be wrong.

45 The Ideal World Objections

There is another kind of case in which an act's wrongness may depend on the number of people who act in this way. It may be true that

> (8) if enough people acted in some way, these people's acts would have good effects, but when fewer people act in this way the effects would or might be very bad.

It may then be true that

> (9) we ought to act in this way if enough people are doing that, but in other cases such acts are wrong.

Kant's Law of Nature Formula, many writers claim, requires some such acts even when they are clearly wrong.

Consider first the maxim 'Never use violence'. Kant's formula, it is sometimes claimed, requires us to act on this maxim, since there is no other conflicting maxim on which we could rationally will everyone to act. If that were true, Kant's formula would require us never to use violence.

Pacifism has considerable intuitive appeal. And many people (one of them my father) have been pacifists on Kantian grounds. But like Kant's belief that we must never lie, pacifism is too simple. Return to the time of the Second World War. If everyone outside Germany had been pacifists, that would have allowed Hitler to dominate the world, with effects that would have been likely to be even worse than this terrible war. If Kant's Law of Nature Formula implied that it was wrong to fight against Hitler's armies, that would count against this formula.

Suppose next that, in

> *Mistake*, several people's lives are in danger. You and I must choose betwen two ways of acting. The possible outcomes are these:

		I	
		do A	do B
You	do A	we save everyone	we save no one
	do B	we save no one	we save some people

We ought both to do A, since that is our only way to save everyone. But suppose that, because you misunderstand our situation, you do B. Despite knowing that you have made this mistake, I do A, with the result that we save no one. I know that, by doing A, I shall prevent us from saving some people whom we would have saved if I had done B. But as a Kantian, I believe that I ought to do A, since that is the only thing that I could rationally will us both to do.

If Kant's formula implied that I ought to do A, despite knowing that you have done B, that implication would be wholly unacceptable. While pacifism has some plausibility, it would be absurd to claim that I ought here to do A, thereby letting some people die whom we could have saved.

These examples illustrate another objection to Kant's Law of Nature Formula. Kant's 'standard of conduct', Korsgaard writes,

> is designed for an ideal state of affairs: we are always to act as if we were living in the Kingdom of Ends, regardless of possible disastrous results.

Korsgaard takes this problem to be raised by the fact that some people act wrongly. But as *Mistake* shows, this objection to Kant's formula is not raised only by deliberate wrong-doing. Though this case is artificially simple, there are many actual cases of this kind. It is often true that, if we did what we could rationally will everyone to do, as Kant's formula

is claimed to require, our acts would predictably have bad effects of a kind that would make them wrong. Discussing such cases, Hill writes:

> The problem is that acting in this world by rules designed for another can prove disastrous.

According to what we can call this

> *Ideal World Objection*: Kant's formula mistakenly requires us to act in certain ways even when, because some other people are *not* acting in these ways, our acts would make things go very badly, and for no good reason.

In discussing this objection, it will be enough to consider cases in which, as in *Mistake*, it would be best if all of the relevant people acted in the same way. Consider this maxim:

> M1: Do whatever I could rationally will everyone to do.

According to the Ideal World Objection, compared with willing that everyone acts on M1, we could not rationally will that no one does. If this claim were true, Kant's formula would require us to act on M1 even when, as in *Mistake*, our acts would predictably have very bad effects.

This claim is not, however, true. Here is a better maxim:

> M2: Do whatever I could rationally will everyone to do, unless some other people haven't acted in this way, in which case do whatever I could rationally will that, in these circumstances, other people do.

I could rationally will it to be true that everyone acts on M2. In *Mistake*, we would both act on M2 if we both did A, since that is how we could save everyone's lives. But I know that you haven't acted in this way, since you have mistakenly done B. Given your mistake, I could not rationally will that I do A, thereby preventing us from saving anyone. To follow M2, I must do B, thereby enabling us to save at least some people. Since Kant's formula permits me to act on M2 rather than M1, this formula permits me to respond to your mistake in what is obviously the right way.

Return next to the pacifist maxim 'Never use violence'. According to the Ideal World Objection, Kant's formula requires us to act on this maxim, since there is no other conflicting maxim on which we could rationally will everyone to act. As before, that is not so. Here is a better maxim:

> Never use violence, unless some other people have used aggressive violence, in which case use restrained violence when that is my only possible way to defend myself or others.

Everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which no one ever uses violence. So Kant's formula does not require us to be pacifists, but permits us to use restrained violence to resist aggression.

Similar claims apply to all such cases. Kant's formula never requires anyone to act on unconditional maxims like M1 or the pacifist maxim. Everyone could rationally will it to be true that everyone acts on conditional maxims like M2 or the maxim of resisting aggression. In acting on such maxims, as Kant's formula permits, we could respond in the best ways to the wrong acts or mistakes of other people.

There is, however, another problem. Kant's Law of Nature Formula merely *permits* us to act on these better maxims. Consider this maxim:

> Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can.

As before, everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which no one ever uses violence. But in the real world some people have used aggressive violence. Since this maxim passes Kant's test, Kant's formula permits the rest of us to act upon it, by killing as many people as we can. Consider next:

> Keep my promises, and help those who are in need, unless some other people haven't acted in these ways, in which case copy them.

This maxim also passes Kant's test. Everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which everyone kept their promises and helped those who were in need. In the real world, however, some people haven't acted in these ways. Since this maxim passes Kant's test, Kant's formula mistakenly permits the rest of us to copy these other people, by breaking all our promises and never helping those who are in need.

To state this problem in a simpler way, we can turn to

> M3: Do what everyone could rationally will everyone to do, unless some other people haven't acted in these ways, in which case do whatever I like.

Since everyone could rationally will it to be true that everyone acts on M3, this maxim passes Kant's test. We know that, in the real world, some people haven't acted on M3, since these people haven't done what everyone could rationally will them to do. So, in permitting us to act on M3, Kant's formula permits the rest of us to do whatever we like.

According to the Ideal World Objection, Kant's formula sometimes requires us to act as if we were in an ideal world even when, in the real world, such acts would have disastrous effects, and would be clearly wrong. We can answer that objection by applying Kant's formula to conditional maxims, as we often need to do for other reasons. But we have now found that, when applied to such maxims, Kant's formula requires too little. According to this

> *New Ideal World Objection*: Once a few people have failed to do what we could rationally will everyone to do, Kant's formula ceases to imply that any act is wrong.

If this objection cannot be answered, it would be at least as damaging.

Similar claims apply to some other moral principles or theories. According to one version of *Rule Consequentialism*, or

> RC: Everyone ought to follow the rules whose being followed by everyone would make things go best.

We *follow* some rule when we succeed in doing what this rule requires us to do. It is often objected that RC requires us to follow these *ideal rules* even when we know that, because some other people are not following these rules, our acts will have disastrous effects. This objection can be answered. Consider

> R1: Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever, given the acts of others, would make things go best.

This is one of the ideal rules, since everyone's following R1 would make things go best. So RC does *not* require us to follow those ideal rules whose being followed by only some people would have disastrous effects. But consider

> R2: Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever you like.

Since R2 is *also* one of the ideal rules, RC permits us to follow this rule. We know that, in the real world, some people have not followed the ideal rules. So in permitting us to follow R2, RC permits the rest of us to do whatever we like. Similar objections apply to most other versions of Rule Consequentialism, such as those theories which appeal to the rules whose being *accepted* by everyone, or by most people, would make things go best. And similar objections apply to some Contractualist moral theories.

To answer this new objection to Kant's Law of Nature Formula, we should again revise this formula. When we apply this formula to some maxim, it is not enough to ask whether we could rationally will it to be true that *everyone* acts upon it. Kant's formula could become:

> *LN4*: It is wrong for us to act on some maxim unless we could rationally will it to be true that this maxim be acted on by everyone, and by *any other number* of people, rather than by no one.

For some maxim to pass this wider test, we must be able rationally to will that this maxim be acted on, not only by *everyone* rather than by no one, but also by *most* people rather than by no one, by *many* people rather than by no one, by a *few* people rather than by no one, and by any other number of people rather than by no one. We must be able rationally to will that, *whatever* the number of people who *don't* act on this maxim, *everyone else* does.

If we widen Kant's formula in this way, it condemns the bad maxims that we have discussed. One example is:

> Do not use violence, unless some other people have used aggressive violence, in which case kill as many people as I can.

Though we could rationally will it to be true that *everyone* acts on this maxim, we could not rationally will that any other number of people act upon it. If anyone uses aggressive violence, everyone else would act on this maxim by killing as many people as they can.

When we consider many maxims and acts, this revision of Kant's formula would make no difference. There are many acts that are right whatever the number of people who act in this way. In such cases there are unconditional maxims on which we could rationally will any number of people to act. Some examples are the maxims 'Help those who are in need' and 'Never injure others merely for my own convenience'. As we have seen, however, when we consider some other kinds of act, what we could rationally will is that people act on conditional maxims which tell us to take into account the acts of others. Some such maxims could take this form:

> Do A, unless the number or proportion of A-doers is or will be below some threshold, in which case do B, or below some other threshold, in which case do C.

Some of these thresholds could be defined as the numbers or proportions of A-doers below which acts of kind A would cease to have certain good effects, or would start to have certain bad effects.

Similar claims apply to Rule Consequentialism. The formula stated above could become

> RC2: Everyone ought to follow the rules whose being followed by any number of people rather than by no one would make things go best.

Some of these rules could take such conditional forms. These rules would tell us to act in the ways that would make things go best, given the number or proportion of people who are following these rules. Similar claims would apply to those versions of RC which appeal to what would happen if people *accepted* certain rules.

This revision makes Rule Consequentialism in some ways closer to Act Consequentialism. That is most importantly true when we ask what proportion of their income or wealth the world's rich people ought to give to the more than a billion people who now live on around $2 a day. When applied to this question, most versions of Rule Consequentialism are not very demanding. These theories appeal to claims about what would be true if *all* or *most* people accepted or followed certain principles. Things might go best if all or most rich people gave to the poor some fairly modest proportion of their wealth or income, such as one fifth, or even one tenth. That would make a great difference, since the richest nations now give less than one per cent. If we revise Rule Consequentialism by changing 'all' or 'most' to 'any number of people', and we appeal to conditional rules of the kind just mentioned, Rule Consequentialism would often be much more demanding. If most rich people are not giving what it would be best for the rich to give, the best rule would require the others to give a great deal.

In revising Kant's Law of Nature Formula in this way, we give up the idea expressed in the question 'What if everyone did that?' But this idea can be successfully applied only to certain kinds of case. In each-we dilemmas, if we are free-riders who fail to contribute to some public good, we can be rightly challenged with the question 'What if everyone did that?' But in many other cases, it is enough to reply 'Most people won't'.

Kant's Moral Belief Formula appeals to a different idea, which might be successfully applied to all kinds of case. Though we cannot plausibly assume that everyone ought to act on the same maxims, or in the same ways, we *can* plausibly assume that everyone ought to have the same moral beliefs. So when people object to one of our moral beliefs, saying 'What if everyone thought like you?', it is *not* enough simply to reply 'Most people won't'. If we could not rationally will it to be true that everyone believes some kind of act to be permitted, this fact might, as Kant assumes, show such acts to be wrong.

We can now turn to some simpler and more fundamental questions.

14

Impartiality

46 The Golden Rule

When describing how his Formula of Universal Law explains our duty to benefit others, Kant writes

> I want everyone else to be beneficent toward me; hence I ought also to be beneficent toward everyone else.

This may remind us of

> *The Golden Rule*: We ought to treat others as we would want others to treat us.

This rule expresses what may be the most widely accepted fundamental moral idea, which was independently discovered in at least three of the world's earliest civilizations. Though Kant calls his formula 'the supreme principle of morality', he dismisses the Golden Rule as 'trivial' and unfit to be a universal law. Does this rule deserve Kant's contempt?

In rejecting the Golden Rule, Kant writes:

> It cannot be a universal law, because it does not contain the ground of duties toward oneself, nor that of duties of love toward others (for many a man would gladly agree that others should not benefit him if only he might be excused from benefiting them); and finally it does not contain the ground of duties owed to others, for a criminal would argue on this ground against the judge who punishes him.

According to one of Kant's objections, the Golden Rule does not imply that we have duties to benefit others. Many people, Kant claims, would gladly agree never to be benefited by others.

This objection backfires. These people ought to benefit or help others, the Golden Rule implies, if they themselves would want to be helped. Kant does not deny that these people would want to be helped. He makes the different claim that these people would agree not to be helped *if they would thereby be excused* from helping others. To state this claim in Kantian terms, these people would will it to be true that the maxim of not helping others be a universal law. That does not imply that, according to the Golden Rule, these people have no duty to help others. It is *Kant's* formula, not the Golden Rule, that permits us to act on maxims that we could will to be universal laws.

Kant's objection might be revised. He might ask us to consider people who do *not* want to be helped by others, whether or not they would thereby be excused from helping others. Kant might then claim that, since these people do not want to be helped, the Golden Rule fails to imply that they have a duty to help others.

As before, however, this objection would apply to Kant's own formula. According to this formula, these people ought to help others if they could not will it to be true that the maxim of not helping others be a universal law. If these people do not even want to be helped, they could more easily will that this maxim be such a law. No one could will such a law, Kant claims, because such a person would thereby 'rob himself of all hope of the assistance that he wishes for himself'. This claim does not apply to people who *don't* wish to be helped.

Kant might reply that, in not wishing or wanting to be helped, these people would be irrational. And he might then argue that, when applied to such people, his formula does better than the Golden Rule. Kant might claim that, since the Golden Rule appeals to these people's desires, which are irrational, this rule fails to imply that these people have a duty to help others. In contrast, because these people could not *rationally* will it to be true that they would never be helped, Kant's formula does imply that they have this duty.

This objection to the Golden Rule has no force. We can first explain why, in most of its stated versions, this rule does not appeal to how

we would *will* that others treat us. We are not absolute monarchs or dictators, who can successfully will it to be true that other people act in some way. Since we do not have such power over others, we can only want or wish it to be true that other people act in some way. Kant's formula asks us to imagine or suppose that we have the power to choose, or will it to be true, that other people act in some way. The Golden Rule could take the same form. This rule need not appeal to our desires, but could appeal to how, if we had the choice, we would will that we ourselves be treated—or how we would be *willing* to be treated. Some familiar statements of the Golden Rule, such as 'Do as you *would be* done by', already take this form.

The Golden Rule can also appeal to what we would *rationally* choose, or will. It is true that, as commonly stated, this rule does not use the concept *rational*. But of Kant's many statements of his formula, only two use this concept, and none explicitly appeal to what we could rationally will. Given some of Kant's other claims, Kant clearly intends us to ask what we could rationally will or choose. The Golden Rule could take the same form. This rule could be stated as

> G2: We ought to treat others only in ways in which we would rationally be willing to be treated by others.

When we apply the Golden Rule, it is sometimes enough to ask whether we would be willing, in the actual world, to be treated in some way. Torturers, for example, would not be willing to be tortured. But when considering many kinds of act, we must ask how we would be willing to be treated in some merely imaginary case. When we could feed someone who is starving, for example, it is not enough to ask whether we would be willing to be given no food. If we have just eaten well, and have a well-stocked kitchen, our answer to that question might be Yes. We should ask whether, even if we were starving, we would be willing to be given no food.

Consider next some white racist who, in the worst period of racial discrimination in the Southern USA, excludes black people from his hotel. This man might claim to be obeying the Golden Rule. He might say:

> We ought to treat others only as we would be willing to be treated by others. I admit to my hotel anyone who is not black.

> I would be willing to be treated in this way. I *am* treated in this way. Since I am not black, I am admitted to every hotel.

This speech misunderstands the Golden Rule. On this rule, this man ought to treat black people only as he would be willing to be treated *if he were going to be in their position*. He must imagine either that (1) all hotels are owned by black people who exclude white people, or that (2) he himself is black. Though (1) would be merely a change in his circumstances, (2) would be a change in him. When we apply the Golden Rule to many other cases, the imagined change would have to be in ourselves, since we must imagine being relevantly *like* the people whom our acts would affect, by having these people's desires, attitudes, and other physical or psychological features. For example, for some man to imagine being treated as he treats women, he may have to imagine that he is a woman. Similar claims apply to sadomasochists.

In a fuller statement, then, the Golden Rule could be

> G3: We ought to treat others only in ways in which we would rationally be willing to be treated, if we were going to be in these other people's positions, and would be relevantly like them.

The phrase 'would be willing' can be misleading. In applying G3, we should not ask how, if we were in these other people's positions, we would *then* be willing to be treated. We should ask how we would *now* be willing to be treated later, if we were later going to be in these people's positions. (If I similarly said 'Would you want your organs to be used after you are dead?', I would be asking you, not to predict your *post mortem* desires, but to make a decision now.)

Kant gives another objection to the Golden Rule. By appealing to this rule, Kant claims, 'a criminal could argue against the judge punishing him'. Kant must be assuming here that this criminal could say: 'Since you would not want to be punished, you ought not to punish me.' This objection takes the Golden Rule to be

> G4: We ought to treat *each* other person as we would rationally be willing to be treated, if we were going to be in this person's position, and we would be relevantly like this person.

Kant would be right to reject *this* rule. Suppose that, in

> *Case One,* I could save either Blue's life, or Brown's.

By appealing to G4, Blue could argue that I ought to save her life. I would not be willing to be left to die if I were going to be in Blue's position. Brown could similarly argue that I ought to save her life. So G4 mistakenly implies that, whatever I do, I shall be acting wrongly, by failing to treat either Blue or Brown as I ought to do. Suppose next that, in

> *Case Two*, I have a small loaf of bread, and meet two starving people.

By appealing to G4, each person could argue that I ought to give her my whole loaf.

When Jesus appealed to the Golden Rule, was he appealing to G4? Was he intending to imply that it would be wrong for me to share my loaf between these people? The answer is clearly No. The Golden Rule should be taken to mean, not G4, but

> G5: We ought to treat *other people* as we would rationally be willing to be treated if were going to be in the positions of *all* of these people, and would be relevantly like them.

In this better form, however, this rule is harder to apply. How are we to imagine that we shall be in the positions of two or more people?

Several suggestions have been made. Suppose that, in

> *Case Three*, I could either save Green's life, or save Grey from going blind.

On Nagel's proposal, I should imagine that, like an amoeba, I shall later divide and become two people, one in Green's position and the other in

Grey's. On Hare's proposal, I should imagine that I shall later live lives that would be just like those of Green and Grey, not simultaneously, but one after the other. On Harsanyi's proposal, I should imagine that I shall have an equal chance of being in either Green's position or in Grey's. On Rawls's proposal, I should imagine that I shall be in one of these people's positions, but with no knowledge of the probabilities.

When we apply the Golden Rule to certain questions, it might make a difference which of these proposals we adopt. But in most cases these proposals would have the same implications. In *Case Three* for example, in whichever of these ways I imagine that I shall be in the positions of both Green and Grey, I would not be willing to be saved from blindness in one of these positions rather than being saved from death in the other.

Of those who have appealed to the Golden Rule, many may not have considered the difference between G4 and G5. But if these people had compared these claims, and seen what they imply, they would have regarded G5 as better stating the moral idea that they had in mind.

Return now to Kant's claim that, by appealing to the Golden Rule, a criminal could argue that his judge ought not to punish him. On the better reading of the Golden Rule, as expressed in G5, judges could reject this argument. These judges should ask how they would rationally be willing to be treated if they were going to be, not only in some criminal's position, but also in the positions of all of the other people whom their decision might affect. These other people include the possible victims of the crimes that would be more likely to be committed if this criminal is not punished, either because this criminal would be free and able to commit some other crime, or because he and other potential criminals would be less likely to be deterred. Since this is how judges ought to apply the Golden Rule, this rule does not mistakenly imply that no one should be punished.

According to Kant's remaining objection in the passage quoted above, the Golden Rule cannot be a universal law because this rule does not cover our duties to ourselves. We might reply that, since this rule applies only to our treatment of other people, it does not claim to cover our duties to ourselves. As Kant elsewhere suggests, however, this feature of

the Golden Rule may make it misdescribe some of our duties to others. Suppose that, in

> *Case Four*, I could either save my own life or save Grey from going blind.

If the Golden Rule tells me only how I ought to treat *other people,* this rule might mistakenly imply that I ought to save Grey from blindness at the cost of my life. This might be what I would be willing to have done if I were going to be only in Grey's position.

To meet this objection, this rule could become

> G6: We ought to treat *everyone* as we would rationally be willing to be treated if we were going to be in all of these people's positions, and would be relevantly like them.

The word 'everyone' here refers to all of the people whom our acts might affect. In many cases, *we* are one of these people. On this version of the Golden Rule, when applied to *Case Four*, I ought to do what I would be willing to have done if I were going to be, not only in Grey's position, but also in mine. As in *Case Three,* I would not be willing to be saved from blindness in one of these positions rather than being saved from death in the other. This revision better states the Golden Rule's assumption that everyone matters equally. It is not surprising that, in most statements of this rule, we are told only to treat *others* as we would be willing that we ourselves be treated. There is little danger that we shall ignore our own well-being. But this reference to others is, in a way, misleading, since *we* are among the people whose well-being we ought to consider in the impartial way that this rule requires.

Kant's contempt for the Golden Rule is not, I have argued, justified. But Kant's Formula of Universal Law might still be, as Kant believed, a better principle. Is that so?

These principles often have the same implications. And as candidates for the supreme principle of morality, both meet the most obvious requirements. Both principles succeed in most of the cases in which Kant's Impossibility Formula so spectacularly fails. Most of us could

not rationally will it to be true that everyone acts on maxims of self-interested killing, injuring, coercing, lying, and stealing. Nor would we be willing to be treated in these ways if we were going to be in the positions of the affected people.

Kant's Formula of Universal Law is in two ways similar to the Golden Rule. In their best forms, both principles appeal to claims about what it would be rational for people to choose. And both principles assume that everyone matters equally, and has equal moral claims. The 'intuitive idea' behind Kant's formula, O'Neill writes, is that 'we should not single ourselves out for special consideration or treatment'.

These principles mainly differ in the ways in which they make our moral thinking more impartial. Both principles tell us to carry out certain thought-experiments, by asking questions about some imagined cases. To apply the Golden Rule, we ask 'What if that was done to me?' To apply the law of nature and moral belief versions of Kant's formula, we ask 'What if everyone did that?' and 'What if everyone believed such acts to be permissible?'

When we apply the Golden Rule, our thought-experiment is fairly simple. As when making many ordinary decisions, we ask what would happen in the actual world if we acted, on one occasion, in each of certain possible ways. We don't even need to decide what are the morally relevant descriptions of these particular possible acts. But we try to think about these possibilities, not only from our own point of view, but also from the points of view of all of the other people whom our act might affect. We ask what we would rationally be willing to do, and have done to us, if we were going to be in all of these people's positions, and would be relevantly like them.

Kant's thought-experiments are in several ways harder. When we apply Kant's Law of Nature Formula, we must first decide what is the maxim on which we would be acting. In my revised version of this formula, we must decide what is the morally relevant description of our act. We then compare two possible worlds, or two ways in which the future history of our world might go. We ask what would happen both if everyone acted on some maxim, and if no one did, because everyone acted on some other maxim. Similarly, when we apply Kant's Moral Belief Formula, we ask what would happen both if everyone had some

moral belief, and if no one did, because everyone had some other moral belief. These four possible worlds may all be very different from the actual world, and it would often be hard to predict what these worlds would be like. We may also have to consider various other possible maxims on which everyone might act, or possible moral beliefs that everyone might have. In another way, however, Kant's formulas are easier to apply than the Golden Rule. When we ask in which of these worlds we could rationally choose to live, we think about these worlds only from our own point of view.

Kant's formulas and the Golden Rule can be usefully compared with two other principles. According to another old idea, we should make our moral reasoning impartial in a different and simpler way. We should ask what it would be rational for us to choose, or prefer, neither from our own point of view, nor from the points of view of those other people whom our acts might affect, but from the imagined point of view of some detached observer, who is not involved in the case we are considering. On a variant of this idea, we ask what it would be rational for us to choose, or prefer, when we imagine some other relevantly similar case, in which everyone involved would be strangers to us. We can call this the *Impartial Observer Formula*.

We can also achieve impartiality by applying Kant's Consent Principle. By asking whether everyone could rationally consent to some possible act, we give equal weight to everyone's reasons for refusing consent.

There are various objections to the Golden Rule. It can be difficult to imagine that we shall be in other people's positions and shall be relevantly like these other people. And what we must try to imagine would often be deeply impossible. But that is not, as some writers claim, a decisive objection. Some thought-experiments are useful even though they ask us to imagine something that is deeply impossible. Einstein usefully asked what he would see if he were travelling at the speed of light. Though we could not possibly *be* the horse whom we are whipping, or the trapped and starved animal whose fur we are wearing, we can imagine such things well enough for moral purposes.

Another objection to the Golden Rule has more force. As Rawls points out, if we imagine that we shall be in the positions of all of the people whom our acts might affect, we shall be led to ignore the

fact that, in the real world, our acts would affect different people. One person's burdens cannot be compensated by benefits to other people. In ignoring this 'separateness of persons', we are ignoring facts that may give us decisive reasons to accept principles of distributive justice.

In these and some other ways, the Golden Rule is theoretically inferior to both the Impartial Observer Formula and Kant's Consent Principle. But this rule may be, for practical purposes, the best of these three principles. By requiring us to imagine ourselves in other people's positions, the Golden Rule may provide what is psychologically the most effective way of making us more impartial, and morally motivating us. That may be why this rule has been the world's mostly widely accepted fundamental moral idea.

Of these four ways of making us more impartial, Kant's Formula of Universal Law is, I shall argue, the least successful. This formula fails to condemn many wrong acts. As we shall see, however, these problems have a Kantian solution.

47 The Rarity and High Stakes Objections

When people act wrongly, they may be doing something that cannot often be done. Some of these people could rationally will it to be true that everyone acts like them, since such acts would be too rare to have significant effects on them. I have called this the *Rarity Objection*. Consider, for example,

> *Unjust Punishment*: Unless *White* goes to the police and confesses, *Black* will be convicted and punished for some crime that White committed. Though White knows this fact, he does nothing.

Suppose that White acts on the maxim 'Let others be punished for my crimes'. To apply Kant's Law of Nature Formula, we ask whether White could rationally will it to be true that everyone acts on this maxim. In answering this question, for the reasons that I gave above, we cannot appeal to our belief that White's act would be wrong. Nor can we appeal to the *deontic* reason that the wrongness of this act

might provide. If we appeal only to other, non-deontic reasons, we may have to admit that White could rationally will it to be true that everyone acts on his maxim. We can suppose that, if White lets Black be punished for White's crime, White would avoid many years in prison. If everyone else acted on White's maxim when it applied to them, that would increase the risk that White would later be punished for someone else's crime. But this extra risk would be small, and would be clearly outweighed by the certain benefit to White of avoiding these many years in prison. Kant's formula therefore permits White to let Black be punished for White's crime, though this act is clearly wrong. Nor does Kant's Moral Belief Formula condemn this act, since White could rationally will it to be true that everyone believes such acts to be morally permitted.

For another example, consider

> *Murderous Theft*: While travelling across some desert, *Grey* and *Blue* have both been bitten by some snake. Blue has prudently brought some drug that is an antidote to this snake's lethal poison. Grey cannot save his life except by stealing Blue's drug, with the foreseen result that Blue dies.

Grey knows, we can assume, that no one else would discover that he stole Blue's drug, nor would his life be ruined by remorse. Since Grey is young, he can expect that his act would give him many more years of life worth living. Blue can also expect such a life, and is much younger. On these assumptions, all plausible moral views imply that it would be wrong for Grey to save his life by stealing Blue's drug.

Suppose first that, if Grey stole this drug, he would be acting on the maxim 'Steal when that is my only way to save my life'. Grey could rationally will it to be true that everyone acts on this maxim, whenever it applies to them. It is unlikely that, in such a world, anyone else would treat Grey in this way; and this risk would be clearly outweighed by the certain benefit to Grey if he saves his life. On these assumptions, this case also illustrates the Rarity Objection, since Kant's formulas would permit Grey's murderous theft.

Suppose instead that, in stealing Blue's drug, Grey would be acting on the Egoistic maxim

> E: Do whatever would be best for me.

Could Grey rationally will it to be true that everyone rather than no one acts on this maxim? That depends on the alternative. As I have said, we could not rationally will it to be true that everyone acts on some maxim if there is some other, significantly better maxim on which everyone could act. One such maxim might be

> E2: Do whatever would be best for me, except when such acts would impose much greater burdens on others.

If everyone always acted on E rather than E2, that would be much worse for most people. That is why, as I have claimed, the Egoistic maxim usually fails Kant's test. Most egoists could not rationally choose to live in a world of egoists.

Grey, however, is one of the exceptions. Grey knows that, if everyone acted on E rather than E2, he would often bear burdens that would be imposed on him by the egoistic acts of others. But we can plausibly suppose that, even in such a world, the rest of Grey's life would be worth living. If that is so, Grey could rationally will it to be true that everyone acts on E rather than E2. If everyone acted on E2, Grey would not steal Blue's drug, and would die. If we ignore deontic reasons, we must agree that Grey has sufficient reasons to prefer, not the partly moral world in which he would die, but the egoistic world in which, by stealing Blue's drug, Grey would save his own life. So Kant's Law of Nature Formula mistakenly permits Grey's murderous theft. For similar reasons, so does Kant's Moral Belief Formula.

These claims illustrate a different objection to Kant's formulas. These formulas fail here, not because few other people could act on Grey's egoistic maxim, but because Grey's wrong act gives him a benefit that is unusually great. We can call this the *High Stakes Objection*.

There are some ways in which we might try to answer this objection. For example, we might repeat Rawls's claim that, in asking whether we could rationally choose to live in a world in which everyone acts on some maxim, we should suppose that this maxim has already been acted on for a long enough time for such acts to have had their full effects. We might then argue that Grey could not rationally choose the world in which everyone always acted on the Egoistic maxim, since there is a risk that, in this world, Grey would already be dead, having been earlier killed by some other egoist. This somewhat puzzling argument would not, however, be enough to defend Kant's Law of Nature Formula. We are comparing this formula with three other principles: Kant's Consent Principle, the Impartial Observer Formula, and the Golden Rule. And when applied to the kinds of case that we are now considering, these three other principles clearly do much better.

The chief difference is this. Since Blue is much younger than Grey, Blue's death would be, for her, a much greater loss. In applying these other principles, we take into account Blue's much greater loss. Blue would not have sufficient reasons to consent to Grey's stealing Blue's drug and thereby causing Blue's death. Any rational impartial observer, given the choice, would choose that Grey does not treat Blue in this way. And Grey could not rationally choose that he be treated in this way, if he were going to be, not only in his own position, but also in Blue's. Because these three principles make our moral reasoning impartial, they all rightly condemn Grey's murderous theft.

When we apply Kant's Law of Nature Formula, in contrast, we ignore Blue's well-being, since we think about this case only from Grey's point of view. We ask whether Grey could rationally will it to be true that he saves his life, and lives in a world of egoists. For Kant's formula to condemn Grey's act, the answer must be No. We must claim that Grey could not rationally choose the world in which he saves his life, because he has decisive non-deontic reasons to prefer the world in which he dies. Compared with the claims to which we can appeal when we apply our other three principles, this claim is much harder to defend.

48 The Non-Reversibility Objection

There is another, similar, but practically more important objection to Kant's formulas. The Golden Rule makes us more impartial by requiring us to treat everyone as we would rationally be willing to be treated if we were going to be in the positions of all these people, and would be relevantly like them. Kant's Law of Nature Formula makes us more impartial in a less direct way. When we apply this formula, rather than asking 'What if that was done to me?' we ask 'What if everyone did that?'

This question has some value. When we act wrongly, as Kant points out, we often make unfair exceptions for ourselves, doing things that we would not want or will other people to do. Kant's Law of Nature Formula rightly condemns such acts. And as I have claimed, this formula is especially helpful when we are considering each-we dilemmas.

Kant's question is not, however, enough. In many cases, if we act wrongly, we would benefit ourselves in ways that would impose much greater burdens on others. The Golden Rule condemns such acts, since we would not rationally be willing to have other people do such things to us. But when we apply Kant's formula to our acting on some maxim, we don't ask whether we could rationally will it to be true that *other* people do these things to *us*. We ask whether we could rationally will it to be true that *everyone* does these things to *others*. And we may know that, even if everyone did these things to others, *no one* would do these things to *us*. When that is true, we *could* rationally will it to be true that everyone acts like us, since we would then get the benefits from our own wrong acts, and the similar wrong acts of others would never impose the greater burdens on us. Kant's formula mistakenly permits such acts. In the simplest cases of this kind, our wrong acts are *not reversible*, since we are doing to others what they could not possibly do to us. So we can call this the *Non-Reversibility Objection*.

Unlike the Rarity and High Stakes Objections, this objection applies to many actual cases. Return first to our white racist. This man cannot claim to be following the Golden Rule. But he might claim to be following Kant's formulas. He might say:

> When I exclude blacks from my hotel, I could rationally will it to be true that everyone acts in this way. Everyone *does* act in this way. Every hotel owner excludes blacks. And I could rationally will it to be true that everyone believes such acts to be right. If the blacks believed that my acts are right, that would be fine with me.

If this man made these claims, would he have misunderstood Kant's formulas? I am not asking whether he would have misunderstood Kant's moral theory. Kant was in some ways remarkably egalitarian, and there is much in Kant's views that would condemn such racist attitudes and acts. My question is only what is implied by Kant's Law of Nature and Moral Belief Formulas.

When Kant illustrates his formulas, he considers maxims on which most people do not act, and on which, he assumes, no one would want everyone to act. When he imagines some wrong-doer asking 'Could I will that my maxim be a universal law?', Kant assumes that this person's maxim *isn't* such a law. But in some cases, like that of this white racist, this assumption fails. This man's maxim is already a universal law. When this man acts on the maxim 'Exclude blacks from my hotel', he is doing what, in his social world, all hotel owners do.

When wrong-doers act on such maxims, it may not help to ask 'What if everyone did that?' Kant's Law of Nature Formula permits such people's acts if they could rationally will it to be true that they and others continue to act as they are now doing. If it is bad for these wrong-doers that they and others are acting in some way—as might be true, for example, in some state of anarchy, or a war of all against all—these people could not rationally will the continuation of the existing state of affairs, or *status quo*. Kant's formula would then rightly condemn these people's acts. In many cases, however, the *status quo* is good for the people who are acting wrongly. And this state of affairs may be good for these people partly *because* their bad maxim is universal, or widely acted upon. Those to whom some maxim applies may be some powerful and privileged group, who are acting in ways that preserve their advantages over other people. Kant's Law of Nature

Formula permits such people's acts if they could rationally will it to be true that they keep their privileged positions.

As before, in trying to argue that these people could *not* rationally choose to keep their privileged positions, we should not appeal to the wrongness of these people's acts, since Kant's formula would then achieve nothing. Nor could we usefully claim that these people are rationally required to give great weight to everyone else's well-being. Kant, rightly, does not appeal to such claims. For Kant's formula to support the view that these people's acts are wrong, we must be able to claim that, for other reasons, these people could not rationally will it to be true that they keep their advantages over other people. At least in the case of many of these people, we could not plausibly defend this claim.

Nor would it help to turn to Kant's Moral Belief Formula. Just as these people could rationally will it to be true that everyone in their position acts like them, they could rationally will it to be true that everyone believes such acts to be morally permitted. These people would have no relevant reason to prefer that everyone believes their acts to be wrong.

Consider, for example, those men who benefit themselves by treating women as inferior, denying women various rights and privileges, and giving less weight to women's well-being. Such acts are wrong, Kant's formulas imply, if these men could not rationally will it to be true either that everyone acts like them, or that everyone believes such acts to be justified. These claims do not provide a good objection to these men's acts. For most of history, most people — including most women — have treated women as inferior, and believed such treatment to be justified. Since we cannot appeal to the wrongness of such treatment, we would have to admit that many men could have rationally willed that they keep their privileged position.

Turn next to slave-owners. For Kant's formulas to condemn slavery, we would have to argue that slave-owners could not have rationally willed it to be true that they keep their slaves, and that everyone, including the slaves, believes slavery to be justified. Since we cannot appeal to the wrongness of slavery, these claims might be hard to defend. It would be much better to appeal to Kant's Consent Principle, or to the

Golden Rule. Women and slaves could not rationally consent to being treated as inferior, or as mere property. Nor would men or slave-owners be willing to be treated in these ways, if they were going to be in the positions of women or slaves.

Similar claims apply to many of the ways in which powerful people benefit themselves by oppressing or exploiting those who are weak. Kant's formulas condemn these people's acts only if they could not rationally will it to be true either that they and others continue to profit in these ways, or that everyone believes such exploitation to be justified. Since we cannot appeal to the unjustifiability of such exploitation, we could not plausibly defend these claims.

For one last example, we can return to global inequality. On any plausible moral view, those who control much the greatest shares of the world's resources ought to transfer much of their wealth or income to the poorest people in the world. Most rich people transfer nothing. To argue that Kant's formulas condemn these people's acts, we would have to claim that these rich people could not rationally will it to be true either that they and others continue to give nothing to the poor, or that everyone believes that, in giving nothing, the rich are acting rightly. Since we cannot relevantly appeal to the wrongness of these people's acts, or to altruistic rational requirements, we could not plausibly defend these claims. These rich people could rationally will it to be true that they continue to act as they do, and that everyone believes their acts to be morally justified.

When Korsgaard discusses Kant's Formula of Universal Law, she writes:

> the kind of case around which the view is framed, and which it handles best, is the temptation to make oneself an exception, selfishness, meanness, advantage-taking, and disregard for the rights of others. It is this sort of thing, not violent crimes born of despair or illness, that serves as Kant's model of immoral conduct. I do not think we can fault him on this, for this and not the other is the sort of evil that most people are tempted by in their ordinary lives.

Kant's formula does not, I have argued, best handle selfishness, meanness, and advantage-taking. In both its law of nature and moral belief versions, Kant's formula fails to condemn many of the acts with which some people take advantage of others—as when men, the rich, and the powerful take advantage of women, the poor, and the weak. And since Kant presents his formula as the supreme principle of morality, we *can* fault this formula for its failure to condemn such acts. These kinds of selfishness and advantage-taking are precisely the sorts of evil that men, the rich, and the powerful are tempted by, and often commit, in their ordinary lives.

49 A Kantian Solution

It might be claimed that, in presenting these objections to Kant's Formula of Universal Law, I have misinterpreted this formula. Nagel suggests that, when we ask whether we could rationally will it to be true that everyone acts on our maxim, Kant intends us to imagine that we are going to be in everyone else's positions, and that we shall be relevantly like all these other people. This suggestion makes Kant's formula like a greatly inflated version of the Golden Rule, which requires us to try to imagine that we shall be in the positions of billions of other people.

None of Kant's claims about his formula support Nagel's interpretation. And there are contrary passages, such as Kant's discussion of the rich and self-reliant man who has the maxim of not helping others who are in need. When Kant claims that this man could not rationally will that his maxim be a universal law, he writes:

> many cases could occur in which . . . by such a law of nature arisen from his own will, he would rob *himself* of all hope of the assistance that he wishes for *himself*.

If Kant intended this man to imagine that he was going to be in the positions of the other people who need help, he would surely say that here.

Nagel defends his interpretation with the claim that, if Kant did not intend us to imagine that we were going to be in everyone else's

positions, Kant's formula would be open to serious objections. But even the greatest philosophers can overlook objections.

Rawls proposes another interpretation of Kant's formula. When we apply this formula, Rawls suggests, Kant intends us to imagine that we know nothing about ourselves or our circumstances. We should ask what we could rationally will if we were behind a *veil of ignorance*, not knowing whether we are men or women, rich or poor, fortunate or in need of help. Like Nagel, Rawls supports this interpretation with the claim that it seems needed to defend Kant's formula from objections. But even if Kant ought to have used the idea of a veil of ignorance, that doesn't show that he did. In his discussions of his Formula of Universal Law, Kant never suggests that we ought to imagine that we know nothing about ourselves or our circumstances.

On a third interpretation of Kant's formula, suggested by T. C. Williams, Kant intends us to judge our maxims from the imagined point of view of an impartial observer. Williams similarly defends his interpretation with the claim that it is needed to defend Kant's formula from objections. But when Kant discusses his formula, he never asks us to imagine that we are impartial observers.

Scanlon proposes a fourth interpretation. When we apply Kant's formula, Scanlon suggests, Kant intends us to ask whether *everyone* could rationally will that our maxim be a universal law. But this cannot be what Kant means. Kant writes:

> I ought never to act except in such a way that *I* could also will that my maxim be a universal law.

Kant gives many different statements of his formula, none of which refers to what everyone could will.

These proposals would be better made, not as claims about what Kant means, but as ways of revising Kant's formula so that it can avoid objections of the kind that we have been considering.

Of these proposed revisions, Scanlon's, I believe, is the best. According to the moral belief version of Kant's formula, or

> MB: It is wrong for us to act on some maxim unless *we ourselves* could rationally will it to be true that everyone believes that such acts are morally permitted.

On Scanlon's proposal, this would become

> MB4: It is wrong for us to act on some maxim unless *everyone* could rationally will it to be true that everyone believes that such acts are morally permitted.

This revision is also suggested by some of Kant's claims about two of his other principles, the Formulas of Autonomy and of the Realm of Ends. For example, Kant refers to

> the idea of the will of every rational being as a will giving universal law.

Though Kant never appeals to what everyone could rationally will, that may be only because he assumes that this revision of his formula would make no difference. Kant may assume that what any one person could rationally will must be the same as what everyone else could rationally will. On this assumption, MB and MB4 would always coincide.

This assumption, I have claimed, is false. What could be rationally willed by many of those who are men, rich, or powerful could *not* be rationally willed by many of those who are women, poor, or weak. Since there can be such differences between what different people could rationally will, MB and MB4 sometimes conflict, and we must choose between them. If Kant had seen the need to make this choice, he would have rightly chosen MB4.

Remember next that we ought to revise Kant's formula so that it applies, not to the agent's maxim, but to the morally relevant description of what this person is doing. Our revised formula can therefore become

> MB5: It is wrong to act in some way unless everyone could rationally will it to be true that everyone believes that such acts are morally permitted.

With similar revisions, Kant's Law of Nature Formula would become:

> LN5: It is wrong to act in some way unless everyone could rationally will it to be true that everyone acts in this way, in similar circumstances, whenever they can.

As I explain in a note, it is enough to appeal to MB5.

When people believe that some kind of act is morally permitted, they accept some principle that permits such acts. So MB5 can become

> *the Formula of Universally Willable Principles*: An act is wrong unless such acts are permitted by some principle whose universal acceptance everyone could rationally will.

In Scanlon's words, 'to answer the question of right and wrong what we must ask is . . . "What general principles of action could we all will?" '

This formula makes our moral reasoning impartial in a way that avoids the Rarity, High Stakes, and Non-Reversibility Objections. Since this formula does not appeal to the agent's maxim, it avoids the Mixed Maxims Objection. Since this formula allows us to appeal to conditional principles, it also avoids the Threshold Objection. We need another revision to avoid the New Ideal World Objection, but that revision would raise some complications that we can here ignore.

After considering some similar objections, as I have said, some people have come to believe that Kant's Formula of Universal Law cannot help us to decide which acts are wrong. When applied to such questions, Wood calls this formula 'radically defective' and 'pretty worthless', Herman claims that it cannot be made to work, Hill doubts that it can provide 'even a loose and partial action guide', and O'Neill claims that it often gives either unacceptable guidance or no guidance at all. Since these claims are about Kant's actual formula, they are, as I have argued, justified. Whether some act is wrong does not depend on the agent's maxim, and Kant's formula cannot succeed if this formula appeals only to what the agent could rationally will. But we can revise Kant's formula by dropping Kant's appeal to the concept of a maxim in the sense that covers policies, and appealing instead to principles,

and to what everyone could rationally will. All these objections then disappear.

If we appeal to the principles that everyone could rationally choose to be the principles that everyone accepts, our view is of the kind that is called *Contractualist*. Several writers, such as Rawls and Scanlon, propose what have been called *Kantian* versions of Contractualism. But the Formula of Universally Willable Principles is, I believe, the version of Contractualism that is closest to Kant's own view. So we can restate this formula, and give it a shorter name. According to

> *the Kantian Contractualist Formula*: Everyone ought to follow the principles whose universal acceptance everyone could rationally will.

This formula might be what Kant was trying to find: the supreme principle of morality.

15

Contractualism

50 The Rational Agreement Formula

Most Contractualists ask us to imagine that we and others are trying to reach agreement on which moral principles everyone will accept. According to what we can call

> *the Rational Agreement Formula*: Everyone ought to follow the principles to whose being universally accepted it would be rational for everyone to agree.

Some Contractualists appeal instead to the principles to whose being universally *followed*—or *successfully* acted upon—it would be rational for everyone to agree. Most of my claims would apply to such versions of Contractualism, to which I shall return. I shall say that we *choose* the principles to whose universal acceptance we agree. We choose rationally, most Contractualists assume, if our choice would be best or expectably-best for ourselves. We can start with that assumption.

Though there are some principles whose universal acceptance would be best for everyone, there are others whose acceptance would be best only for certain people. What would be best for men, for example, would not always be best for women. It may seem that, when people's interests conflict, there would be no principle whose choice would be rational for everyone in self-interested terms. But the Rational Agreement Formula applies only to principles that it would be rational for *everyone* to choose. There would be no point in our choosing principles whose acceptance would be best for ourselves, if some other people could not rationally choose these principles.

What we could rationally choose would also depend on the effects of our failing to reach agreement. Most Contractualists tell us to suppose that, if we failed to agree, no one would accept any moral principles, so that no one would believe that any acts were wrong. Such a world would be likely to be bad for everyone. In this amoral *No-Agreement World*, as Hobbes memorably wrote, our lives would be 'solitary, poor, nasty, brutish, and short'. That would give everyone strong self-interested reasons to try to reach agreement.

We can suppose that, to make agreement easier to achieve, there would be discussions, and a series of straw votes. But there would have to be some final vote. We must all know that, if we failed to reach agreement in this last round, we would have lost our last chance, since we could not try again. In earlier rounds, it would be rational for us to vote tactically. We could declare that we intended to choose principles that favoured ourselves, and we would vote for these principles, thereby trying to persuade others to vote for these principles as well. Only in the decisive final vote would it be rational for each of us, given our need to reach agreement, to make our full concessions to others.

Morality, some Contractualists believe, is best regarded as a mutually advantageous bargain. This need not be an actual bargain. When people's interests conflict, it would be rational for everyone to agree on certain principles to resolve these conflicts. By appealing to this fact, these writers argue, we can justify these principles in the actual world, in which there has been no such agreement. We ought to treat each other as we would have rationally agreed to do. That is a plausible claim.

To justify certain principles in this way, however, we must defend the claim that everyone *would* have rationally reached agreement on these principles. And this claim would be hard to defend. When Gauthier discusses his proposed version of the Rational Agreement Formula, he tells us to 'suppose that after each party advances his initial claim, agreement is reached in a single round of concessions'. But we cannot simply *suppose* that such agreement would be reached. Given our need to reach agreement, it would be rational for each of us to try to predict which principles everyone else would choose, and to choose these principles ourselves. In some cases, each of us

might be able to predict what other people would choose. Suppose, for example, that we are trying to reach agreement on how some fixed set of resources would be shared between us. It might be uniquely rational for everyone to choose that everyone should get equal shares, since we could each predict that everyone else would make this choice. But when we are choosing most other moral principles, this coordination problem would have no such obvious solution. In trying to predict what other people would choose, each of us would be groping in the dark. So in the decisive final vote, there would be many conflicting principles that it would be equally rational for everyone to choose. The Rational Agreement Formula would then fail, since there would be no set of principles that everyone ought rationally to choose.

This version of Contractualism faces another objection. The No-Agreement World would be less bad for certain people, such as those who have greater abilities, and those who are rich in the non-legal sense that they control more resources. In a world without morality, people with such advantages would be better able to fend for themselves. As everyone would know, these people would have less need to reach this Contractualist agreement. That would give them greater bargaining power. These people could declare that, in the decisive final vote, they will choose certain principles that would allow them to keep their advantages, and would give them further benefits. Such threats might be credible, since these people would be more prepared than others to run the risk of bringing about the No-Agreement World. When certain questions were being discussed, moreover, it might be better for some people if there was no agreement. One example is the question of how much of their resources the rich ought to give to the poor. If there was no agreement on this question, so that no one accepted any principle about what the rich ought to give, that would be much the same as everyone's believing that the rich were permitted to give nothing. That might be fine with the rich. In these and similar ways, those who had greater bargaining power might be able to use that power to make it rational for others to accept principles that favoured these powerful people.

Some writers accept this implication of the Rational Agreement Formula. That is true of *Hobbesian* Contractualists, like Gauthier, who

defend only a minimal version of morality. Gauthier claims that, since morality presupposes mutual benefit, it would not be wrong for us to impose great harms on certain other people, if the existence of these people does not benefit us. On this view, for example, when Europeans founded colonies in North America, they were morally permitted to kill the native inhabitants. Nor can this view directly support requirements to care for some people who are congenitally handicapped. Such conclusions, Gauthier concedes, conflict strongly with most people's moral beliefs. But Gauthier rejects appeals to such intuitive beliefs, or to our 'considered moral judgments', which he claims that moral theories ought to ignore.

I have rejected Gauthier's claim that, when we apply the Rational Agreement Formula, it is Gauthier's minimal morality that everyone ought rationally to choose. As I argue in Appendix B, we ought also to reject Gauthier's assumptions about rationality. And we ought, I believe, to reject Gauthier's moral view. As Locke said of Hobbes, Gauthier's minimal morality does not admit 'a great many plain duties'. Similar claims apply, I believe, to other Hobbesian theories. Hobbesian Contractualists give unsound arguments for unacceptable conclusions.

51 Rawlsian Contractualism

Though Rawls also appeals to the Rational Agreement Formula, he defends more acceptable conclusions. Most of Rawls's claims are about the *justice* of what he calls the *basic structure*, or main institutions, of those societies that are nation-states. These claims are not relevant here. My remarks will only be about Rawls's Contractualist account of morality, which he calls *rightness as fairness*.

When applied to morality, I shall argue, Rawls's version of Contractualism fails. But if we removed the Contractualism from Rawls's great *Theory of Justice*, the result would be a liberal egalitarian view that is both in itself very appealing and well supported by some of Rawls's Non-Contractualist claims and arguments.

In considering Rawlsian Moral Contractualism, we can start with Rawls's assumptions about rationality and reasons. Rawls accepts a desire-based subjective theory, claiming that we ought rationally to try to achieve the aims that, after fully informed and procedurally rational deliberation, we would most want to achieve. Of those who accept this theory, many believe that it coincides with Rational Egoism, which claims that we ought rationally to try to do whatever would be best for ourselves. These people mistakenly assume that, after such deliberation, each of us would always care most about our own well-being in the rest of our lives as a whole.

Rawls does not make that assumption. He considers cases in which justice requires us to act in ways that would be bad for us. Even in such cases, Rawls claims, it might be rational for us to do what justice requires. We would be acting rationally if we would be doing what, all things considered, we most wanted to do. In his words,

> If a person wants with deliberative rationality to act from the standpoint of justice above all else, it is rational for him so to act.

Since Rawls's theory about reasons is desire-based, however, Rawls cannot claim that it would be rational for *everyone* to act justly. When he discusses people whose informed desires would be better fulfilled if they acted unjustly, Rawls claims that these people would not have sufficient reasons to do what justice requires.

On subjective theories, as I have argued, we cannot have reasons to want anything as an end, or for its own sake. If people don't care about something, and they would not care even after fully informed and procedurally rational deliberation, we cannot claim that they have reasons to care. Rawls would accept these claims. He also writes:

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently.

Similarly, when Rawls discusses the view that

> something is right . . . when an ideally rational and impartial spectator would approve of it,

he writes:

> Since this definition makes no specific psychological assumptions about the impartial spectator, it yields no principles to account for his approvals . . .

Rawls here assumes that we have no reasons to care about anything. If Rawls believed that we have such reasons, he would not claim that, if we knew only that someone was *ideally rational*, we could draw no conclusions about what this person would approve. Rawls's claim would instead be that, since this person was ideally rational, he would approve what he had most reason to approve. For example, he would approve of acts that relieved suffering, or saved people's lives.

As a Contractualist, Rawls appeals to the principles that it would be rational for everyone to choose, if we were all trying to reach agreement on the principles that we would all accept. On Rawls's desire-based theory, what it would be rational for people to choose depends on what they would in fact want. Since Rawls cannot predict what people would want, he adds a motivational assumption. He tells us to suppose that, when we were choosing moral principles, everyone's main aim would be to promote their own interests. On this assumption, Rawls's desire-based theory coincides with Rational Egoism. If we cared most about our own interests, it would be rational for us, according to desire-based or aim-based theories, to make the choices that we could expect to best promote these interests. Rawls's motivational assumption therefore allows him to appeal to claims about self-interested rationality. In his words,

> In choosing between principles each tries as best he can to advance his interests.

Rawls revises the Rational Agreement Formula by adding a *veil of ignorance*. According to

> *Rawls's Formula*: Everyone ought to follow the principles to whose universal acceptance it would be rational in self-interested terms for everyone to agree, if everyone had to reach this agreement without knowing any particular facts about themselves or their circumstances.

In explaining why he adds this veil of ignorance, Rawls appeals to the two objections to Hobbesian Contractualism mentioned above.

First, if everyone knew particular facts about themselves and their circumstances—such as their sex, age, abilities, and the resources that they control—we could not hope to work out what it would be rational for everyone to choose. In Rawls's words, 'the bargaining problem . . . would be hopelessly complicated'. There would be no principles to whose universal acceptance it would be rational for everyone to agree. Rawls's veil of ignorance solves this problem. If no one knew any of these facts about how they differed from other people, it would be rational for everyone to choose the same principles, so agreement would be guaranteed. It would be enough to ask what it would be rational for any one person to choose, since the same answer would apply to everyone.

Second, as Rawls points out, if we knew nothing about ourselves or our circumstances, that would make us impartial. We would not know the facts that might give us greater bargaining power. Nor could anyone choose principles that were biased in their own favour. Though we would be choosing principles for self-interested reasons, our ignorance would ensure that, in choosing principles, we would give equal weight to everyone's well-being.

One of Rawls's main aims, he writes, is to produce a systematic theory which provides an alternative to all forms of Utilitarianism. It is surprising that, in trying to achieve this aim, Rawls proposes his version of Moral Contractualism, which appeals to a combination of self-interested rationality and impartiality. We should expect such a theory to support some view that is, or is close to being, Utilitarian. As Rawls himself points out, Utilitarianism is, roughly, self-interested rationality plus impartiality.

Rawls is aware of this problem. According to one version of Rawls's Formula, when we imagine that we are behind the veil of ignorance, we would assume that we had an equal chance of being in anyone's position. On that assumption, Rawls claims, it would be rational for everyone to choose the principle whose acceptance would make the average level of well-being as high as possible. By choosing this *Utilitarian Average Principle*, each of us would maximize our own expectable level of well-being.

Rawls rejects what we can call this *Equal Chance Formula*. If we were behind the veil of ignorance, Rawls claims, we ought not to assume that we had an equal chance of being in anyone's position. According to Rawls's preferred version of his formula, which we can call the *No Knowledge Formula,* we would have no knowledge of the probabilities. That would make it rational for us, Rawls argues, to choose certain non-Utilitarian principles.

For Rawls's Contractualist theory to achieve his aims, he must defend his rejection of the Equal Chance Formula. When describing his veil of ignorance, Rawls writes

> there seem to be no objective grounds . . . for assuming that one has an equal chance of turning out to be anybody.

This remark treats our imagined state behind the veil of ignorance as if it would be some actual state of affairs, whose nature we would have to accept. But Rawls is proposing a thought-experiment, whose details he is free to choose. He could tell us to *suppose* that we have an equal chance of being anyone. So Rawls must give some other objection to the Equal Chance Formula. Rawls himself points out that, since there are different Contractualist formulas, which have different implications, he must defend his choice of his particular formula. This formula, he writes, must be the one that is 'philosophically most favoured', because it 'best expresses the conditions that are widely thought reasonable to impose on the choice of principles'. Could Rawls claim that, compared with the Equal Chance Formula, his No Knowledge Formula better expresses these conditions?

The answer, I believe, is No. Rawls's veil of ignorance is intended to ensure that, in choosing principles, we would be impartial. To achieve this aim, Rawls need not tell us to suppose that we have no knowledge

of the probabilities. If we supposed that we had an equal chance of being in anyone's position, that would make us just as impartial. Since there is no other difference between the Equal Chance and No Knowledge Formulas, Rawls's No Knowledge Formula cannot be claimed to be in itself more plausible.

When Rawls discusses what he calls the 'Kantian interpretation' of his theory, he suggests another defence of his No Knowledge Formula. Kantian Contractualism, Rawls writes,

> aims for the thickest possible veil of ignorance . . . The Kantian rationale . . . starts by allowing the parties no information and then adds just enough so that they can make a rational agreement.

By supposing that we know as little as possible, Rawls suggests, we would make our reasoning as similar as possible to the reasoning of our noumenal selves in Kant's timeless noumenal world, and we would thereby best express our freedom and autonomy.

This defence of the No Knowledge Formula does not, I believe, succeed. If we start by supposing that, behind Rawls's veil of ignorance, we would have *no* information, and we ought then to add *just enough* information to make a rational choice possible, we ought to appeal to a more extreme version of the No Knowledge Formula. In making our choices, for example, we need not know that different people have different abilities, or that we live in a world with scarce resources. Even if we did not know such facts, we could know enough to make a rational decision. We would then be closer to achieving Rawls's aim of 'the thickest possible veil of ignorance'. But this version of Contractualism could not be claimed to be the one that, in Rawls's words, 'best expresses the conditions that are widely thought reasonable to impose on the choice of principles'. We cannot reasonably require that those who are choosing moral principles be as ignorant as possible. It is *well*-informed not *ill*-informed choices to which we can more plausibly appeal. Rawls also writes that, on this Kantian version of his view, 'we start from no information at all; for by negative freedom Kant means being able to act independently from the determination of alien causes'. True beliefs are not well regarded as alien causes.

Remember next that, as Rawls claims, the Equal Chance Formula 'leads naturally' to the Utilitarian Average Principle. Since Rawls cannot justify his rejection of Equal Chance version of Rawlsian Contractualism, Rawls's theory does not, as he hopes, provide an argument against all forms of Utilitarianism.

Rawls might reply that we can have another kind of reason to reject some formula, or moral theory. We can justifiably reject some formula, however plausible it seems, if this formula's implications conflict too strongly with some of our best considered and firmest moral beliefs. Rawls assumes that Utilitarianism conflicts with some of these beliefs, such as the belief that slavery is always wrong. He might therefore claim that we can justifiably reject the Equal Chance Formula on the ground that, in leading to the Utilitarian Average Principle, this formula has unacceptable implications.

If Rawls made this claim, however, his Contractualism would still provide no argument against Utilitarianism. Rawls would be appealing to our non-Utilitarian beliefs to justify our rejecting the Equal Chance Formula and appealing to his No Knowledge Formula. So he could not also claim that, by rejecting the Equal Chance Formula and appealing to his No Knowledge Formula, we could justify our non-Utilitarian beliefs. If we defend some argument only by appealing to certain beliefs, we cannot then defend these beliefs by appealing to this argument. That defence would be circular, by assuming what it was trying to justify.

Rawls might next retreat to the claim that, though the Equal Chance Formula supports Utilitarianism, his No Knowledge Formula supports plausible non-Utilitarian principles. If that were true, Rawls's appeal to his formula would at least show that Veil of Ignorance Contractualists do not have to accept Utilitarian conclusions.

Rawls's Formula does not, however, support plausible non-Utilitarian principles. When he applies his formula, Rawls argues that, if we had no knowledge of the probabilities, we ought rationally to assume the worst, and try to make our worst possible outcome as good as possible. We ought therefore to choose the principles whose acceptance would make the worst off people as well off as possible. Since this argument tells us to *max*imize the *min*imum level of well-being, we can call it the *Maximin Argument*.

This argument has been widely criticized. Even if it were sound, however, it would not support an acceptable non-Utilitarian moral view. Suppose first that we must decide how to use some scarce medical resources, treating various young people who all have some disease. In one of two possible outcomes,

> *Blue* would live to the age of 25, and a thousand other people would all live to 80.

In the other outcome,

> *Blue* would live to 26, and these other people would all live to 30.

People would be relevantly worse off, we can next suppose, if their lives would be shorter. On the Maximin Argument, we ought then to choose the second of these outcomes, giving Blue her extra year of life, since that is what would be best for the person who would be worst off. That is an indefensible conclusion. Though we can plausibly give some priority to benefiting those people who would be worse off, this priority should not be absolute. It would be wrong to give Blue one more year of life, rather than giving fifty more years to each of a thousand other people—people who, without these extra years, would all die almost as young as Blue. When applied to this and many other cases, the Maximin Argument has implications that are much too extreme.

Rawls accepts what I have just claimed. Though he applies his Maximin Argument to the basic structure of society, Rawls agrees that, when we apply this argument to other questions about distributive justice, this argument's implications are much too extreme. Utilitarian theories, Rawls claims, fail to provide an acceptable general principle of distributive justice. But as Rawls admits, his version of Contractualism also fails to provide such a principle.

We can now turn to other moral questions. On Rawls's Maximin Argument, when we choose between different moral principles, we ought rationally to choose the principles whose acceptance would be best for those who would be worst off. There are many moral questions to which, even if it were sound, the Maximin Argument could not be plausibly applied. Suppose that we are comparing different principles

about when we could justifiably fail to keep our promises, or tell lies, or impose risks on other people. It would be hard to decide which are the principles about such questions whose acceptance would be best for the worst off people. Nor could this be the right way to choose between such principles. Suppose that, if we all accepted one of two forms of the practice of promising, or one of two principles about imposing risks on others, that would give much greater benefits to most people. These facts would not be, as the Maximin Argument implies, morally irrelevant.

Even if Rawls did not appeal to this argument, there is another way in which Rawls's Formula fails to support plausible non-Utilitarian principles. Rawls's version of Contractualism forces us to ignore most non-Utilitarian considerations. According to Utilitarians, when we are choosing between acts or principles, it is enough to know the size and number of the resulting benefits and burdens. Most of us believe that there are several other morally important facts and considerations. We have such beliefs, for example, about how benefits and burdens should be distributed between different people, and about responsibility, desert, deception, coercion, fairness, gratitude, and autonomy. When we apply Rawls's version of Contractualism, all such considerations are irrelevant, except insofar as they affect our own well-being. Though Rawlsian moral reasoning differs from Utilitarian reasoning, it differs only by subtraction. When Rawls describes how people would choose moral principles from behind his veil of ignorance, he writes that they

> decide solely on the basis of what best seems calculated to further their interests so far as they can ascertain them.

Rawls merely denies these people most of the knowledge that self-interested calculations need. Since Rawls's imagined contractors choose principles for purely self-interested reasons, there is no way in which non-Utilitarian considerations could possibly enter in.

When he first presents his theory, Rawls writes

> It is perfectly possible . . . that some form of the principle of utility would be adopted, and therefore that contract theory leads eventually to a deeper and more roundabout justification of Utilitarianism.

He also writes

> for the contract view, which is the traditional alternative to Utilitarianism, such a conclusion would be a disaster.

Rawls might be able to deny that his version of Contractualism justifies any form of Utilitarianism. But his claim would have to be that, even if his theory led to some Utilitarian conclusion, it is not plausible enough to justify this conclusion.

52 Kantian Contractualism

To reach a more plausible and successful version of Contractualism, we should return to a different formula, and a different view about reasons and rationality. According to

> the Kantian Contractualist Formula: Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.

Remember next that, according to

> the Rational Agreement Formula: Everyone ought to follow the principles to whose universal acceptance it would be rational for everyone to agree.

These formulas both require unanimity, since they both appeal to the principles whose universal acceptance everyone could rationally choose. But unlike the Rational Agreement Formula, the Kantian Formula does not use the idea of an *agreement*. When we apply the Agreement Formula, we imagine that we are all trying to reach agreement on which principles everyone would accept. Such agreement would be needed, since everyone would accept only the principles that, in this *single* thought-experiment, *everyone* chose. According to the Kantian Formula, in contrast,

> Everyone ought to follow the principles that everyone could rationally choose, if each person supposed that everyone would accept the principles that *he* or *she herself* chose.

In applying this formula, we carry out *many* thought-experiments, one for each person. In making these separate choices, none of us would need to reach agreement with other people, since each of us would have this power to choose which principles everyone would accept. The Kantian Formula requires unanimity in a quite different way. This formula appeals to the principles that, in these many separate thought-experiments, everyone would have sufficient reasons to choose.

Though Rawls rightly rejects the Rational Agreement Formula, the Kantian Formula is, I believe, more plausible than Rawls's Formula, and better achieves Rawls's aims.

Rawls's veil of ignorance is in part intended to eliminate inequalities in bargaining power. The Kantian Formula achieves this aim in a better way. Since there is no need to reach agreement, there is no scope for bargaining, so no one would have greater bargaining power. When we ask which principles everyone could rationally choose, we can therefore suppose that everyone knows all of the relevant, reason-giving facts, and could therefore respond to all these reasons.

Consider next one of Rawls's reasons for rejecting Utilitarianism. Utilitarians believe that it would be right to impose great burdens on a few people, whenever such acts would give a greater sum of benefits to others. In such cases, Rawls claims, justice

> does not allow that the sacrifices imposed on the few are outweighed by the larger sum of advantages enjoyed by the many.

According to several writers, Utilitarians reach such unacceptable conclusions because they merely add together different people's benefits and burdens. In Nagel's phrase, different people's claims are all 'thrown into the hopper', and merged into an impersonal sum. Some of these writers suggest that, to protect people from having such great burdens imposed on them, we should appeal instead to the idea of a unanimous agreement. On this proposal, by requiring such an agreement, we give everyone a *veto* against being made to bear such burdens, thereby achieving what we can call the *anti-Utilitarian protective aim*.

Vetos, however, can be misused. Precisely *by* requiring such unanimous agreement, the Rational Agreement Formula makes it harder to achieve this protective aim. This formula gives further advantages, not to those who *most* need morality's protection, but to those who *least* need such protection, because their greater abilities, or their control of more resources, gives them greater bargaining power.

Rawls's Formula does little to achieve this protective aim. Though Rawls's veil of ignorance eliminates bargaining power, it also prevents anyone from knowing whether they are one of the few people on whom some Utilitarian principle would require or permit us to impose great burdens. Rawls appeals to the principles whose choice would be rational in self-interested terms. And as I have claimed, Rawls has no relevant objection to the Equal Chance Formula. So he cannot plausibly deny that, from behind the veil of ignorance, we could rationally choose some Utilitarian principle, or some similar but somewhat more cautious principle, running the small risks of bearing some great burden for the sake of much more likely benefits.

The Kantian Formula requires unanimity without appealing either to a veil of ignorance or to a need to reach agreement. Partly for this reason, this formula better achieves the protective aim. If Utilitarians appealed to this formula, they would have to claim that we could rationally choose their principle even if we knew that we were one of the few people on whom these great burdens would be imposed. In at least some cases, we could plausibly reject this claim.

The Kantian Formula has other advantages. Though Rawls's veil of ignorance ensures impartiality, it does that crudely, like frontal lobotomy. The disagreements between different people are not resolved, but suppressed. Since no one knows anything about themselves or their circumstances, unanimity is guaranteed. In the thought-experiments to which the Kantian Formula appeals, there is no veil of ignorance. Everyone would know how their interests conflict with the interests of others. Since unanimity is not guaranteed, it would be morally more significant if unanimity *could* be achieved, because there are some principles that, even with full information, everyone could rationally choose.

Whether there are such principles depends on what we ought to believe about reasons and rationality. If the best theory were either

Rational Egoism, or some desire-based or aim-based subjective theory, the Kantian Formula would not succeed. In the thought-experiments to which this formula appeals, there would be no set of principles whose choice would be rational for everyone in self-interested terms. Nor would there be some set of principles whose universal acceptance would best fulfil everyone's fully informed desires or aims.

We ought, I believe, to reject all subjective theories. And though Rational Egoism is, in being objective and value-based, a theory of the right kind, this theory is too narrow. According to objective theories of the kind that I believe to be the truest or best, we have strong reasons to care about our own well-being, and in a temporally neutral way. But our own well-being is not, as Rational Egoists claim, the one supremely rational ultimate aim. We could rationally care as much about some other things, such as the well-being of others.

Return next to the fact that, since Rawls appeals to the principles that it would be rational to choose for self-interested reasons, there is no way in which, when we apply the Rawlsian Formula, non-Utilitarian considerations can enter in. When we apply the Kantian Formula, we can appeal to every kind of non-deontic reason, so this formula can support non-Utilitarian principles.

For the Kantian Formula to succeed, what we can call its *uniqueness condition* must be sufficiently often met. It must be true that, at least in most cases, there is some relevant principle, and only one such principle, that everyone could rationally choose. If there was no such principle, there would be no principle that the Kantian Formula would require us to follow. This formula might then fail, by failing to disallow acts that are clearly wrong. If everyone could rationally choose two or more seriously conflicting principles, this formula might again fail, in similar though more complicated ways. It would not matter, though, if everyone could rationally choose any of several similar principles. Such principles would be different versions of some more general, higher-level principle, and the choice between these lower-level principles could then be made in some other way. The uniqueness condition would, I believe, be sufficiently often met.

To illustrate the Kantian Formula, we can apply it to an easy question. Suppose that

> some quantity of unowned goods can be shared between different people,
>
> no one has any special claim to these goods, such as a claim based on their having greater needs, or their having produced these goods, or their being worse off than others,

and

> if these goods were equally distributed, that would produce the greatest sum of benefits.

It is clear that, in such cases, everyone should be given equal shares. Kantians might argue:

> (A) Everyone could rationally choose the principle that, in such cases, gives everyone equal shares.
>
> (B) No one could rationally choose any principle that gave them and the other people in some group less than equal shares.
>
> (C) Only the principle of equal shares gives no one less than equal shares.

Therefore

> (D) This is the only principle that everyone could rationally choose.

If we accept Rational Egoism, we must reject this argument's first premise. On this theory, everyone ought rationally to choose some principle that gave to themselves more than equal shares. We must also reject (A) if we accept a subjective theory about reasons. There are many people whose fully informed desires or aims would not be best

fulfilled by their choosing the principle of equal shares. But I believe that, as (A) claims, everyone could rationally choose this principle, since we would all have sufficient reasons to make this choice. We would not be rationally required to choose some principle that gave us *more* than equal shares. As (B) claims, no one could rationally choose any principle that gave them and the other people in some group *less* than equal shares, thereby producing a smaller sum of unequally distributed benefits. As (C) claims, only the principle of equal shares gives no one less than equal shares. So, as this argument shows, this is the only principle that everyone could rationally choose. The Kantian Formula rightly implies that, in such cases, everyone should be given equal shares.

53 Scanlonian Contractualism

We can now introduce another version of Contractualism. According to

> *Scanlon's Formula*: Everyone ought to follow the principles that no one could reasonably reject.

In a fuller statement:

> Some act is wrong just when such acts are disallowed by some principle that no one could reasonably reject, or when any principle permitting such acts could be reasonably rejected by at least one person.

Though 'reasonable' sometimes means the same as 'rational', Scanlon's Formula uses this word in a different, partly moral sense. We are unreasonable in this sense if we give too little weight to other people's well-being or moral claims.

Some people claim that, because Scanlon appeals to this partly moral sense of 'reasonable', his formula is empty. If we accepted Scanlon's Formula, these people say, that would make no difference to our moral thinking, since everyone could claim that the moral principles which they accept could not be reasonably rejected.

This objection overlooks the fact that, when we apply some Contractualist formula, we cannot appeal to our beliefs about which acts are wrong. Suppose again that, in

> *Means*, Grey and White trapped in slowly collapsing wreckage. Grey is in no danger. I could save White's life, but only by using Grey's body as a shield, without her consent, in some way that would destroy Grey's leg.

Many people would believe that it would be wrong for me to save White's life in this way. If we accept this view, we might appeal to

> *the Harmful Means Principle*: It is wrong to impose such a serious injury on someone as a means of benefiting other people.

According to another, conflicting view, which we can call

> *the Greater Burden Principle*: We are permitted to impose a burden on someone if that is the only way in which someone else can be saved from some much greater burden.

Scanlon makes various claims about what would be reasonable grounds for rejecting moral principles. According to one such claim,

> it would be unreasonable . . . to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others.

We impose a burden on someone, in Scanlon's intended sense, when we fail to give this person some benefit. White could argue that, as Scanlon's claim implies, Grey could not reasonably reject the Greater Burden Principle. Though my acting on this principle would impose a burden on Grey, my acting on the Harmful Means Principle would impose a much greater burden on White. Losing a leg is a much smaller burden than failing to have our life saved.

Grey might reply that, in her opinion, White could not reasonably reject the Harmful Means Principle. But why would this rejection be unreasonable? Grey might say that she has a right not to be seriously injured without her consent as a means of benefiting someone else. But in claiming that she has this right, Grey would be implicitly appealing to her belief that it would be wrong for me to injure her in this way.

When we apply Scanlon's Formula, we cannot appeal to such *deontic* beliefs. Grey might claim that

> (1) my act would be wrong, because no one could reasonably reject the Harmful Means Principle, which disallows such acts.

But Grey could not defend (1) with the claim that

> (2) no one could reasonably reject this principle because such acts are wrong.

As I have said, if we combined such claims, we would be going round in a circle, getting nowhere. Grey must argue in some other way that no one could reasonably reject the Harmful Means Principle.

As this example shows, Scanlonian Contractualism is far from being empty. When White rejects the Harmful Means Principle, White can appeal to the fact that, compared with losing a leg, dying is a much greater burden. This is one of the kinds of fact that, on Scanlon's view, can provide reasonable grounds for rejecting some moral principle. When Grey defends the Harmful Means Principle, she cannot appeal to any such fact. Grey's problem is that, unlike the Greater Burden Principle, the Harmful Means Principle is best defended by appealing to our intuitive beliefs about which acts are wrong. Many of us would believe it to be wrong to inflict a serious injury on someone, without this person's consent, even when that is our only way to save someone else's life. But when we apply Contractualist formulas, we cannot appeal to such beliefs.

Like Rawls, Scanlon proposes his Contractualism partly as a way of avoiding Act Utilitarianism, or *AU*. In one way, as we have just seen, Contractualism makes AU easier to defend. Most of us reject AU because this view requires or permits many acts that seem to us to be wrong. As Scanlon writes,

> the implications of Act Utilitarianism are wildly at variance with firmly held moral convictions.

But when we apply some Contractualist formula, and follow the Deontic Beliefs Restriction, we cannot appeal to such convictions.

Even without appealing to such convictions, however, Scanlonian Contractualists can reject Act Utilitarianism. To illustrate Scanlon's Formula, it is worth considering some examples. Suppose that, in

> *Transplant,* I am in hospital, to have some minor operation. You are my doctor. You know that, if you secretly killed me, my transplanted organs would be used to save the lives of five other people.

According to

> AU: We ought always to do, or try to do, whatever would benefit people most.

This principle requires you to save these five people by killing me, since that is how you would benefit people most. Most of us would believe this act to be wrong.

We can plausibly defend this belief by appealing to one version of Scanlon's Formula. Suppose we all knew that, whenever we were in hospital, our doctors might secretly kill us so that our organs could be used to save other people's lives. Even if that risk would be very small, this knowledge would make many of us anxious, and would worsen our relation with our doctors. This relation is of great importance, since we often rely on the judgment of our doctors, and their concern for our well-being, and they may be people whom we expect to help us through the ending of our lives. By appealing to such facts, we could reasonably reject AU. If all doctors followed this principle in such cases, a few more people's lives would be saved. But the saving of these extra lives would be outweighed by these ways in which it would be bad for us and others if, as we all knew, our doctors believed that it could be right to kill us secretly in this way. We can call this the *Anxiety and Mistrust Argument*.

This argument illustrates another way in which, if we appeal to a Contractualist formula, this makes a difference to our moral reasoning. If we consider *Transplant* on its own, we could ignore this argument. Since you could save the five by secretly killing me, your act would produce no anxiety or mistrust. But when we apply some Contractualist

formula, such as the Kantian or Scanlonian Formulas, we don't consider particular acts on their own. We ask which are the principles that everyone could rationally choose, or that no one could reasonably reject, if we were choosing the principles that everyone would accept. In answering *this* question, we must take into account the effects of everyone's accepting, and being known to accept, these principles. That makes it irrelevant that, in *Transplant*, your act would be secret, and would therefore produce no anxiety or mistrust.

We can reasonably reject some principle, Scanlon claims, only if we can propose some better alternative. If we reject AU, what alternative should we propose?

It may help to compare *Transplant* with two other cases. Suppose again that, in

> *Tunnel*, by switching the points on some track, you could redirect a driverless, runaway train, so that it kills me rather than five other people,

and that in

> *Bridge*, you could save the five only by using remote control to make me fall in front of the train, thereby killing me, but also triggering the train's automatic brake.

For one alternative to AU, we might return to

> *the Harmful Means Principle*: It is wrong to impose a great injury on one person as a means of benefiting other people.

What is morally important, on this view, is how your saving of the five would be causally related to the act with which you kill me. It would be wrong for you to save the five in both *Transplant* and *Bridge by* killing me. But it would not be wrong for you to kill me in *Tunnel*, since you would here be killing me, not as a means of saving the five, but only as the foreseen side-effect of redirecting the train. Many of us would accept these claims, believing my act to be wrong in *Bridge* but permissible in *Tunnel*. When we apply Scanlon's Formula, can we plausibly defend this distinction?

The answer, I suggest, is No. When we consider cases like *Tunnel* and *Bridge*, we have strong reasons to care whether we would live or die, but no strong reasons to care how our death might be causally related to the saving of other people's lives. In making this claim, I am not assuming that only outcomes matter. We can have reasons to care about how some outcomes are produced. But when someone else could act in some way that would both kill us but also save several other people's lives, we would have no strong reason to prefer to be killed as a side-effect of the saving of these people's lives rather than as a means. It would be in one way better to be killed as a means, since our death would then at least do some good. Given these facts, Scanlon's Formula seems to count against the view that there is an important moral difference between your acts in *Tunnel* and *Bridge*. If I could *not* reasonably reject some principle that would permit you to kill me in *Tunnel*, it seems doubtful that I could reasonably reject every principle that would permit you to kill me in *Bridge*. Scanlon's Formula seems to imply that these acts are either both wrong, or both morally permitted.

Consider next another alternative to AU, which is suggested by the anxiety and mistrust argument. According to what we can call

> *the Emergency Principle*: Doctors must never kill their patients as a means of saving more lives. In certain *non-medical emergencies*, however, everyone is permitted to do whatever would save the most lives.

These non-medical emergencies are cases that involve unintended and immediate threats to people's lives, such as some fire, flood, avalanche, or driverless run-away train. The Emergency Principle condemns your saving the five by killing me in *Transplant*, since you are here my doctor. But this principle permits you to save the five in a way that kills me, in both *Tunnel* and *Bridge*, because these are non-medical emergencies, and in these cases I would be a stranger to you.

Compared with the Harmful Means Principle, Scanlon's Formula seems more strongly to support the Emergency Principle. What is morally important, this principle assumes, is not the *causal* relation between your saving of the five and your killing of me, but the *personal* relation between you and me in *Transplant*, and the other differences between medical and non-medical emergencies. These are the kinds of

fact to which, when applying Scanlon's Formula, we can more plausibly appeal. We have reasons to want our doctors to believe that they must never kill their patients as a means of saving other people's lives—or, we can add, even as a side-effect. While our relation to our doctors is of great importance, we have no such personal relation to those who might kill us or save our lives in these rare non-medical emergencies. And we have reasons to want such people to believe that, in such cases, they ought to save as many lives as possible. We would know that, if our lives were threatened in such an emergency, we would be more likely to be one of the people whose lives would be saved.

54 The Deontic Beliefs Restriction

Suppose that, after thinking hard about these imagined cases, we believe that you would be morally permitted to kill me, in *Tunnel*, as a foreseen side-effect of saving the five, but that it would be wrong for you, in *Bridge*, to kill me as a means. We may then accept the Harmful Means Principle, which draws this distinction. Suppose next that, for the reasons I have just given, we cannot successfully defend this principle by appealing to Scanlon's Formula. This and other similar principles are best defended by appealing to our intuitive beliefs about which acts are wrong. But when we apply Contractualist formulas, we cannot appeal to these beliefs. Nor can we appeal to these beliefs when we apply Kant's Formula of Universal Law.

We might now challenge this Deontic Beliefs Restriction. When we try to answer moral questions by applying these Kantian or Contractualist formulas, why should we ignore our beliefs about which acts are wrong?

Kantians and Contractualists might reply that, if we appealed to such deontic beliefs, their formulas would be circular, in a way that made them useless. As I have said, there is no point in claiming both that

> acts are wrong when any principle permitting them would fail some Kantian or Contractualist test,

and that

> principles would fail this test when and because the acts they permit are wrong.

But this is not a good enough reply. Even if these formulas would be useless unless we follow the Deontic Beliefs Restriction, that does not show that we ought to think about morality by applying these formulas.

Another reply appeals to a distinction that is *meta-ethical*, in the sense that it makes claims about the nature and justifiability of moral beliefs and claims. According to *Intuitionists*, Rawls writes, there are certain independent truths about which acts are wrong, and about which facts give us reasons. Two examples are the truths that slavery is wrong, and that we have reasons to prevent or relieve suffering. These truths are *independent* in the sense that they are not created or constructed by us. According to a different view, which Rawls calls *Constructivism*, there are no such truths. On this view, what is right or wrong depends entirely on which principles it would be rational for us to choose in some Kantian or Contractualist thought-experiment. In Rawls's phrase, it's for us to decide what the moral facts are to be. If we are Constructivist Contractualists, and we believe that it would be rational to choose principles that permit slavery, we ought to conclude that slavery is not wrong. Though slavery may seem to us to be wrong, Constructivists reject appeals to such intuitive moral beliefs, which some of them claim to involve prejudice, or cultural conditioning, or to be mere illusions.

I shall here assume that we ought to reject these *sceptical*, anti-intuitionist views. Rawls does not commit himself to Constructivism, and he often assumes that there are some independent moral truths, such as the truth that slavery is wrong. When we try to achieve what Rawls calls reflective equilibrium, we should appeal to all of our beliefs, including our intuitive beliefs about the wrongness of some kinds of act. As Scanlon writes:

> this method, properly understood, is . . . the best way of making up one's mind about moral matters . . . Indeed, it is the only defensible method: apparent alternatives to it are illusory.

If Kantians and Contractualists accept that our moral reasoning should appeal to such intuitive beliefs, they must defend the Deontic Beliefs Restriction in some other way.

There is one straightforward and wholly satisfactory defence. In describing this defence, we can first distinguish between two senses in which some property of an act, or some fact about this act, might make this act wrong. When some property of an act makes this act wrong, it does not *cause* it to be wrong. In one trivial sense, wrongness is the property that *non-causally* makes acts wrong. That is like the sense in which blueness is the property that makes things blue, and illegality is the property that makes acts illegal. It is in a different and highly important sense that when acts have certain other properties—such as that of causing pointless suffering, or being a lying promise—these facts may non-causally make these acts wrong. Causing pointless suffering isn't the same as being wrong. But if some act causes pointless suffering, this fact may make this act wrong by making it have the different property of being wrong. Moral theories should try to describe the properties or facts that, in this sense, can make acts wrong.

Scanlon once claimed that his Contractualism gives an account, not of what *makes* acts wrong, but of wrongness itself, or of *what it is* for acts to be wrong. This claim was, I believe, a mistake. To see why, we can first restate the Kantian Contractualist Formula. According to

> KF2: An act is wrong just when such acts are disallowed by one of the principles whose universal acceptance everyone could rationally will.

Suppose next that, in

> the *Kantian* sense, 'wrong' means 'disallowed by the principles whose universal acceptance everyone could rationally will'.

If Kantian Contractualists used 'wrong' in this sense, they could claim to be giving an account of one kind of wrongness. On this view, when acts are disallowed by such a principle, that's *what it is* for these acts to be wrong in this Kantian sense. But KF2 would then be a *concealed tautology*, one of whose open forms would be

> KF3: An act is disallowed by such a principle just when such acts are disallowed by such a principle.

And this claim is not worth making. Kantian Contractualists ought instead to use 'wrong' in one or more non-Kantian senses. KF2 would not then be trivial, since this claim would mean that, when some act is disallowed by such a principle, that makes this act wrong in such other senses. For example, Kantian Contractualists might claim

> KF4: When some act is disallowed by one of the principles whose universal acceptance everyone could rationally will, that makes this act wrong in the senses of being unjustifiable to others, blameworthy, and an act that gives its agent reasons to feel remorse and gives others reasons for indignation.

If we are Kantian Contractualists, we should not claim that our formula describes the *only* property or fact that makes acts wrong in these other senses. There are other wrong-making properties or facts that would often have more importance. Our claim should instead be that this formula describes a *higher-level* wrong-making property or fact, under which all other such properties or facts can be subsumed, or gathered. When some act is a lying promise, for example, this fact may make this an act that is disallowed by one of the principles whose universal acceptance everyone could rationally will. According to this version of Kantian Contractualism, both of these facts could then be truly claimed to make this act wrong.

Scanlon's theory should, I believe, take the same form. According to

> Scanlon's Formula: An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject.

If Scanlon was here using 'wrong' in another Contractualist sense, to mean 'disallowed by such an unrejectable principle', he could claim that his formula gives an account of this Contractualist kind of wrongness, or of *what it is* for acts to be wrong in this sense. But his formula would then be another concealed tautology, one of whose open forms would be the claim that acts are disallowed by such unrejectable principles

just when these acts are disallowed by such principles. We could all accept that claim, whatever our moral beliefs. Scanlon's claim should instead be that, if some act is disallowed by some principle that could not be reasonably rejected, that makes this act wrong in one or more non-Contractualist senses.

Scanlon now accepts that his view should take this form. We can therefore say that, on Scanlon's theory, when acts have certain other properties, that makes these acts disallowed by some unrejectable principle, and these facts can all be truly claimed to make these acts in other senses wrong.

If Contractualists make such claims, they can defend the Deontic Beliefs Restriction without rejecting our moral intuitions as worthless. On these versions of Contractualism, it is only *while* we are asking what Contractualist formulas imply that we should not appeal to our beliefs about the wrongness of the acts that we are considering. We can appeal to these beliefs at a later stage, when we are deciding whether we ought to accept these formulas. As when considering any other claim about which acts are wrong, we could justifiably reject any Contractualist formula if this formula's implications conflict too often and too strongly with our intuitive moral beliefs.

On this version of Scanlon's view, he does not reject appeals to our intuitive beliefs. On the contrary, Scanlon shows that, as well as having such beliefs about which acts are wrong, we have and can usefully appeal to intuitive beliefs about what are reasonable grounds for rejecting moral principles. That is Scanlon's greatest contribution to our moral thinking.

16

Consequentialism

55 Consequentialist Theories

Before we ask what is implied by the best versions of Contractualism, it may help to return to the relation between what is good and what is right.

Pain is bad, some of us truly believe, in the sense of being something that we have reasons to want to avoid. But some great philosophers did not have such beliefs. Hume, for example, does not use 'good' or 'bad' in reason-implying senses. This may be why Hume claims that it cannot be unreasonable, or contrary to reason, to prefer our own acknowledged lesser good to our greater good. If Hume had used 'lesser good' to mean 'what we have less reason to prefer', he could not have believed that no such preference could be unreasonable. Hume often uses 'good' and 'evil' merely to mean 'pleasure' and 'pain'.

While Hume would have thought it trivial to claim that pain is evil, Kant sometimes rejects this claim. For example, Kant writes:

> good or evil is, strictly speaking, applied to actions, not to the person's state of feeling . . . Thus one may laugh at the Stoic who in the most intense pains of gout cried out, 'Pain, however much you torment me, I will still never admit that you are something evil (*kakon*, *malum*)', nevertheless, he was right.

When Kant claims that pain cannot be evil, he means that pain cannot be morally bad. Like Hume, Kant seems sometimes to be unaware of, or to forget, the reason-implying sense in which it is bad to be in pain.

So does Ross. If some event would be bad, Ross assumes, we have a prima facie duty to prevent this event, if we can. Because we have no such duty to prevent ourselves from being in pain, Ross concludes that our own pain is not bad. More exactly, Ross, suggests, our pain *is* bad, but only from other people's point of view. Ross reaches this strange conclusion because he ignores the reason-implying senses in which things can be non-morally good or bad.

As well as being bad *for* the person who is in pain, pain is also *impersonally* bad. In Nagel's words, 'suffering is a bad thing, period, and not just for the sufferer'. Many people believe that, though outcomes can be good or bad for particular people, there is no sense in which outcomes could be impersonally good or bad. But, as I have said, we can explain such a sense. When we are comparing different possible outcomes, and we claim that some outcome would be

> *impersonally best* in the *impartial-reason-implying* sense, we mean that this is the outcome that, from an impartial point of view, everyone would have most reason to want, or to hope will come about.

When we consider possible events that would involve and affect only strangers, our actual point of view is impartial. But we also have impartial reasons when our point of view is not impartial, as is true, for example, when we could relieve either our own or someone else's pain. All pain is bad in the sense that we all have reasons to regret anyone's being in pain, whatever that person's relation to us. And we all have reasons to want everyone's life to go well.

If we accept some subjective theory about reasons, or Rational Egoism, we must deny that outcomes could be in this sense good or bad. On these theories, there are no outcomes that everyone has some reason to want, or to regret. It could not be in this sense bad if some plague or earthquake killed many people, since this outcome would not be bad for everyone, nor would everyone have desire-based or aim-based reasons to want such people not to be killed. But we ought, I have argued, to reject these theories.

In what follows, I shall use 'best' in the impartial-reason-implying sense. There are often two or more possible outcomes that might be

called 'equal-best'. Since that phrase misleadingly suggests precision, it would be better to call such outcomes not worse than any of the others. To save words, however, I shall use 'best' to refer to all such outcomes.

Though any plausible moral theory could appeal to facts about the goodness of outcomes, certain theories take such facts to be fundamental. According to what I am now calling

> *Consequentialism*: Whether our acts are right or wrong depends only on facts about how it would be best for things to go.

Consequentialist theories can differ in several ways, since they can make conflicting claims both about what is good and bad, and about how the rightness of our acts depends on facts about what would be best.

Some Consequentialists are *Utilitarians*, who believe that

> (A) things go best when they go in the way that would, on the whole, benefit people most, by giving them the greatest total sum of benefits minus burdens.

Other Consequentialists believe that the goodness of outcomes depends in part on other facts. Some people, for example, believe that

> (B) how well things go depends in part on how benefits and burdens are distributed between different people.

On two such views, one of two outcomes might be better, though it would involve a smaller sum of benefits minus burdens, because these benefits and burdens would be more equally distributed, or because more of the benefits or fewer of the burdens would go to people who were worse off.

The word 'Consequentialist' is in one way misleading, as is talk of the goodness of *outcomes* and of the acts that *make* things go best. These words suggest that, on these theories, all that matters is the future, and the effects of our acts. Consequentialists can reject those claims. The goodness of some outcomes might depend in part on facts about the past. It might be better, for example, if benefits went to people who

had earlier been worse off, or if we kept our promises to those who are dead, or if people are punished only if they earlier committed some crime. And some acts, intentions, and motives may be in themselves good or bad, whatever their effects. Kind acts may be good, for example, even when they fail, and it may often be in itself bad when people are deceived or coerced. When we ask whether it would be best if something happened, or if someone acted in some way, we are asking what, from an impartial point of view, everyone would have most reason to want, or to hope. This sense of 'best' leaves it entirely open which are the ways in which we would have most reason to want things to go.

There is, however, one kind of value to which Consequentialist theories cannot appeal. Some Consequentialists believe that

> (C) when people act rightly for the right reasons, these acts are in themselves good, and wrong acts are in themselves bad.

As I explain in a note, the rightness or wrongness of our acts cannot depend on whether these acts are in these ways good or bad. But that is not a serious objection to these theories.

All Consequentialists appeal to claims about what would make things go best. We can call this the *Consequentialist Criterion. Direct* Consequentialists apply this criterion directly to everything: not just to acts, but also to rules, laws, customs, desires, emotions, beliefs, the distribution of wealth, the state of the Earth's atmosphere, and anything else that might make things go better or worse. When these people apply this criterion to acts, they are *Act Consequentialists*. Some of these people claim that

> (D) everyone ought always to do whatever would in fact make things go best.

Others claim that

> (E) everyone ought always to do, or try to do, whatever would be most likely to make things go best, or more precisely what would make things go *expectably-best*.

If (D) uses 'ought' in the fact-relative sense and (E) uses 'ought' in the evidence-relative or belief-relative senses, these claims do not conflict.

In most of what follows we can ignore the difference between these claims. And I shall often use 'best' to mean 'best or expectably-best'.

Indirect Consequentialists apply the Consequentialist Criterion directly to some things but only *indirectly* to others. *Rule Consequentialists* apply this criterion directly to rules or principles, but only indirectly to acts. Some of these people believe that

> (F) everyone ought to follow the principles whose universal acceptance would make things go best.

On this view, though the best principles are the ones whose universal acceptance would make things go best, the best or right acts are not the acts that would make things go best, but the acts that are required or permitted by the best principles. It would be wrong to do what would make things go best when such acts are disallowed by one of the best principles. *Motive Consequentialists* similarly claim that, though the best motives are the ones whose being had by everyone would make things go best, the best or right acts are not the acts that would make things go best, but the acts that would be done by people with the best motives. These theories overlap with those systematic forms of *virtue ethics* which appeal to the character-traits and other dispositions that best promote human flourishing or well-being. There could be many other forms of Indirect Consequentialism.

56 Consequentialist Maxims

Some Consequentialists might apply their criterion directly to maxims, and only indirectly to acts. Of the possible maxims on which everyone might act, some would be

> *optimific* in the sense that, if everyone acted on these maxims, things would go in the ways that would be impartially best.

According to what we can call

> *Maxim Consequentialism*: Everyone ought to act only on these optimific maxims.

It is worth returning briefly to one of Kant's formulas. Some Kantians might argue:

> (G) Each of us is permitted to act on some maxim if we could rationally will it to be true that everyone acts on this maxim.
>
> (H) Some people could rationally will it to be true that everyone acts on the optimific maxims.

Therefore

> These people are permitted to act on these maxims.

(G) is Kant's Law of Nature Formula. If (H) is true, Kant's formula permits some people to be Maxim Consequentialists, who act on these optimific maxims.

In assessing this argument, we must appeal to some view about reasons and rationality. According to wide value-based objective views of the kind that I believe we should accept, (H) is true. If everyone acted on the optimific maxims, things would go in ways that would both be impartially best and be best *for* some people. These *fortunate* people would have both impartial and personal reasons to will it to be true that everyone acts on these maxims, and at least some of these people would not have any stronger conflicting reasons.

When we apply Kant's formula, some writers claim, we ought to appeal only to a rational requirement to avoid inconsistency, or contradictions in our will. On this assumption, (H) is true. There would be some people who could rationally will it to be true that everyone acts on the optimific maxims, since that would involve no inconsistencies or contradictions in these people's wills. Other writers claim that we are rationally required to will what would best fulfil our true needs as rational agents. On this assumption, there would again be some fortunate people who could rationally will it to be true that everyone acts on the optimific maxims. Things would go best in such a world in part because many people's true needs as agents would be best fulfilled.

(H) is also true on subjective theories about reasons. Of the fortunate people, some would care strongly about the well-being of others, and would want things to go in the ways that would be best. Some of these

people would have desires that would be best fulfilled if everyone acted on the optimific maxims.

Rational Egoists might reject (H). We are rationally required, these people believe, to choose whatever would be best for ourselves. It would be best for each person, Rational Egoists might claim, if everyone acted on certain maxims that were not optimific, because some of these acts would give this person extra benefits, in ways that imposed greater burdens on others. But this claim, I believe, is false. As before, some of the fortunate people would care strongly about the well-being of others, and if things went in the ways that would be impartially best, that would be best for some of these people.

Similar claims apply to any other plausible or widely accepted view about reasons and rationality. On all such views, there would be some people who could rationally will it to be true that everyone acts on the optimific maxims. Kant's original Law of Nature Formula, we can therefore claim, permits some people to be Maxim Consequentialists.

It is an objection to Kant's formula that it permits only *some* people to be Maxim Consequentialists, since such moral claims ought to apply to everyone. We can call this *the Relativism Objection*. To answer this objection, we can revise Kant's formula so that it appeals, not to what the agent could rationally will, but to what everyone could rationally will. This revised formula has implications that apply to everyone.

We have other strong reasons, I have argued, to revise Kant's formulas in this and certain other ways. These revisions lead us to the Kantian Contractualist Formula. So we can now ask what this formula implies.

57 The Kantian Argument

Of the principles that everyone might accept, some might be

> *UA-optimific* in the sense that these are the principles whose universal acceptance would make things go best.

According to the *universal acceptance* version of Rule Consequentialism, or

> *UARC*: Everyone ought to follow these optimific principles.

When we consider some kinds of case, there might be two or more optimific principles that were significantly different. Rule Consequentialists would then have to choose between these principles in some other way. This question is best considered later. So we can here suppose that there is only one set of UA-optimific principles.

Kantians could argue:

> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.
>
> (B) Everyone could rationally choose whatever they would have sufficient reasons to choose.
>
> (C) There are some UA-optimific principles.
>
> (D) These are the principles that everyone would have the strongest impartial reasons to choose.
>
> (E) No one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons.

Therefore

> (F) Everyone would have sufficient reasons to choose these optimific principles.
>
> (G) There are no other significantly non-optimific principles that everyone would have sufficient reasons to choose.

Therefore

> (H) It is only these optimific principles that everyone would have sufficient reasons to choose, and could therefore rationally choose.

Therefore

> Everyone ought to follow these principles.

This argument is valid. (A) is the Kantian Contractualist Formula. So if this argument's other premises are true, this formula requires everyone to follow these optimific principles. We can call this *the Kantian Argument for Rule Consequentialism*.

When we apply the Kantian Formula, we ask which principles each person could rationally choose, if this person supposed that he or she had the power to choose which principles would be accepted by everyone, both now and throughout the future. This formula appeals to the principles that, in these many imagined cases, everyone could rationally choose. We should assume that, in making these choices, everyone would know all of the relevant facts. On that assumption, as premise (B) claims, everyone could rationally choose what they would have sufficient reasons to choose.

We are supposing that, as (C) claims, there is some set of principles that are UA-optimific. Of all the principles that everyone might accept, these are the principles whose universal acceptance would make things go best in the impartial-reason-implying sense. If everyone accepted these principles, things would go in the ways in which everyone would have the strongest impartial reasons to want things to go. That is true by definition. So, as premise (D) claims, these are the principles whose universal acceptance everyone would have the strongest impartial reasons to choose.

According to premise (E), no one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons. This premise needs to be defended. If we were choosing principles from an impartial point of view, it is the optimific principles that everyone would have most reason to choose. But in the thought-experiments to which this Kantian Formula appeals, we would *not* be choosing principles from an impartial point of view. Our choices would affect our own lives, and the lives of those other people to whom we have close ties, such as our close relatives and those we love. So we might have strong personal and partial reasons *not* to choose the optimific principles.

To decide whether everyone could rationally choose these principles, we must know what the alternatives would be. It will be enough here to consider other principles that would be *significantly* non-optimific, in the sense that their universal acceptance would make the future history of the world go, in certain ways, *much* worse. We need not compare the optimific principles with any principles that are only *slightly* non-optimific, since their acceptance would make things go in ways that would be only slightly worse. As before, we should first try to get the main outlines right. Details can wait.

58 Self-Interested Reasons

In asking whether premise (E) is true, we should consider the strongest reasons that anyone might have not to choose that everyone accepts the optimific principles. Of our reasons not to choose these principles, some might be provided by facts about our own well-being. If everyone accepted the optimific principles, that would be very bad for certain people. These people would have strong self-interested reasons not to choose these principles.

I might be such a person. Suppose again that, in

> *Lifeboat*, I am stranded on one rock, and five people are stranded on another. Before the rising tide covers both rocks, you could use a lifeboat to save either me or the five. I and the five are all strangers to you and to each other, and we are in other ways relevantly similar. We are all young, and we would all lose, in dying, many years of happy life.

Any optimific principle would require you to save the five, since it would be worse if more people died. According to one such principle, which we can call

> *the Numbers Principle*: When we could save either of two groups of people, who are all strangers to us and are in other ways relevantly similar, we ought to save the group that contains more people.

Suppose next that my rock is nearer to you. According to

> *the Nearness Principle*: In such cases, we ought to save the group that is nearer to us.

If everyone accepted the Numbers Principle rather than the Nearness Principle, there would be many other cases in which some people would act on this principle, so many more people's lives would be saved. This fact would give me strong impartial reasons to choose that everyone accepts the Numbers Principle. But I would know that, if I made this choice, you would act on this principle by saving the five, and I would die, thereby losing many years of happy life. This fact would give me strong self-interested reasons to choose the Nearness Principle, since you would then save my life. According to premise (E), these self-interested reasons would not be decisively stronger than, or outweigh, my impartial reasons to choose the Numbers Principle. Is that true?

On Subjectivist theories about reasons, the answer depends on my desires or aims. If I cared enough about the well-being of other people, I could rationally choose that everyone accepts the Numbers Principle. But if we are Subjectivists, we must reject the Kantian Formula. In most cases, there would be no principles that *everyone* would have sufficient desire-based or aim-based reasons to choose. As I have argued, however, we ought to reject Subjectivism, and accept some Objectivist view, which appeals to value-based object-given reasons.

According to one such view,

> *Rational Egoism*: We always have most reason to do whatever would be best for ourselves.

On this view, premise (E) is false. I could not rationally choose that everyone accepts the Numbers Principle, since that choice would be worse for me. But we ought, I believe, to reject this view.

According to a view at the opposite extreme,

> *Rational Impartialism*: We always have most reason to do whatever would be impartially best.

On this view, we would be rationally required to sacrifice our life if we could thereby save several strangers. If that were true, cases like *Lifeboat* would provide no objection to premise (E). I would be rationally required to choose that everyone accepts some optimific principle, such as the Numbers Principle. But we ought also, I believe, to reject this view.

According to

> *wide value-based objective views*: When one of two possible acts would make things go in some way that would be impartially better, but the other act would make things go better either for ourselves or for other people to whom we have close ties, we often have sufficient reasons to act in either way.

On such views, we are often rationally permitted but not rationally required to give significantly greater weight, or strong priority, both to our own well-being and to the well-being of those to whom we have close ties, such as our close relatives and those we love. We ought, I believe, to accept some view of this kind.

On the views that seem to me most plausible, if we could save either our own life or the lives of several strangers, we would have sufficient reasons to act in either way. In *Lifeboat*, I could rationally choose that you save me; but I could also rationally choose instead that you save the five. So I could rationally choose that everyone accepts the Numbers Principle.

According to some more egoistic objective views, we are rationally required to give strong priority to our own well-being. I would not have sufficient reasons to give up my life unless I would thereby save as many as a hundred or a thousand other people. But in the thought-experiment to which the Kantian Formula appeals, I would have the power to choose which principles everyone would accept, both now and in all future centuries. The principles I chose would be accepted by many billions of people. If I chose that everyone accepts the Numbers Principle rather than the Nearness Principle, my choice would affect how people would later act in very many other cases of this kind. Though I would die, my choice would indirectly save at least a million other people. Millions of

people now die each year whose lives could have been easily saved. So even on these more egoistic views, I would have sufficient reasons to give up my life to save these very many other people.

This case is only one example. But if, as I believe, I could rationally choose this optimific principle even at the cost of my own life, similar claims apply to all of the many cases in which, because the stakes are lower, no one's choice of an optimific principle would involve so great a sacrifice of their own well-being.

Suppose next that my belief is mistaken. We ought, I have claimed, to reject Rational Egoism. But there is another, more plausible view that is relevant here. On this view, we could often rationally choose to bear some significant burden when we could thereby save many other people from similar burdens. That is not true, however, when this burden would be as great as dying young, and thereby losing many years of happy life. I could not rationally choose the Numbers Principle, because I could not rationally choose to give up my life, however many other people's lives my choice would save. We can call this view *High Stakes Egoism*.

If this view were true, *Lifeboat* would provide an objection, not only to premise (E) of the Kantian Argument for Rule Consequentialism, but also to the Kantian Contractualist Formula. Just as I could not rationally choose any principle that required you to save the five rather than me, the five could not rationally choose any principle that required you to save me rather than them. In this and other such cases, there would be no principle that everyone could rationally choose, so there would be no principle that the Kantian Formula would require us to follow. If we could save either one stranger or a million others, this formula would permit us to act in either way. That is an unacceptable conclusion.

High Stakes Egoism is, I believe, false. But it is worth describing how, if this view were true, we could respond to this objection to the Kantian Formula.

Contractualists appeal to the principles that it would be rational for everyone to choose, if we were choosing in some way that would make our choices sufficiently impartial. Rawls suggests that, to achieve such impartiality, we should appeal to the principles that it would be rational for everyone to choose from behind some *veil of ignorance*, which prevented us from knowing particular facts about ourselves

or our situation. I have claimed that, when we apply the Kantian Contractualist Formula, we have no need for such a veil of ignorance. There would always be some relevant principle that, even with full knowledge, everyone could rationally choose.

We are now supposing that, in one kind of case, my claim is mistaken. In these cases, we could save the lives of either of two groups of strangers, one of which contains more people. According to High Stakes Egoism, when the people in these groups were choosing principles that apply to such cases, they would be rationally required to give absolute priority to the saving of their own lives. The Kantian Formula would here fail because these people's choices would be wholly self-interested. To avoid this objection, we could revise this formula. When we apply the Kantian Formula to such cases, we might appeal to the principles that these people could rationally choose from an impartial point of view. Or we might partly follow Rawls by adding a *local* veil of ignorance, so that these people did not know whether they were in the smaller or the larger group. On both these versions of the Kantian Formula, these people could all rationally choose some optimific principle that would require us to save the group that contained more people.

The Kantian Formula might be more sweepingly revised, by appealing to principles that would *all* be chosen either from an impartial point of view, or from behind a *global* veil of ignorance. But that would make this formula less appealing in ways that I describe in Section 52. And there would be no need for such a revision. High Stakes Egoism applies only to cases in which, if we chose some optimific principle, this choice would impose on us some very great burden, such as dying young or having to endure prolonged agony. We could rationally choose to accept some lesser injury, such as becoming deaf, or losing a leg, when our choice would indirectly save many other people from such injuries. So we could still claim that, in nearly all cases in which people's interests conflict, there would be some principle that, even with full knowledge and from their actual partial point of view, all of these people could rationally choose.

If we ought to reject High Stakes Egoism, as I believe, the Kantian Formula does not need to be even partly revised in such a way.

59 Altruistic and Deontic Reasons

Of our reasons not to choose the optimific principles, others might be provided by facts about certain other people's well-being. Suppose that, in

> *Second Lifeboat*, you could save either your child or five strangers.

We may believe that, even if you could rationally give up your *own* life to save five strangers, you could not rationally give up your *child's* life to save these strangers, nor could you rationally choose that we all accept some optimific principle that would require this act. This case may then seem to provide an objection to premise (E).

The optimific principles would *not*, however, require you to save these five strangers rather than your child. Suppose that we all accepted and acted on some principle that required us to give no priority to saving our own children from death or lesser harms. In such a world, things would go in some ways better, since more children's lives would be saved and fewer children would be harmed. But these good effects would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would give no such priority to saving our own children from harm, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and other similar facts, the optimific principles would in many cases permit us, and in many others require us, to give strong priority to our own children's well-being.

This objection could be transferred, however, to a different kind of case. Suppose that, in

> *Third Lifeboat*, it is I who could save either your child or five other children. These six children are all strangers to me.

Any optimific principle would require *me* to save the other five children. And we might claim that

> (I) you could not rationally choose that everyone accepts such an optimific principle, since you would have decisive reasons to choose instead that I accept some principle that would require me to save your child.

You would have such decisive reasons, we might claim, because you would have a duty to make the choice that would save your child's life.

There are other ways in which, by appealing to our moral beliefs, we might argue that we could not rationally choose that everyone accepts certain optimific principles. We may believe that, if everyone accepted these principles, that would sometimes lead us or others to act wrongly. The wrongness of such acts, we might claim, would give us decisive reasons not to choose that everyone accepts these principles.

As I have often said, however, when we apply the Kantian Formula or any other Contractualist formula, we cannot appeal to our beliefs about which acts are wrong. If we claim that

> some act is wrong because we could not all rationally choose any principle that permits such acts,

it would be pointless also to claim that

> we could not all rationally choose any such principle because such acts are wrong.

It would be similarly pointless to claim both that

> everyone ought to follow certain principles because these are the only principles that everyone could rationally choose,

and that

> these are the only principles that everyone could rationally choose because these are the principles that everyone ought to follow.

If we combined these claims, the Kantian Formula would achieve nothing. So when we apply this formula, we must ignore our beliefs about which acts are wrong. We can appeal to these beliefs only at a

later stage, after we have worked out what this formula implies, and we are asking whether, given these implications, we ought to accept this formula.

Since we cannot appeal to our beliefs about your duties to your child, could we defend (I) in some other way? We could most plausibly appeal, I believe, to your love for your child. Rather than trying to ignore your duties to your child, it will be simpler to change our example. Suppose that, in

> *Fourth Lifeboat*, I could save either someone whom you love or five other people. These six people are all strangers to me.

Any optimific principle would require me to save the other five people. It might now be claimed that

> (J) you could not rationally choose that everyone accepts some optimific principle, since you would have decisive reasons to choose that I accept some other principle which required me to save the person whom you love.

Though this claim is plausible, it is not, I believe, true.

It may seem absurd to deny that you would have decisive reasons to choose this other principle. Could Romeo or Isolde have rationally chosen to let Juliet or Tristan die? While discussing a similar example, Williams writes:

> deep attachments to other persons . . . cannot embody the impartial view, and . . . also run the risk of offending against it . . . yet unless such things exist, there will not be enough substance or convictions in a man's life to compel his allegiance to life itself. Life has to have substance if anything is to have sense, including adherence to the impartial system; but if it has substance, then it cannot grant supreme importance to the impartial system . . .

I am not appealing, however, to the kind of impartial system that Williams here movingly rejects. As I have just said, the optimific

principles would often either permit or require us to give strong priority to the well-being of those to whom we have close ties. And in claiming that we could rationally choose that everyone accepts these principles, I am not assuming that we are rationally required to give equal weight to everyone's well-being. I assume only that, though we are rationally permitted to give strong priority to the well-being of ourselves and certain other people, we are also rationally permitted to give great weight to the well-being of strangers.

As my claims about *Lifeboat* imply, the person whom you love could rationally choose that everyone accepts some optimific principle. Though this person would then die, this choice would indirectly save very many other people's lives. This fact would give this person sufficient reasons to make this choice.

When someone whom we love could rationally choose to bear some burden for the sake of benefits to others, this fact does not imply that *we* could rationally choose that this person bears this burden. We might be rationally required to give to the well-being of those we love much more weight than we are rationally required to give to our own well-being. We might not have sufficient reasons to save five, or fifty, or even five hundred strangers rather than saving someone whom we love. But in *Fourth Lifeboat* you would know that, if you chose that everyone accepts some optimific principle, your choice would indirectly save the lives of a much greater number of other people. You would have sufficient reasons, I believe, to make the choice that would save these many other people. It is, I agree, absurd to imagine Romeo or Isolde choosing to let Juliet or Tristan die. If you were Romeo or Isolde, you would not *in fact* make the choice that would save these many other people. But we often know that people won't in fact do what they have sufficient reasons to do. Since you would have sufficient reasons to choose some optimific principle, *Fourth Lifeboat* does not, I believe, provide an objection to premise (E), or to the Kantian Formula.

Suppose next that my belief is mistaken. It might be claimed that, when the stakes are as high as this, we ought rationally to give absolute priority to the well-being of those we love. If that were true, there would be no principle applying to such cases that everyone could

rationally choose, so there would be no principle that, according to the Kantian Formula, everyone ought to follow. This formula would not require me to save even a million strangers rather than the person whom you love. That is another unacceptable conclusion. This objection is like the one that appeals to High Stakes Egoism. As before, the Kantian Formula could be revised by adding some local veil of ignorance. But this revision is not, I believe needed.

60 The Wrong-Making Features Objection

On some value-based objective theories, there are some things that are worth doing, and some other aims that are worth achieving, in ways that do not depend, or depend only, on their contributions to anyone's well-being. Scanlon's examples are 'friendship, other valuable personal relations, and the achievement of various forms of excellence, such as in art or science'. These we can call *perfectionist* aims.

On such views, it would be in itself good in the impartial-reason-implying sense if we and others had these valuable personal relations, and achieved these other forms of excellence. The optimific principles might require us to try to achieve some perfectionist aims, and to help other people to do the same. Since these are views about how it would be best for things to go, these claims could not give us reasons to reject the optimific principles.

On some views, however, we might also have some *personal* and *partial* perfectionist reasons. These are not self-interested reasons, since to achieve some perfectionist aim we may have to sacrifice much of our well-being. But these reasons might conflict with our reasons to make things go impartially better in such perfectionist ways. Suppose that I could save either the only copy of my great nearly finished novel or the only copies of five similarly great novels by other writers. I might have personal perfectionist reasons not to choose any optimific principle that would require me to save these other people's novels rather than saving mine. But these reasons would not, I believe, outweigh my impartial reasons to choose this principle. I could rationally give up my novel to save these five other similarly great novels. If my belief were mistaken,

we could again revise the Kantian Formula. But that would make little difference, since such cases would be rare.

There is another, more important possibility. Suppose that some optimific principle requires certain acts that we believe to be wrong. When we apply the Kantian Formula, we cannot appeal either to our belief that certain acts are wrong, or to the *deontic* reasons that the wrongness of these acts might provide. But we can appeal to the features of these acts that, in our opinion, *make* them wrong. And we might claim that

> (K) these wrong-making features give us decisive *non-deontic* reasons not to act in these ways, and not to choose that everyone accepts the optimific principle that requires such acts.

If there were certain acts of which (K) was true, that would provide an objection to premise (E) of the Kantian Argument for Rule Consequentialism, since there would be an optimific principle that we would not have sufficient reasons to choose. We can call this *the Wrong-Making Features Objection*.

This objection rightly assumes that, of the features that can make acts wrong, some would also give us decisive non-deontic reasons. If certain acts would cause pointless suffering, for example, this fact would give us decisive reasons not to act in these ways. These reasons would not be deontic, since they would not be provided by the fact that these acts would be wrong. The wrongness of these acts would at most give us further reasons not to act in these ways. But (K) could not be truly applied to these acts, since the optimific principles would not require us to cause pointless suffering.

(K) seems most likely to be true when applied to acts that would have good effects, but would also, we believe, violate some principle about the wrongness of treating people in some way. Return to

> *Bridge*, in which you cannot save the five except by causing me to fall in front of the runaway train, thereby killing me.

Suppose we believe that this act would be wrong, and that its wrong-making feature is the fact that

> (L) you would be killing me as a means of saving these other people.

To state one version of objection (K), we might claim both that

> (M) the optimific principles would require us, in cases like *Bridge*, to kill one person as a means of saving several others, since we would thereby make things go better,

and that

> (N) the wrong-making feature of such acts would give us a decisive non-deontic reason not to act in this way, and not to choose any optimific principle that would require such acts.

(M) is not obviously true. For various reasons that I mention above and below, the optimific principles would often permit or even require us *not* to do what would make things go best. But we can here suppose that (M) is true. It will be enough to ask whether claims like (M) and (N) could *both* be true.

For the optimific principles to require certain acts, it must be true that

> (O) when we consider these acts from an impartial point of view, we would have most reason to want everyone to act in these ways.

If we did not have such impartial reasons, it would not be better in the impartial-reason-implying sense if everyone acted in these ways, so the optimific principles would not require such acts. Our point of view is impartial when we are considering cases that involve people who are all strangers to us. That is true of nearly all actual cases, since nearly everyone is a stranger to us. So we can also claim that if

> (P) the optimific principles require certain acts,

it must be true that

> (Q) we would have most reason to want nearly everyone to act in these ways.

On the objection we are now considering,

> (R) some of these acts have certain features that would give everyone decisive non-deontic reasons not to act in these ways.

At least in most cases, I believe, (P), (Q), and (R) could not all be true. When applied to *Bridge*, for example, these claims would imply that

> (S) you would have a decisive non-deontic reason not to save the five by killing me,

but that

> (T) you would also have most reason to want or hope that some stranger would arrive and act instead of you, saving the five by killing me.

On this view, though everyone would have decisive non-deontic reasons not to kill someone as a means of saving more lives, what everyone would have most reason to want, from an impartial point of view, is that everyone who can act in this way *does* kill someone as a means of saving more lives. These two kinds of reason could not, I believe, be so directly opposed. We could not have such impartial reasons to want everyone to do what everyone had such decisive non-deontic reasons *not* to do. So (S) and (T) could not both be true.

Similar claims apply to other cases. Of the features that make certain acts wrong, most give us non-deontic reasons not to act in these ways. At least in most cases, these features also give us reasons to want other people not to act in these ways. That is most obviously true of those wrong acts that harm other people, since we all have impartial reasons to want other people not to be harmed. But similar claims would apply to acts that had other wrong-making features. Suppose, for example, that it would be wrong to deceive or coerce other people as a means of producing certain benefits. The wrong-making features of these acts might give everyone decisive non-deontic reasons not to act in these ways. If that were true, could it also be true that, from an impartial point of view, we would have most reason to want everyone to act in these

ways? Our answer should, I believe, be No. If the nature of deception and coercion gave everyone decisive non-deontic reasons, in such cases, *not* to deceive and coerce others, we could not also have such impartial reasons to *want* everyone, in such cases, to deceive or coerce others. That would be a strangely schizophrenic or internally conflicting view. And if we did not have such impartial reasons, the optimific principles would not require such acts. I defend these claims further below.

There may, however, be one kind of exception. Suppose that, in

> *Lesser Evil*, you know that, unless you save the five by killing me, *Grey* and *Green* will save the five by each killing two other people.

Of those who believe it to be wrong to kill someone as a means of saving other people, most would believe that such an act would be wrong even if, as in *Lesser Evil*, this act is the only way to prevent more acts of the same kind. Even if this act would be wrong, however, we would have impartial reasons to want you to act in this way. Though it would be bad if you killed me as a means, it would clearly be even worse if Grey and Green both acted wrongly in this way, by each killing two people as a means. So if we learnt that you had acted wrongly in this way, thereby preventing the wrong acts of both Grey and Green, we ought to regard this fact as, in a sober way, good news. Similar claims apply if we set aside our beliefs about which acts are wrong, as we must do when applying the Kantian Formula. If everyone had such decisive non-deontic reasons *not* to act in some way, we could not, I have claimed, have impartial reasons to *want* everyone to act in this way. That would be a schizophrenic view. But we might have impartial reasons to want *no one* to act in this way *except* when such an act is the only way to prevent more such acts. That would not be a schizophrenic view.

According to the objection that we are now discussing

> (U) The optimific principles require us to act in certain ways, though these acts have wrong-making features that give everyone decisive non-deontic reasons not to act in these ways, and not to choose that everyone accepts these principles.

As I have argued, we can reply that

> (V) if these acts had such features, the optimific principles would *not* require us to act in these ways, except perhaps when such an act would be the only way to prevent more such acts.

If (V) is true, as I believe, this objection would at most apply to only a few cases, such as *Lesser Evil*. I shall now argue that, even in these cases, this objection would fail. If you expect that you would agree, you might skip the next section.

61 Decisive Non-Deontic Reasons

If you saved the five, in *Bridge* or *Lesser Evil*, you would be doing that by killing me. We can next ask whether, as this objection claims, this feature of your act *would* give you a decisive non-deontic reason not to act in this way. We can first reconsider *Tunnel*: the case in which, if you redirected a runaway train,

> (W) you would save the five, but in a way that also killed me.

This fact, we can plausibly believe, would give you a strong non-deontic reason not to act in this way. It would be awful to do what you knew would kill an innocent person. This may be why many people believe that you would merely be morally permitted, rather than morally required, to save the five by redirecting this train. But, as these people would, I believe, agree, the awfulness of killing someone would not give you a *decisive* non-deontic reason not to act in this way. If you would be morally permitted to redirect this train, though you would thereby kill me, the fact that you would be saving several people's lives would give you a sufficient reason to act in this way.

Similar claims apply to *Bridge*, in which if you caused me to fall onto the track

> (L) you would be killing me as a means of saving the five.

It would again be awful to save the five by killing an innocent person. This feature of this act might give you a strong non-deontic reason

not to act in this way. As in *Tunnel*, however, this non-deontic reason could not decisively outweigh your reason to do what would save several people's lives. If *Bridge* is significantly different from *Tunnel*, as many people would believe, this difference could not, I believe, be that, since you would be killing me as a means, you would have a decisive *non*-deontic reason not to act in this way. This feature of this act might give you a decisive reason not to act in this way. But it could do that, I believe, only *by making this act wrong*. This decisive reason would have to be *deontic*. If that is true, the objection we are now considering fails. You would not have a decisive *non*-deontic reason not to act in this way.

Similar remarks apply to other kinds of case. I suggest that

> (X) if the optimific principles require certain acts that we believe to be wrong, the features or facts that, in our opinion, make these acts wrong would not give us decisive *non*-deontic reasons not to act in these ways. What might be true is only that, by making these acts wrong, these facts would give us decisive deontic reasons not to act in these ways.

The optimific principles would require several kinds of act that many people believe to be wrong. These principles might, for example, require some of us to use artificial contraceptives, or to perform or have an abortion, or to help someone to die in a swifter, better way, or to steal from certain rich people and give what we steal to the poor. If we had decisive reasons not to act in these ways, these reasons, I suggest, would have to be provided by the wrongness of these acts.

We should expect (X) to be true. If the optimific principles require some kind of act, we must have strong impartial reasons to want everyone to act in this way. If we did not have such reasons, it would not be better if everyone acted in these ways, so the optimific principles would not require such acts. Since we would have strong impartial reasons to want everyone to act in this way, we should expect that these reasons could not be decisively outweighed except by the fact that such acts would be wrong. I defend (X) further in Appendix C.

Though I am strongly inclined to believe that (X) is true, it is again worth supposing that I am mistaken. Suppose that the optimific principles

require certain acts that we believe to be wrong, and that the features that, in our opinion, make these acts wrong *would* give us decisive *non*-deontic reasons not to act in these ways. These beliefs would not, by themselves, provide an objection to premise (E). This objection must claim that

> (Y) these wrong-making features would also give us decisive non-deontic reasons not to choose that everyone accepts the optimific principle that requires such acts.

Only (Y) would count against (E), by implying that there is some optimific principle that we would not have sufficient reasons to choose.

(Y) is a claim, not about our reasons for acting in certain ways, but about our reasons for choosing that everyone accepts some principle. These are quite different questions. Consider, for example, some kind of act that would be bad for us, but would give some greater benefit to others. Even if we had strong reasons not to act in this way, we might have decisive reasons both to want everyone to act in this way, and to choose that everyone accepts some principle that requires such acts. If everyone acted in this way, for example, that might be better for everyone, including us.

(Y) seems most likely to be true when applied to acts that violate some deontological constraint. Our main example is *Bridge*. We are supposing both that, in this case, the optimific principles would require you to save the five, and that this act would be made to be wrong by the fact that you would be killing me as a means. According to (Y), this fact would give you a decisive non-deontic reason not to choose that everyone accepts any such optimific principle. We should ask what this reason might be.

Since this reason must be non-deontic, it could not be provided by the wrongness of such acts. We might appeal again to the awfulness of saving several people's lives by killing an innocent person. The awfulness of such an act, we can plausibly believe, would give you a strong non-deontic reason to want not to be morally required to act in this way. But in a case like *Tunnel*, as we have seen, this reason would not be decisive, since you would have sufficient reasons to save the five in a way that would also kill me. And if the optimific principles required

you, in *Bridge*, to save the five by killing me, this would have to be because the relevant facts gave you impartial reasons to want everyone, in such cases, to act in such ways. These facts would also give you reasons to want everyone to accept some principle that requires them to act in this way. These impartial reasons could not, I believe, be *decisively* outweighed by your personal *non*-deontic reason to want *yourself not* to be required to act in this way.

In defending this belief, I shall make some wider claims, which apply to all cases. If the optimific principles require us to act in some way, the relevant facts must give us impartial reasons to want everyone, in relevantly similar cases, to act in this way. Only then would it be better if everyone acted in this way. Since we would be considering nearly all these cases from an impartial point of view, we would have most reason to want nearly everyone to act in this way. If we choose that everyone accepts the principle that requires such acts, our choice would indirectly bring it about that most people *would* do what we had most reason to want nearly everyone to do. These facts would give us strong impartial reasons to choose that everyone accepts this principle. According to premise (E), these reasons would not be decisively outweighed by any relevant conflicting reason. We are now asking whether, as (Y) claims, there are some cases in which (E) is false.

It will help to remember here the other kinds of case that raise the strongest objections to (E). If we choose that everyone accepts some optimific principle, this choice might be very bad either for ourselves or for certain people to whom we have close ties, such as those we love. In *Lifeboat*, for example, if I chose that everyone accepts the Numbers Principle, you would save the five rather than me, and I would lose many years of happy life. This fact would give me a very strong personal reason not to choose the Numbers Principle. But this reason would not, I believe, be decisive. By choosing that everyone accepts this optimific principle, I would indirectly save very many other people's lives, and this fact would give me sufficient reasons to make this choice.

We are now considering a different kind of reason. In the cases to which (Y) might apply, the relevant facts would give us strong impartial reasons both to want everyone to act in some way, and to choose that everyone accepts some optimific principle that requires such acts. But

these impartial reasons, (Y) claims, would be decisively outweighed by some conflicting non-deontic reason. Any such reason would have to be much stronger than the personal reasons I have just mentioned, such as our reasons to want not to die young, losing many years of happy life. Only if this reason was much stronger could it decisively outweigh these conflicting impartial reasons. There is, I believe, only one third kind of reason that might be clearly stronger than, and decisively outweigh, both such strong personal reasons and such strong impartial reasons. If we would have some decisive reason not to make some choice, despite the fact that this choice would either (1) be much better for ourselves or those we love, or (2) would make things go impartially much better, this reason would have to be provided by the fact that this choice would be morally wrong. We could not have decisive *non*-deontic reasons not to make this choice. If that is so, as I believe, (Y) could not be true, so this objection to (E) fails.

62 What Everyone Could Rationally Will

According to premise (E), no one's impartial reasons to choose the optimific principle would be decisively outweighed by any relevant conflicting reasons. In defending (E), I have appealed to several claims that I believe to be true, and then argued that, even if I am mistaken, (E) would still be true, or could be made true by some acceptable revision of the Kantian Formula. Premise (E) is in this way *robust*.

It is worth supposing that I have made yet another mistake. Suppose that, in some cases, (Y) *is* true, because we would have a decisive non-deontic reason not to choose that everyone accepts some optimific principle. Suppose also that this objection could not be met by any similar revision of the Kantian Formula. In such cases, (E) would be false. The Kantian Argument could not show that the Kantian Formula always requires us to follow the optimific principles. We would have to revise this argument's conclusion.

This argument would then be in a different way robust, since this revision would be slight. For the reasons given above, if there were cases in which (Y) was true, such cases would be rare. (Y) might be true

only in cases like *Lesser Evil*, in which some optimific principle required some act as the only way to prevent more such acts. Since such cases would be rare, the Kantian Argument would show that, in nearly all actual cases, the Kantian Formula requires us to follow the optimific principles. Kantian Contractualism would then be, in its implications, very close to Rule Consequentialism. There might be less disagreement between these theories than there is between some different versions of Rule Consequentialism.

Remember next that, in *supposing* that (Y) is sometimes true, I am supposing that several of my earlier claims are mistaken. (Y), I believe, is never true. If that is so, this argument's conclusion does not need to be revised.

There is, I believe, no other strong objection to (E). If that is so, we ought to accept premises (B) to (E). Everyone would have strong impartial reasons to choose the optimific principles, and these reasons would not be decisively outweighed by any relevant conflicting reasons.

Since we ought to accept these claims, we ought to accept this argument's first conclusion. As (F) claims, everyone would have sufficient reasons to choose that everyone accepts the optimific principles.

According to this argument's remaining premise:

> (G) There are no other, significantly non-optimific principles whose universal acceptance everyone would have sufficient reasons to choose.

Compared with (E), this premise is much easier to defend. If everyone accepted any such other principle, things would go in ways that would be impartially much worse. That is what is meant by the claim that these other principles are significantly non-optimific. These facts would give everyone strong impartial reasons not to choose that everyone accepts any such principle. Since most people would have no conflicting personal reasons, most people could not rationally make this choice. And in nearly all these cases, if everyone accepted any such non-optimific principle, things would also go much worse for some unfortunate

people. It is even clearer that *these* people could not rationally choose that everyone accepts this principle, since these people would have both strong impartial reasons and strong personal reasons not to make this choice. In *Earthquake*, for example, White could not rationally choose that we all accept some non-optimific principle that required me to save Grey's leg rather than White's life. And in *Lifeboat*, none of the five could rationally choose that we all accept some non-optimific principle that required you to save me rather than saving all of the five. So, as (G) claims, there are no significantly non-optimific principles that everyone would have sufficient reasons to choose.

(B), (F), and (G) together imply

> (H) It is only the optimific principles whose universal acceptance everyone would have sufficient reasons to choose, and could therefore rationally choose.

When combined with (H), the Kantian Formula implies that everyone ought to follow these principles. I defend these claims further in a note.

We can now restate this argument more briefly. Kantians could claim:

> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally choose, or will.
>
> (C) There are some principles whose universal acceptance would make things go best.
>
> (F) Everyone could rationally will that everyone accepts these principles.
>
> (H) These are the only principles whose universal acceptance everyone could rationally will.

Therefore

> UARC: These are the principles that everyone ought to follow.

(A) is the Kantian Contractualist Formula, and UARC is one version of Rule Consequentialism. We are assuming (C). I have, I believe,

successfully defended (F) and (H). So this Kantian Formula requires everyone to follow these Rule Consequentialist principles.

This argument, we may suspect, must have at least one Consequentialist premise. If that were true, this argument would be uninteresting. We would expect Consequentialist premises to imply Consequentialist conclusions. And such an argument would not give Non-Consequentialists any reason to change their view.

This argument's premises are not, however, Consequentialist. The argument assumes that outcomes can be better or worse in the impartial-reason-implying sense. But Non-Consequentialists can accept that assumption. Many Non-Consequentialists believe, for example, that it would be worse if more people suffer, or die young. These people reject Consequentialism, not because they deny that outcomes can be in this sense better or worse, but because they believe that the rightness of acts does not depend only on facts about how it would be best for things to go. This argument also assumes that there are some principles whose universal acceptance would make things go best. But this assumption is not Consequentialist. We may believe that there are such optimific principles, but also believe that we ought to reject some of these principles, because they require or permit some acts that are wrong.

Since this argument does not have any premise that assumes the truth of Consequentialism, it is worth explaining how this argument validly implies its Consequentialist conclusion.

Consequentialists appeal to claims about what would be best in the impartial-reason-implying sense. These are claims about what, from an impartial point of view, everyone would have most reason to want, or choose. The strongest objections to Consequentialism are provided by some of our intuitive beliefs about which acts are wrong.

Contractualists appeal to the principles that it would be rational for everyone to choose, if we were all choosing in some way that would make our choices sufficiently impartial. Some Contractualists claim that, to achieve such impartiality, it is enough to appeal to the principles that it would be rational for everyone to choose, if everyone needed to reach

agreement on these principles. Other Contractualists, such as Rawls, add a veil of ignorance. Kantian Contractualists achieve impartiality by appealing to what everyone could rationally choose, if each person supposed that he or she had the power to choose which principles everyone would accept. Impartiality is here achieved, without any need to reach agreement or any veil of ignorance, by the requirement of unanimity. In arguing that there are principles that everyone could rationally choose, I have appealed to another feature of Contractualism. When we apply any Contractualist formula, we cannot appeal to our intuitive beliefs about which acts are wrong.

We can now explain how, without having any Consequentialist premise, this argument validly implies its Consequentialist conclusion. As I have just said:

> Consequentialism appeals to claims about what it would be rational for everyone to choose from an impartial point of view. The strongest objections to Consequentialism are provided by some of our intuitive beliefs about which acts are wrong.
>
> Contractualism appeals to claims about what it would be rational for everyone to choose, in some way that would make these choices impartial. In Contractualist moral reasoning, we cannot appeal to our intuitive beliefs about which acts are wrong.

Since both kinds of theory appeal to what it would be rational for everyone impartially to choose, and Contractualists tell us to ignore our Non-Consequentialist moral intuitions, we should expect that valid arguments with some Contractualist premise could have some Consequentialist conclusion.

We can draw another conclusion. There are, I have claimed, some decisive objections to Kant's Formula of Universal Law. To avoid these objections, Kant's Formula must be revised. In its best revised form, this formula requires us to follow the principles whose universal acceptance everyone could rationally will, or choose. There are, I have argued,

no significantly non-optimific principles that everyone could rationally choose. So this formula cannot succeed unless it is true that, as I have also argued, everyone could rationally choose the optimific principles. Kant's Formula of Universal Law cannot succeed unless, in this revised form, this formula implies Rule Consequentialism.

17

Conclusions

63 Kantian Consequentialism

Return next to Act Consequentialism, or

> AC: Everyone ought always to do, or try to do, whatever would make things go best.

Is this principle UA-optimific, by being the principle whose universal acceptance would make things go best?

As Sidgwick argued, the answer is No. If everyone always tried to do whatever would make things go best, these attempts would often fail. When predicting the effects of possible acts, people would often make mistakes, or deceive themselves in self-benefiting ways. It would be easy, for example, to believe that we were justified in stealing or lying, because we falsely believed that the benefits to us would outweigh the burdens that our acts would impose on others. If we were all Act Consequentialists, that would also undermine or weaken some valuable practices or institutions, such as the practice of trust-requiring promises. If everyone had the motives of an Act Consequentialist, that would be bad in other ways. For it to be true that everyone nearly always tried to make things go best, most of us would have to lose too many of the strong loves, loyalties, personal aims, and other motives in which much of our happiness consists, and that also make our lives in other ways worth living. For these and other such reasons, we can claim that

> (A) if everyone accepted AC, things would go worse than they would go if everyone accepted certain other principles.

These other, *UA-optimific* principles would partly overlap with the principles of common sense morality. These principles would often require us, for example, not to steal, lie, or break our promises, even when such acts would predictably make things go best. These principles would permit us to give some kinds of strong priority to our own well-being. And they would often permit us, and often require us, to give some kinds of strong priority to the well-being of certain other people, such as our close relatives and friends, and those to whom we may be related in various other ways, such as our pupils, patients, clients, colleagues, customers, neighbours, and those whom we represent. Since AC is not the principle whose universal acceptance would make things go best, the Kantian Formula does not require us to be Act Consequentialists.

We have been discussing the *universal acceptance* version of Rule Consequentialism, or UARC. According to a different version of this theory, which we can call

> *UFRC*: Everyone ought to follow the principles of which it is true that, if they were *universally followed*, things would go best.

Such principles we can call *UF-optimific*. We *follow* some principle when we succeed in doing what this principle requires. For example, we would be following AC if we always did whatever would make things go best.

We have also been discussing what we can now call the *acceptance version* of Kantian Contractualism, or *AKC*. According to a different version of the Kantian Formula, which we can call

> *FKC*: Everyone ought to follow the principles whose being universally followed everyone could rationally will, or choose.

The Kantian Argument discussed above could be revised to show that

> (B) it is only the UF-optimific principles whose being universally followed everyone could rationally will.

This other version of the Kantian Formula therefore requires us to follow these principles.

According to some writers, the Act Consequentialist principle is UF-optimific. For example, Kagan claims that

> (C) if everyone always followed AC, by doing whatever would make things go best, things would go best.

This claim may seem undeniable. And if this claim were true, this version of the Kantian Formula would require us to be Act Consequentialists.

(C) is not, I believe, true. When we ask whether things would go best if everyone followed AC, we should consider all of the ways in which such a world would differ from the other possible worlds in which everyone followed various other principles. We should take into account, not only the effects of people's acts, but also the effects of people's intending to act in these ways, and having the motives that would lead them to act in these ways. For some of the reasons that Sidgwick gave, we can claim that

> (D) if everyone always followed AC, things would go worse than they would go if everyone always followed certain other principles.

If everyone always did whatever would make things go best, everyone's *acts* would, in most cases, have the best possible effects. Things would go better than they would go if everyone always tried to do whatever would make things go best, but such attempts often failed. But the good effects of everyone's acts would again be outweighed, I believe, by the ways in which it would be worse if we all had the motives that would lead us to follow AC. As before, in losing many of our strong loves, loyalties, and personal aims, many of us would lose too much of what makes our lives worth living. So this version of the Kantian Formula does not require us to be Act Consequentialists.

This formula does, however, require us to follow the principles that are UF-optimific. And compared with the UA-optimific principles, these principles are more similar to AC. So this version of the Kantian Formula supports a moral view that is significantly closer to Act Consequentialism.

To cover both versions of the Kantian Formula, we can restate Kantian Contractualism as

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.

Principles could be *universal laws* by being either universally accepted, or universally followed.

Since these different versions of KC and RC have different implications, we might have to choose between them. In making this choice, we would have to consider several questions that I shall not consider here. But I shall mention one possibility. We ought, I have claimed, to distinguish different senses of 'ought' and 'wrong', which we can use in different parts of our moral theory, to answer different questions. It is worth drawing other such distinctions. For example, it is one question what *we ought all ideally to do* if we suppose that we would all succeed. Our answers to this question will be our *ideal act theory*, or what some call our *full compliance theory*. It is another question what we ought to do when we know that some other people will act wrongly. Some call this our *partial compliance theory*. We can also ask what we ought to try to do when we take into account various other facts, such as facts about the mistakes that people would be likely to make, and facts about people's motives, desires, and dispositions. Another question is which motives we ought to have, and what we ought to be disposed to do. Our answers to this question would be our *motive theory*, which would itself have ideal and non-ideal parts. If we are Kantian Contractualists and Rule Consequentialists, we may not need to choose between at least some of these different versions of KC and RC, since we might appeal to these different versions, and use these different senses of 'ought' and 'wrong', in such different parts of our moral theory.

There may be another complication. I have supposed that there is one set of principles that are UA-optimific, and another set that are UF-optimific. If there were two or more such sets, which were significantly different, we would have to choose between these sets of

principles in some other way. There are several possibilities, which I shall not consider here.

We can now return to another part of Kant's view. According to what I have called Kant's

> *Formula of the Greatest Good*: Everyone ought to strive to promote a world of universal virtue and deserved happiness.

We can best promote this world, Kant claims, by following the moral law, as described by Kant's other formulas. Some of these formulas, I have argued, are best revised and combined in Kantian Contractualism. So Kant might have claimed:

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.
>
> (E) What everyone could rationally will to be such laws are the principles whose being universal laws would make things go best, by bringing the world closest to its ideal state.
>
> (F) This ideal state would be a world of universal virtue and deserved happiness.

Therefore

> Everyone ought to follow the principles whose being universal laws would best promote this ideal world.

This argument would give Kant's moral theory its most unified and harmonious form. Kant's Formula of the Greatest Good would describe a single ultimate end or aim which everyone ought to try to achieve, and Kantian Contractualism would describe the moral law whose being universally accepted or followed would best achieve this aim.

Of this argument's premises, KC is Kantian Contractualism. The Kantian Argument in Chapter 16 could be turned, with some revisions, into

a defence of (E). (F) is Kant's description of the ideal world that he calls the Greatest Good.

We ought, I have argued, to revise (F). It would be bad, Kant claims, if people had more happiness, or less suffering, than they deserve. But Kant also claims

> (G) If all of our decisions were merely events in time, no one could deserve to suffer.

We ought, I have argued, to accept this claim. As I have said, we can add:

> (H) All of our decisions *are* merely such events.

Therefore

> (I) No one could deserve to suffer.

Nor could anyone deserve to be less happy. If we subtract Kant's claims about desert, Kant's ideal world would be a world of universal virtue and happiness. In considering worlds that are not ideal, we would again have to decide which worlds would be closer to the ideal. It would always be better, I believe, not only if there was less suffering and more happiness, but also if more of this happiness came to people who were less happy, or who suffered more. We might add that our well-being does not consist merely in happiness and avoiding suffering, and that how well things go depends in part on other facts that are not about anyone's well-being.

Kant's claims about his ideal world raise another question. In asking how we could get closest to Kant's ideal, we must compare the goodness of virtue and happiness. On one view, the goodness of virtue is infinitely greater, so that if anyone became slightly more virtuous, or slightly less vicious, this change would be better than the achievement of any amount of happiness, however great, or the prevention of any amount of suffering. For this view to seem plausible, I believe, we must assume that we have some kind of freedom that could make us responsible

for our acts in some desert-implying way. If there could be no such freedom, as I have claimed, we ought to accept a very different view. If someone is morally bad, by being a cruel murderer for example, this would be bad for the murderer, his victim, and others, and this would also be a bad state of affairs, which we would all have reasons to regret, and try to prevent. But the badness of someone's being a cruel murderer is, I believe, relevantly similar to the badness of someone's being insane. Such badness can be easily outweighed by the badness of great suffering.

This rejection of desert may seem to take us far from Kant's view. But Kant sometimes makes such claims, as when he refers to

> the supreme end, the happiness of all mankind.

And, in an early lecture, Kant said:

> If we conduct ourselves in such a way that, if everyone else so conducted themselves, the greatest happiness would arise, then we have so conducted ourselves as to be worthy of happiness.

Kant here asserts a hedonistic version of Rule Consequentialism.

I shall now sum up these conclusions. Moral principles could be universal laws by being either universally accepted or universally followed. Kantians, I have claimed, can argue:

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.
>
> (J) There are certain principles whose being universal laws would make things go best.
>
> (K) These are the only principles that everyone could rationally will to be universal laws.

Therefore

> RC: Everyone ought to follow these optimific principles.

KC and RC are the most general statements of Kantian Contractualism and Rule Consequentialism. We are supposing that (J) is true. I have, I believe, successfully defended (K). So Kantian Contractualism implies Rule Consequentialism.

Since that is true, these theories can be combined. According to what we can call

> *Kantian Rule Consequentialism*: Everyone ought to follow the optimific principles, because these are the only principles that everyone could rationally will to be universal laws.

64 Climbing the Mountain

Remember next that, according to

> Scanlon's Formula: Everyone ought to follow the principles that no one could reasonably reject.

Kantians might argue:

> (A) If we could not rationally will that one of two principles be a universal law, there must be facts which give us a strong objection to this principle.
>
> (B) If everyone *could* rationally will that the other principle be such a law, no one's objection to this alternative could be as strong.
>
> (C) Since our objection to the first principle is stronger than anyone's objection to this alternative, we could reasonably reject this principle.
>
> (D) When there is only one relevant principle that everyone could rationally will to be a universal law, no one's objection to this principle could be as strong as the strongest objections to every alternative.

(E) No one could reasonably reject some principle if there are stronger objections to every alternative.

Therefore

(F) When there is only one relevant principle that everyone could rationally will to be a universal law, no one could reasonably reject this principle.

(G) Since there are stronger objections to every alternative, these alternatives could all be reasonably rejected.

Therefore

(H) When there is only one relevant principle that everyone could rationally will to be a universal law, this is the only relevant principle that no one could reasonably reject.

(I) There is only one set of principles that everyone could rationally will to be universal laws.

Therefore

These are the only principles that no one could reasonably reject.

We can call this *the Convergence Argument*. If this argument is sound, Kantian and Scanlonian Contractualism can be combined. The principles that no one could reasonably reject are the same as the principles that everyone could rationally will to be universal laws.

This argument applies, not to Scanlon's present theory, but to what I believe to be the best version of Scanlonian Contractualism. I defend this belief, and discuss this argument further, in Chapters 21 to 23.

This combined theory, as I have argued, can also include Rule Consequentialism. According to what we can call this

Triple Theory: An act is wrong if and only if, or *just when*, such acts are disallowed by some principle that is

(1) one of the principles whose being universal laws would make things go best,

(2) one of the only principles whose being universal laws everyone could rationally will,

and

(3) a principle that no one could reasonably reject.

More briefly,

TT: An act is wrong just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable.

We can call these the *triply supported* principles. If some principle could have any of these three properties without having the others, we would have to ask which of these properties had most moral importance. But these three properties, I have argued, are had by all and only the same principles. If that is true, we could claim

(J) Moral principles are not reasonably rejectable just when they are uniquely universally willable, and they are uniquely so willable just when they are optimific.

We could also claim

(K) When some principle is optimific, that makes it one of the only principles that are universally willable,

and

(L) When some principle is one of the only principles that are universally willable, that makes it one of the principles that no one could reasonably reject.

We might add:

(M) When acts are disallowed by some principle that is optimific, universally willable, and not reasonably rejectable, that makes these acts unjustifiable to others.

> (N) Such acts would be blameworthy, and would give their agents reasons to feel remorse, and give others reasons for indignation.
>
> (O) Everyone has reasons never to act in these ways. These reasons are always sufficient, and often decisive.

For the reasons that I earlier gave, this Triple Theory should claim to describe, not wrongness itself, but one of the properties or facts that make acts wrong. There are several other, more particular wrong-making properties or facts, such as the properties of causing pointless suffering or coercing others for our own convenience. The Triple Theory should claim to describe a single *higher-level* wrong-making property, under which all other such properties can be subsumed, or gathered. This higher-level property is the complex property of being disallowed by some principle of which (1), (2), and (3) are true. When acts have certain other properties, that makes them acts that would be disallowed by such a triply supported principle, and all these facts could be claimed to make these acts wrong. Each of these facts, we might add, would give everyone further reasons not to act in these ways.

If we accept this Triple Theory, we should admit that, in explaining why many kinds of act are wrong, we would not need to claim that such acts are disallowed by some triply supported principle. In some cases such a claim would be, not merely unnecessary, but also puzzling or offensive. This is like the fact that, after some rape or murder, we ought not to say 'What if everyone did that?' or 'What if everyone believed such acts to be permitted?' Some acts are open to objections that are both clearer and stronger than the objections to these acts that are provided by Kant's formulas, or by any version of Contractualism or Rule Consequentialism.

In many other cases, however, it may help to ask whether some act is permitted or disallowed by some triply supported principle. It may be unclear, for example, whether it would be wrong to break some law, or tell some lie to achieve some good end, or coerce someone in some way for this person's or someone else's good, or steal some object that its owner never uses, or fail to help some people who are in great need, or fail to

vote, or have, in an overpopulated world, more than two children. If any of these kinds of act would be disallowed by one of the principles whose acceptance would make things go best, and by one of the only principles whose being universal laws everyone could rationally will, and by a principle that no one could reasonably reject, these facts would provide some of the strongest objections to these acts.

Remember next that, on the Triple Theory, an act is wrong *just when* such acts are disallowed by the triply supported principles. There are several lower level wrong-making properties, and several principles that disallow acts with these properties. The Triple Theory makes claims about what all these properties and principles have in common. If this theory's claims are true, that would give us deeper explanations of why these principles are justified, and why these acts are wrong. One aim of such a theory, as Scanlon writes, is to provide 'a general criterion of wrongness that explains and links these more specific wrong-making properties'.

For some moral theory to succeed, it must have plausible implications. The Triple Theory has many such implications. But after we have worked out what this theory implies, and we have carefully considered all of the relevant facts and arguments, this theory might conflict with our intuitive beliefs about the wrongness of certain acts. If there are many such conflicts, or these intuitive beliefs are very strong, we could then justifiably reject this theory. If instead these conflicts are significantly less deep, or less common, we could justifiably follow this theory in revising some of our intuitive moral beliefs.

We have such intuitive beliefs, not only about which acts are wrong, but also about which principles or theories might be true. So as well as having plausible implications, any successful principle or theory must be in itself plausible. Only such a principle or theory could *support* our more particular moral beliefs.

Kantian Contractualism passes this test. If some act is disallowed by one of the only principles whose being a universal law everyone could rationally will, this fact can be plausibly claimed to be one of the facts that make this act wrong.

Scanlonian Contractualism may seem to be, not merely plausible, but undeniable. Suppose I claimed:

> Though my act is disallowed by some principle that no one could reasonably reject, I deny that such acts are wrong.

This claim may seem close to a contradiction. Though I am rejecting this principle, I am also conceding, it seems, that this rejection is unreasonable. And if my rejection of this principle is unreasonable, this rejection could not be justified, so I could not defensibly deny that such acts are wrong. If Scanlon's Formula seems undeniable, however, that is because this formula does not explicitly include the Deontic Beliefs Restriction. In a fuller statement, this formula might claim:

> An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject, on grounds *other than* their belief that this principle is mistaken, because it disallows some acts that are not wrong.

It would not be self-contradictory to claim that, even though some kind of act is disallowed by such a principle, this principle *is* mistaken, because such acts are not wrong.

Kantian Contractualism can be combined, I believe, with the best version of Scanlonian Contractualism. But my arguments for this belief may fail. We would then have to choose between these theories.

Kantian Contractualism could still be combined, however, with Rule Consequentialism. I have argued that

> (K) when some principle is optimific, that makes it one of the principles whose being universal laws everyone could rationally will,

and that

> (P) there are no other principles whose being universal laws everyone could rationally will.

If these claims are true, Kantian Contractualism and Rule Consequentialism fit together like two pieces in a jig-saw puzzle.

Of the Triple Theory's components, Rule Consequentialism is, in one way, the hardest to defend. Some Rule Consequentialists appeal to the claim that

(Q) all that ultimately matters is how well things go.

This claim is in itself very plausible, and is not challenged by any of the arguments that I have given. If we reject (Q), that is because this claim supports Act Consequentialism, which conflicts too often, or too strongly, with our intuitive beliefs about which acts are wrong. Rule Consequentialism conflicts much less with these intuitive beliefs. But if Rule Consequentialists appeal to (Q), their view faces a strong objection. On this view, though the best principles are the principles that are optimific, the right acts are *not* the acts that are optimific, but the acts that are required or permitted by the best principles. It would be wrong to act in ways that these principles disallow, even if we knew that these acts would make things go best. We can plausibly object that, if all that ultimately matters is how well things go, it could not be wrong to do what we knew would make things go best.

Rule Consequentialism may instead be founded on Kantian Contractualism. What is fundamental here is not a belief about what ultimately matters. It is the belief that we ought to follow the principles whose being universally accepted, or followed, everyone could rationally will. Because Kantian Rule Consequentialists do not assume that all that ultimately matters is how well things go, their view avoids the objection that I have just described. When acts are wrong, these people believe, that is not merely or mainly because such acts are disallowed by one of the optimific principles. These acts are also wrong because they are disallowed by one of the only set of principles whose being universal laws everyone could rationally will.

If Kantian Contractualism implies Rule Consequentialism, as I have claimed, that does not make the resulting view wholly Consequentialist.

Though this view is Consequentialist in its claims about which *principles* we ought to follow, it is not Consequentialist either in its claims about *why* we ought to follow these principles, or in its claims about which *acts* are wrong. This view, we might say, is only *one-third* Consequentialist.

In this volume I have argued that, with some revisions and additions, Kant's most important claims are these:

> (R) Everyone ought to treat everyone only in ways to which they could rationally consent.
>
> (S) Everyone ought to regard everyone with respect, and never merely as a means. Even the morally worst people have as much dignity or worth as anyone else.
>
> (T) If all of our decisions are merely events in time, we cannot be responsible for our acts in any way that could make us deserve to suffer, or to be less happy.
>
> (U) Everyone ought to follow the principles whose being universal laws would make things go best, because these are the only principles whose being universal laws everyone could rationally will.

We ought, I believe, to accept (S) and (T), and we have strong reasons to accept (R) and (U).

It may be worth explaining why I have spent so long defending (U). Of our reasons for doubting that there are moral truths, one of the strongest is provided by some kinds of moral disagreement. Most moral disagreements do not count strongly against the belief that there are moral truths, since these disagreements depend on different people's having conflicting empirical or religious beliefs, or on their having conflicting interests, or on their using different concepts, or these disagreements are about borderline cases, or they depend on the false assumption that all questions must have answers, or precise answers. But some disagreements are not of these kinds. These disagreements are deepest when we are considering, not the wrongness of particular acts,

but the nature of morality and moral reasoning, and what is implied by different views about these questions. If we and others hold conflicting views, and we have no reason to believe that *we* are the people who are more likely to be right, that should at least make us doubt our view. It may also give us reasons to doubt that any of us could be right.

It has been widely believed that there are such deep disagreements between Kantians, Contractualists, and Consequentialists. That, I have argued, is not true. These people are climbing the same mountain on different sides.

It has also been widely believed that nothing matters, since reasons are given by our desires, and we have no reasons to have these desires. As I have argued, and shall argue further in Part Six, we ought to reject this bleak view.

What now matters most is that we rich people give up some of our luxuries, ceasing to overheat the Earth's atmosphere, and taking care of this planet in other ways, so that it continues to support intelligent life. If we are the only rational animals in the Universe, it matters even more whether we shall have descendants during the billions of years in which that would be possible. Some of our descendants might live lives and create worlds that, though failing to justify past suffering, would give us all, including those who suffered, reasons to be glad that the Universe exists.

APPENDIX A

STATE-GIVEN REASONS

According to what we can call

> *the State-Given Theory:* Whenever certain facts would make it better if we had some belief or desire, these facts give us a reason to have this belief or desire.

To decide whether we have such *state-given* reasons, we can first ask how we might respond to such reasons.

Suppose that, in

> *Case One*, some whimsical Despot credibly threatens that I shall be tortured for ten minutes unless, one hour from now, I both believe that $2 + 2 = 1$, and want to be tortured. Some lie-detector test will reveal whether I really have this belief and desire.

On the State-Given Theory, this man's threat gives me strong state-given reasons to have this belief and desire, since that is my only way to avoid being tortured. But I could not respond to such reasons by choosing to have this belief and desire.

One problem here is that I have *object*-given reasons that count decisively *against* believing that $2 + 2 = 1$, and *against* wanting to be tortured. Suppose that, because I fail to have this belief and desire, this Despot tortures me. Someone might say: 'You idiot! Why didn't you believe that $2 + 2 = 1$?' But this remark would be absurd. I could not help believing that $2 + 2$ does *not* $= 1$. It would also be absurd to claim that I was an idiot in not wanting to be tortured. I might want to be tortured if I knew that this would be my only way to achieve some great good. That might be true, for example, if I have some life-threatening illness, and great pain would trigger some healing process in my body. But this example is not of that kind. This Despot will carry out his threat

unless I want to be tortured, not as a means to some end, but as an end, or for the sake of being tortured. Since I am rational, I could not want to be tortured for its own sake. Given the awfulness of being tortured, I have a decisive object-given reason *not* to have this desire, and I could not help responding to this reason in the non-voluntary way.

Suppose next that this Despot gives me an easier task. In

> *Case Two*, I shall be tortured unless, one hour from now, I believe that a certain closed box is empty.

On the State-Given Theory, this threat gives me a state-given reason to have this belief. And this reason would be unopposed, since I have no object-given epistemic reason *not* to believe that this box is empty. But as before, I could not respond to this alleged state-given reason by choosing to have this belief. Since I am rational, I could not choose to believe that this box is empty simply because I know that it would be better for me if I had this belief.

There are other possibilities. When it would be better for us if we had some belief, there are three main ways in which we might be able to cause ourselves to have this belief. One method is to make this belief true. In *Case Two*, for example, I might be able to open the closed box and take out anything that it contains. That would make me believe that this box is empty, thereby saving me from my Despot's threat.

In some other cases, we might cause ourselves to have some beneficial belief by finding evidence or arguments that gave us strong enough epistemic reasons to have this belief. This method is risky, since we might find evidence or arguments that gave us strong reasons *not* to have this belief. But we might reduce this risk by trying to avoid becoming aware of such reasons. If we are trying to believe that God exists, for example, we might read books written by believers, and avoid books by atheists. While we are acting in this way, it is worth adding, we may be fully rational not only practically but also epistemically. We may always respond rationally to our awareness of any epistemic reason or apparent reason. This may be why we have to take such care to avoid becoming aware of epistemic reasons not to believe what we are trying to believe.

In a third kind of case, it would be better if we had some belief that we know to be false, because we are aware of facts that give us decisive epistemic reasons not to have this belief. If we are rational, we could not have this belief while we are aware of these decisive reasons not to have it. But we might be able to make ourselves have this belief by using some technique like self-hypnosis. We could not choose to give ourselves beliefs whose content makes them too obviously false. When my Despot makes his first threat, I could not make myself believe that $2 + 2 = 1$. No one could both understand this mathematical equation and believe it to be true. But suppose that, in

> *Case Three*, this Despot threatens that I shall be tortured unless, one hour from now, I believe that he is the world's greatest genius.

I might be able to hypnotize myself into having this false belief. I would have to make myself forget my epistemic reasons to believe that this man is *not* a genius. I might also have to make myself forget how and why I had caused myself to have this new, false belief, since my remembering these facts would be likely to undermine this belief. Since I am rational, I could not believe what I knew that I had no epistemic reasons to believe. For similar reasons, I might also have to give myself some false apparent memories of this Despot's brilliant achievements. But if I am a skilled self-hypnotist, I might be able to do these things. I would then rationally come to believe that this man is the world's greatest genius, because these false apparent memories would give me decisive apparent reasons to have this belief.

Most of us do not have such self-hypnotic powers. But we can imagine coming to have them. We could then make ourselves have many false beliefs at will, just as directly as we can perform various other mental acts.

Return now to the view that we can have state-given reasons. State-Given Theorists claim that

> (1) whenever certain facts would make it better if we had some belief, these facts give us a reason to have this belief.

In cases of the kinds that I have just described, we would have no need to appeal to such reasons. It would be enough to claim that we have reasons to want to have such beneficial beliefs, and to cause ourselves to have them, if we can. These would be like any other reasons to want something to happen, and to make it happen if we can. There would be no point in adding that, as well as having reasons to *cause* ourselves to have such beliefs, we would have reasons to *have* them.

We can imagine another change in our psychology. It might become true that, when we believed that it would be better if we had some epistemically irrational belief, we sometimes didn't need to make ourselves have this belief with some voluntary mental act, like self-hypnosis. We might find ourselves coming to have such beneficial beliefs, with supporting sets of false apparent memories, in a non-voluntary way.

It may seem that, in *these* cases, we *could* significantly claim that we had state-given reasons to have these beliefs. As I have said, when we are aware of facts that give us decisive epistemic reasons to *have* some belief, we respond to most of these reasons, not by voluntarily *causing* ourselves to have this belief, but by coming to have this belief, and then continuing to have it, in a non-voluntary way. We might similarly claim that, when we found ourselves coming to have such irrational but beneficial beliefs, we would be responding to practical reasons to *have* these beliefs.

We ought, I suggest, to reject these claims. There would be two other, better ways to describe such cases.

On one description, in coming to have these beneficial beliefs, we would still be responding, though in a non-voluntary way, to our reasons to cause ourselves to have these beliefs. We often find ourselves doing something that we could also voluntarily do. For example, we might find ourselves suddenly trying to catch some object that we have just dropped, or moving our body to regain our balance, or raising our arms when we are falling so as to protect our head. If we saw some hand grenade that was about to explode, we might find ourselves throwing ourselves onto this grenade, to save the lives of those around us. These

would be non-voluntary responses to our reasons to act in certain ways. Suppose that, when my Despot makes his third threat, I find myself coming to believe that this man is a genius. I might here be responding in this non-voluntary way to my practical reason to cause myself to have this beneficial belief. This may be what happens in some actual cases of unconscious self-deception.

We might instead claim that, when we found ourselves coming to have such beneficial beliefs, we would not be responding to any reasons. The truth might be only that, when we believed that it would be better if we had some other belief, this belief would cause us to have this other belief. This would be partly like the way in which, when we believe that we are in danger, this belief causes adrenalin to be released into our blood stream, thereby helping us to respond more effectively to this danger. This release of adrenalin, though beneficial, does not involve a response to some reason. Nor, perhaps, do some cases of wishful thinking.

Return now to the claim that, in such cases, we would be responding to our reasons to *have* these beneficial beliefs. We ought, I have suggested, to reject this claim. If we were *causing* ourselves to have these beliefs, this process might be rational, and involve responses to reasons. We would be responding to reasons for *acting*, which would be provided by the facts that would make it good if we had these beliefs. But if we were merely *passively* coming to have these beliefs, this process would not be rational, or involve any response to reasons. Suppose that I cannot hypnotize myself into believing that my Despot is a genius. As a result, he tortures me. Someone might say: 'You idiot! Why didn't you respond to your reasons to believe this man to be a genius?' When we are aware of facts that give us decisive *epistemic* reasons to have some belief, we are less than fully rational if we fail to respond to these reasons by coming to have this belief. But if we cannot cause ourselves to have some beneficial but irrational belief, we would not be open to the slightest criticism if we failed to have this belief. And if we would be in no way irrational despite our failure to respond to our awareness of certain alleged reasons, this counts against the view that we have any such reasons.

We have other reasons to reject the State-Given Theory. Two reasons, we can say,

> *compete* when we could not successfully respond to both these reasons,

and they

> *conflict* when they support different answers to the same question.

If we have a moral reason to keep some promise, for example, and a self-interested reason to break this promise, these reasons compete, since we couldn't both keep and break this promise. These reasons also conflict, since they support different answers to the question of what we have most reason to do.

Suppose next that we are aware of facts that give us decisive epistemic reasons *not* to have some beneficial belief. According to the State-Given Theory, the benefits of having this belief would also give us state-given reasons to have it. These two sets of reasons would compete, since we could not both have and not have this belief. On one version of this view, these reasons would also conflict. When we ask what we had most reason to believe, these reasons would support different answers to this question. We would have to decide whether our state-given reasons to have this belief were stronger than, or outweighed, our epistemic reasons not to have this belief.

We would not, I believe, have such conflicting reasons. When my Despot makes this third threat, I would be aware of facts that gave me decisive epistemic reasons *not* to believe falsely that this man is the world's greatest genius. If I had a state-given reason to have this belief, this reason would be provided by the facts that would make it bad to be tortured. I might ask whether, compared with being tortured, it would be worse to have such a false belief. But I would here be asking which of two outcomes I had more reason to want to prevent and to try to prevent. That is a question about the strength of two *practical* reasons, like any

other reasons for wanting to prevent and trying to prevent some bad outcome. I could not rationally ask whether my state-given reason to have this false belief is stronger than, or outweighs, my *epistemic* reasons *not* to have it. It makes no sense to compare the strength of my evidence for the falsity of this belief with the badness of my being tortured.

Having seen that such comparisons make no sense, State-Given Theorists might turn to the claim that these two kinds of reason do not conflict, since they support answers to different questions. When we ask whether we ought to have some belief, we might be asking either

> Q1: Is this a belief that I *ought epistemically* to have?

or

> Q2: Is this a belief that I *ought practically* to have?

On this view, in answering Q1, we should consider only epistemic reasons; and in answering Q2, we should consider only practical state-given reasons. Since these are different questions, we cannot ask what we ought to believe, or what we have most reason to believe, *all things considered*.

These claims are partly right. There are, indeed, two questions here. But these claims do not help to show that we can have practical state-given reasons to have beliefs. Q2 needs to be explained, since it is unclear what it means to ask whether we *ought practically* to have some belief. This question could be more clearly stated, I suggest, as

> Q3: What would it be best for me to believe? In other words, what do I have most reason to want to believe, and to cause myself to believe, if I can?

And this question is not about what I have reasons to *believe*. Like other practical questions, this question is about what I have reasons to *want*, and to *do*.

Since Q1 and Q3 are different questions, we never need to compare the strength of practical and epistemic reasons. We *respond* to reasons. And we could never have practical reasons to respond in a certain way, while having epistemic reasons *not* to respond in this same way. When

my Despot makes his third threat, I might respond to my practical reasons by acting in a way that would make me believe that this man is the world's greatest genius. I have no epistemic reasons *not* to act in this way, since epistemic reasons are not reasons for *acting*. I do have decisive epistemic reasons not to *believe* that this man is such a genius, and while I remember the facts that give me *these* reasons, I might respond to them in a non-voluntary way by losing this belief. But I have no practical reasons *not* to respond in this non-voluntary way. My practical reasons are to act in ways that would make me keep this belief until I have passed this Despot's lie-detector test, so that he will not torture me. These practical and epistemic reasons do *compete*, in the sense that I could not successfully respond to both sets of reasons. But these reasons do not conflict.

It is easy to overlook, or misunderstand, the distinctions that I have just drawn. As I have said, theoretical reasoning is a voluntary activity, in which we often engage for practical reasons. When we are doing mathematics, for example, we may have a practical reason to check some part of some proof, or to redo some calculation in a different way. These are reasons for acting in ways that may help us to reach the truth. While we are acting in these ways, for these practical reasons, we shall also respond to many epistemic reasons. While we are checking some proof, for example, we respond to epistemic reasons whenever we see what follows from what, and what must be true. Coming to have some such particular belief is not a voluntary mental act. Theoretical reasoning, we might say, involves both *practical* and *pure* epistemic rationality.

There are other close connections between practical reasons and certain epistemic reasons. Much of our practical reasoning consists in theoretical reasoning about practical questions. When we ask what we have most reason to do, we may be trying to reach some true answer to this question. And some facts may give us both a decisive practical reason to act in some way, and a decisive epistemic reason to believe that we have this practical reason. Return to the case in which your hotel is on fire, and you could save your life only by jumping into some canal. This fact would give you a decisive reason to jump, and a decisive

reason to believe that you ought to jump. But though our practical and epistemic reasons are often very closely related, and these kinds of reason can compete, they cannot ever conflict.

State-Given Theorists also claim that

> (2) whenever certain facts would make it better if we had some desire, these facts give us a reason to have this desire.

Compared with the claim that we can have state-given reasons to have beliefs, this claim is more plausible. We can object that, since beliefs aim at the truth, our reasons to have beliefs must all be epistemic, or truth-related. No such claim applies to desires. So it may seem that, just as we have an object-given reason to have some desire when, and because, *what we want* would be relevantly good, we have a state-given reason to have some desire when, and because, *our wanting something* would be good.

We do not, I suggest, have such reasons. Suppose that, in

> *Case Four*, my Despot declares that I shall be tortured for ten minutes unless, one hour from now, I want him to kill me. If I have this desire, and ask him to kill me, he will refuse, and set me free. As I know, this man always does what he declares that he will do.

Suppose next that the rest of my life would be well worth living. I would then find it difficult to want this man to kill me. But I might be able to hypnotize myself into having this desire during the next few hours. That would be what I had most reason to do, and what I ought rationally to do. This mental act would be a riskless way to avoid some intense pain.

State-Given Theorists might claim that their view explains why I ought to act in this way. They might argue:

> (A) I have a decisive reason to want this Despot to kill me, since that would save me from being tortured.
>
> (B) When we have a decisive reason to have some desire, this fact gives us a decisive reason to make ourselves have this desire, if we have some riskless way of doing that.

(C) I have such a way of making myself want this man to kill me.

Therefore

I ought to make myself have this desire.

Premise (A), however, is false. I have object-given reasons to want this Despot *not* to kill me, and these are also reasons not to want this man to kill me. These reasons are clearly stronger than my alleged state-given reason to want this man to kill me. Losing a life worth living is much worse than being tortured for ten minutes. So I do not have a decisive reason to want this man to kill me.

State-Given Theorists might reply that I don't have any reason not to want this man to kill me. If I had this desire, this man would not kill me but set me free. Since I have a reason to have this desire, and no reason not to have it, I ought rationally to cause myself to have this desire. On this view, all reasons to have desires are state-given, or provided by the benefits of having these desires.

To assess this view, we can suppose that, because my attempt to have this desire fails, this Despot tortures me. Someone might say: 'You idiot! Why didn't you want him to kill you?' But this remark would be unjustified. As before, if I am rational, I could not want this man to kill me merely because I know that, if I had this desire, that would be better for me. This point is clearer in a simpler case. If I learnt that I was fatally ill, it might be better for me if I wanted to die. But that wouldn't show that I had no reason to want not to die. It would be absurd for others to say 'You idiot! Why don't you want to die?' We should admit that, even after this Despot has made his threat, I have decisive object-given reasons to want this man not to kill me.

State-Given Theorists might next suggest that, since these reasons are of different kinds, they do not conflict. On this view, we can ask two questions:

Q4: What do I have the strongest object-given reasons to want?

Q5: What do I have the strongest state-given reasons to want?

But this suggestion fails. We can also ask

> Q6: What do I have most reason to want all things considered?

If we have reasons for and against having the same desire, these reasons *do* conflict, since they support different answers to this wider question. It is irrelevant that these reasons are of different kinds. It might be similarly claimed that moral and self-interested reasons are of different kinds: but, when we ask what we have most reason to do all things considered, these reasons can conflict, by supporting different answers to this question.

In cases of the kind that we are now discussing, there *are* two questions that are worth asking. But these are not questions about two kinds of reason for or against having the same desire. Q6 can be restated as

> Q7: Which desires do I have most reason to have?

We can also ask

> Q8: Which desires do I have most reason to want to have, and to cause myself to have, if I can?

In *Case Four*, I could ask:

> If I wanted this Despot to kill me, would I be wanting something that I have decisive reasons to want?
>
> If I caused myself to have this desire, would I be doing something that I have decisive reasons to do?

My answers should be No and Yes. If I wanted this man to kill me, this desire would be in itself irrational, since I have decisive reasons *not* to want this man to kill me. But it would be rational for me to cause myself briefly to have this irrational desire, since this act would save me from being tortured.

There is another kind of case that gives us reasons to deny that we have state-given reasons to have desires. Suppose that, in

> *Self-defeating Desire*, I have a strong desire to get to sleep, because I need to sleep to improve my performance in some interview tomorrow. But I have one kind of insomnia. Whenever I strongly want to get to sleep, this desire makes me anxious about my failure to become sleepy, thereby keeping me awake. So I shall get the sleep I need only if I lose my desire to get to sleep.

My need for sleep gives me an object-given reason to want to get to sleep. According to the State-Given Theory, this need also gives me a state-given reason *not* to have this desire, since that would be my only way to get to sleep. These reasons would conflict, since they would be reasons for and against having the same desire. On this view, to decide whether I ought to have this desire, I should compare the strength of these two reasons. I should ask what I have most reason to want, all things considered.

I could easily compare the strength of these two reasons. My object-given reason to want to get to sleep is provided by the fact that I need sleep to improve my performance in my interview tomorrow. My alleged state-given reason *not* to have this desire would be provided by this same fact, together with the fact that having this desire would keep me awake. Since these reasons would both get their normative force from my need for sleep, their strength would be precisely equal. Since these reasons would also conflict, they would cancel each other out. The State-Given Theory therefore implies that, on balance, I have no reason to want to get to sleep. If that were true, I would have no reason to have the aim of getting to sleep, and no reason to cause myself to lose this desire, so that I could achieve this aim. These claims are clearly false.

We ought, I suggest, to reject this State-Given Theory. I have no state-given reason not to have my desire to get to sleep. What I have are *object*-given reasons to *want* not to have this desire, and to *cause* myself to lose this desire, if I can. Unlike my alleged state-given reason *not* to have this desire, these reasons do not conflict with my object-given reason to *have* this desire. On this view, we reach the right conclusion. My need for sleep gives me a strong and unopposed reason to want to get to sleep, and this need also gives me a strong and unopposed reason

to cause myself to lose this desire, since that is my only way to fulfil this same desire, thereby getting the sleep I need.

Whenever it would be better if we had certain beliefs or desires, we have reasons to want to have these beliefs or desires, and to make ourselves have them, if we can. But we do not, I suggest, have *state*-given reasons to have beliefs or desires.

We may have state-given reasons to be in some other kinds of state. I might truly claim, for example, that I have a reason to be in Paris next April. But as I have argued, such reasons would have no importance. It would be enough to claim that I have reasons to want to be in Paris next April, and to go there, if I can.

APPENDIX B

RATIONAL IRRATIONALITY AND GAUTHIER'S THEORY

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

1

Gauthier assumes that, to be rational, we must maximize our own expected utility. Though he distinguishes between 'utility' and 'benefit', this distinction does not affect his main arguments. We can regard him as appealing to Rational Egoism.

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

If we appeal to Rational Egoism, it may seem impossible to defend that claim. How can our acts be rational, in self-interested terms, if we know them to be worse for us? But Gauthier *revises* Rational Egoism. On the standard version of this theory, an act is rational if it will maximize our expected benefit—or be *expectably-best* for us. On Gauthier's version, it is rational to benefit ourselves not with our *acts* but with our *dispositions*. A disposition is rational if having it will be expectably-best for us. An act is rational if it results from such a disposition. In making these claims, Gauthier's view is like a version of Indirect Consequentialism.

Besides revising Rational Egoism, Gauthier restricts the scope of morality. To act morally, Gauthier claims, we must honour our agreements. In the cases with which he is concerned, each of us promises

that, at some cost to ourselves, we shall give a greater benefit to others. If we all kept such promises, we would all gain. The cost to each would be outweighed by the greater benefits that each received from others.

Though such agreements are mutually advantageous, it would often be better for each of us if he or she broke this promise. Either we could break it secretly, or the damage to our reputation would be outweighed by what we would gain. We may think that, in self-interested terms, it is rational to break such promises. But Gauthier argues that, if we do, we are fools.

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

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

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

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

2

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

In his earliest statements of his view, Gauthier assumed

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

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

I gave the following example, which I shall here call *Your Fatal Threat*. Suppose that you and I are on a desert island, and we are both transparent. You become a *threat-fulfiller*. By regularly threatening to explode some bomb, you aim to make me your slave. My only way to preserve my freedom is to become a *threat-ignorer*, who is disposed never to give in to your threats. Since I am translucent, I can reasonably expect you to be aware of my disposition, which would be best for me. I manage to acquire this disposition. But I have bad luck. In a momentary lapse, you threaten that, unless I give you a coconut, you will blow us both to pieces. According to (A), it would be rational for me to ignore your threat. This would be rational even though I know that, if I do, you will explode your bomb, killing us both.

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

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

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

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

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

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

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

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

To handle this version of the case, Gauthier must again change his view. It may have been rational for me to become a threat-ignorer. But as Gauthier must agree, it would now be rational for me to try to lose this disposition. If I could cause myself to lose this disposition, it would be irrational to allow myself to keep it. Since that is so, Gauthier cannot

claim that it must still be rational to act upon it. Now that I could soon be free, it would be irrational for me knowingly to bring about my death.

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

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

If he appealed to (C), Gauthier's view would not be challenged by my example. When I see that my disposition has become disastrous, (C) does not imply that it must still be rational for me to act upon it.

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

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

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

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

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

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

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

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

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

Gauthier admits as much in retreating from claim (A). Suppose that, though it was rational for me to acquire some disposition, I have learnt that doing so was a terrible mistake. Gauthier no longer claims that it must still be rational to act upon such dispositions. He agrees that, from the fact that I rationally acquired some disposition, and that I cannot now overcome it, we cannot infer that it is rational for me to act upon it.

3

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

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

Reconsider *Schelling's Case*. Because I am temporarily insane, the robber knows that, even if he starts to injure my children, he would not thereby induce me to unlock my safe. That gives him reasons to give up and leave, which will be much better for me. But while I am in my drug-induced state, and before the robber leaves, I act in damaging and self-defeating ways. I beat my children because I love them. I burn my manuscripts because I want to preserve them.

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

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

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

Hume notoriously claimed that it would not be contrary to reason to prefer our own total ruin to the least uneasiness of some stranger. But Gauthier's view is more extreme. Hume at least required that, for our acts to be rational, we must be trying to achieve our aims. On Gauthier's view, we could be trying to frustrate our aims. When I burn my manuscript, or beat my children, I might be doing what I believe to be irrational, and *because* I believe it to be irrational. My acts could be as crazy as we can imagine. They could still, on Gauthier's view, be rational. That is clearly false.

4

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

Gauthier assumes that, to be rational, we should maximize our own expected utility. He compares two versions of this view. According to the standard version of Rational Egoism, which we can call *E*, we should maximize at the level of our acts. An act is rational if it maximizes our benefits or expected benefits. According to Gauthier's view, we should maximize only at the level of our dispositions. An act is rational if it results from a benefit-maximizing disposition. This view we can now call *G*.

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

When we cannot maximize at both levels, it would be better for us if we had maximizing dispositions. The good effects of these dispositions would outweigh the bad effects of our acts.

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

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

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

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

Gauthier might now say that, if we accept E, we would be *unable* to acquire these dispositions. We would believe that, in some cases, acting

on these dispositions would be irrational. And we might be unable to make ourselves disposed to do what we believe to be irrational. Perhaps, to acquire these dispositions, we must accept Gauthier's view, and believe that it is rational to act upon them.

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

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

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

or is the best view. To reach that conclusion, Gauthier needs another premise.

In the original version of his argument, Gauthier's other premise was—surprisingly—the standard version of Rational Egoism. He assumed that we should start by accepting E. We should believe that an act is rational if it will be expectably-best for us. He then claimed that it would be better for us if we changed our own conception of rationality, by moving from E to G. Since it would be better for us if we made this change, E implies that it would be rational to do so. S tells us to believe that the true theory is not E but G. Gauthier concluded that the true theory *is* G.

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

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

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

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

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

Such an argument, however, would conflate two kinds of rationality. When we claim that it would be rational to have some belief, we usually mean that this belief would be *theoretically* or *epistemically* rational, since we have sufficient epistemic reasons to have it. Such reasons *support* this belief, since they are provided by facts which either entail

this belief, or make it likely that this belief is true. But Gauthier's argument does not appeal to epistemic reasons. His claim would be that, since it is in our interests to believe his view, this belief would be *practically* rational. When we have practical reasons to cause ourselves to have some belief, these reasons do not support this belief, since they are not related, in relevant ways, to this belief's truth.

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

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

Gauthier, however, rightly rejects (D). He imagines a demon who rewards various beliefs about rationality. He then claims that, if there were such a demon, it would be 'rational to hold false beliefs about rationality'. Gauthier here concedes that, though it would be in our interests to hold these beliefs, they would still be false. The fact that they would be in our interests could not make them true.

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

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

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

5

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

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

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

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

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

This distinction, I believe, undermines Gauthier's reply to Lewis. If some policy is justified despite having bad effects, we may agree that, in one sense, these effects 'should occur'. But this only means, 'Things should be such that they occur'. And in accepting that claim, we need not endorse, or welcome, these effects. If we are giving a dinner party, things should be such that we later have to do the washing up. We can

still have reasons to regret having to wash up. Similar claims apply to the acts that result from an advantageous disposition. We can agree that, in one sense, these acts should be performed. Things should be such that these acts will be performed. But we can still, consistently, believe these acts to be regrettable and irrational.

6

Gauthier suggests another argument in favour of his view. This view avoids, he claims, 'some of the unwelcome consequences' of Rational Egoism. The chief such consequence is that, on that theory, it could be a curse to be rational.

This argument does not, I believe, support Gauthier's view. Gauthier admits that, even on his view, it might be a curse to be epistemically rational. That would be true if epistemic irrationality were directly rewarded. This unwelcome consequence, Gauthier claims, could not be avoided by any theory. But that is not true. Gauthier could extend his view. He could similarly claim that our theoretical reasoning is epistemically rational if and only if it is in our interests. On this version of Gauthier's view, epistemic rationality could never be a curse. This revision would not, however, improve Gauthier's view. When crazy reasoning would be in our interests, that does not make it rational.

Epistemic irrationality could be in our interests, as any good theory should admit. So could practical irrationality. Both kinds of irrationality could be rewarded. It is no objection to Rational Egoism that it assumes or accepts these facts.

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

That is not so. I was asking what, in the ordinary sense, it is rational to want and do. And Gauthier claims that *Schelling's Case* 'shows that our ordinary ideas about rationality . . . are sometimes mistaken.' Since

Gauthier is arguing that we should revise our ordinary ideas, he cannot defend his use of 'rational' by making it a mere stipulation, which is true by definition. And that would also make his view trivial.

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

Gauthier proposes a Hobbesian version of Contractualism, and defends a minimal morality, because he believes he can then show that, even in self-interested terms, we are rationally required never to act wrongly. No other moral theory, Gauthier claims, achieves this aim. If Gauthier's argument fails, as I have claimed, we lose our main reason to accept Gauthier's minimal morality.

APPENDIX C

DEONTIC REASONS

In defending premise (E) of the Kantian Argument for Rule Consequentialism, I suggest that

> (X) if the optimific principles require certain acts that we believe to be wrong, the features or facts that, in our opinion, make these acts wrong would not give us decisive *non*-deontic reasons not to act in these ways. What might be true is only that, by making these acts wrong, these facts would give us decisive deontic reasons not to act in these ways.

It may seem that, to defend (X), we could appeal to the claim that

> (1) if these acts were not wrong, we would not have decisive reasons not to act in these ways.

But it may be difficult to defend this claim. If certain facts would make certain acts wrong, it is hard to suppose that such acts are not wrong, since there may be no possible world in which that is true. And even if we could appeal to (1), that would not show that it is the wrongness of these acts that gives us decisive reasons not to act in these ways. There may be facts that would make certain acts wrong if and only if these facts also gave us decisive non-deontic reasons not to act in these ways.

I know of no quick argument for (X), which is why I merely suggest that (X) is true. But one argument against (X) is worth discussing. When some people claim that some act is wrong, these people mean that we have decisive moral reasons not to act in this way. Though these people appeal to *moral* reasons, they would deny that there are any *deontic* reasons. On this view,

> (2) when some act is wrong, this fact is the second-order fact that certain other facts give us decisive moral reasons not to

> act in this way, and the fact that we had these reasons would not give us a *further*, independent or non-derivative reason not to act in this way.

This claim conflicts with (X), since (2) implies that

> (3) if the optimific principles required some acts that are wrong, we would have decisive non-deontic reasons not to act in these ways.

Most of us, I believe, do not use 'wrong' in this *decisive-moral-reason* sense. Since we use 'wrong' in some other sense, we could justifiably reject (2). And (2), I believe, is least plausible in precisely the cases that we are now considering. If the optimific principles did require some acts that are wrong, it is acts of the kind that we are now considering whose wrongness could most plausibly be claimed to give us a further, independent reason not to act in these ways. In some of these cases, we might even claim, the wrongness of these acts would give us our *only* reason not to act in these ways. If some method of contraception would be artificial, for example, this fact, when considered by itself, seems to give us no reason not to act in this way.

This example does not show that (2) is false if, as most of us believe, such methods of contraception are not wrong. In asking whether (2) is true, we cannot usefully consider acts that are clearly wrong, and ask what would be true if such acts were not wrong. As I have said, this counterfactual may be impossible, or at least too hard to imagine. But it may help to consider how certain people have changed their moral view. In describing this change of view, I shall redescribe these people's beliefs so that they apply to my imagined cases rather than to the slightly different versions of these cases which these people actually considered. Suppose first that, in

> *Bomb*, the runaway train is headed for the tunnel in which it would kill the five. You could save the five by throwing a bomb in front of the train. But I am standing nearby, so this bomb's explosion would also kill me.

Many people would believe this act to be wrong. After considering such cases, certain people accepted

> *the Priority Principle*: The negative duty not to kill has priority over the positive duty to save people's lives.

In explaining this principle, these people claimed that

> (4) it would be wrong to save several people's lives in some way that would also kill someone else.

Remember next that, in

> *Tunnel*, you could redirect the runaway train onto another track so that it would kill me rather than the five.

This imagined case has been much discussed, though it has little practical importance, because this case seems to many people a counter-example to the Priority Principle. When they considered *Tunnel*, several supporters of this principle changed their mind. These people ceased to believe (4). On their view, you would be morally permitted to save the five by redirecting the train, even though your act would also kill me. These people then supposed that, in

> *Bridge*, you could save the five only by causing me to fall onto the track, thereby killing me but stopping the train.

This act, these people believed, *would* be wrong. These people concluded that, though it would *not* be wrong to save several people's lives by redirecting some threat so that it would kill fewer people, it *would* be wrong to save these people *by* killing someone else.

According to (2), an act's wrongness does not give us a further, independent reason not to do it. That is true, some people believe, because the claim that some act is wrong adds nothing to the claim that we have decisive moral reasons not to act in this way. If these claims were true, it would always be enough to ask whether we have such decisive reasons not to act in some way. We would never need to ask, as a separate question, whether some act would be wrong.

These claims are, as I have said, least plausible in precisely the kinds of case that we are now discussing. I have just described how, when comparing cases like *Bomb*, *Tunnel*, and *Bridge*, several people changed

their moral view. This was not a change of view about the strength of our reasons to act in certain ways. When these people considered *Tunnel*, they did not first decide that you would have sufficient reasons to save the five by redirecting the train, and then conclude that, since you would have such reasons, this act would not be wrong. What struck them first was that this way of saving the five would not be wrong. Some of these people then concluded that, since this act would not be wrong, the fact that you would be saving several people's lives would give you sufficient reasons to act in this morally permissible way. Similar claims apply to *Bridge*. When these people considered this example, they did not first decide that you would have a decisive reason not to save the five by killing me, and only then conclude that this act would be wrong. These people were struck first by the belief that this act would be wrong, and only then concluded that the wrongness of this act gave you a further, and perhaps decisive reason not to act in such a way.

Some of us, I have claimed, use the word 'wrong' in an indefinable sense, which I express with the phrase 'mustn't-be-done'. It is in cases like *Tunnel*, *Bomb*, and *Bridge* that we can most plausibly believe that certain acts are in this sense wrong. In both *Tunnel* and *Bridge*, you could save the five by acting in a way that would also kill me. From my point of view, being killed as a means in *Bridge* would be no worse than being killed as a side-effect in *Tunnel*. But of these similar acts, many people believe, it is only killing as a means that has the distinctive property of being something that mustn't-be-done. Such acts are *out*, or *impermissible*. And if some act mustn't-be-done, we can plausibly believe, this fact gives us a further, independent reason not to act in this way. These are the cases in which it seems *least* plausible to claim that, when some act is wrong, this fact doesn't give us any further reason not to do it.

If we can justifiably reject (2), as I have just argued, we can reject this argument against (X). I am therefore inclined to believe that, when the optimific principles require certain acts, we would never have decisive *non*-deontic reasons not to act in these ways.

End Notes

Some of these notes can be read on their own, since I quote enough of the passages to which these notes apply. In other notes I quote the first words of some block of text and some relevant later words. I give references in a later set of notes.

xxxiv *Though Sidgwick's book is long, and some of its chapters can now be ignored . . .* We can ignore Book I Chapter II, Book II chapter VI, and Book III Chapter XII. But we should read the sixth Preface.

xxxvi *. . . the first edition . . . contains only a few jokes, some of which Sidgwick later removed.* For example, Sidgwick dropped his claim that, if Utilitarians know that some community's happiness would be enhanced by having not only a great deal of virtue but also a little of 'what is commonly blamed as vice', and they believe that 'others will supply the virtue', they may 'think themselves justified' in 'supplying the vice' (451).

. . . the Cosmos of Duty . . . When asked if he was proud of *The Methods*, Sidgwick replied 'The first word in my book is *Ethics* and the last is *failure*'.

xxxviii *Sidgwick's irony . . . more than mildly adventurous to argue, though in guarded terms . . .* Sidgwick writes: 'And if we consider the matter in its relation to the individual's perfection, it is certainly clear that he misses the highest and best development of his emotional nature, if his sexual relations are of a merely sensual kind: but we can hardly know *a priori* that this kind of relation interferes with the development of the higher (nor indeed does experience seems to show that this is universally the case). And this latter line of argument has a further difficulty. For the common opinion that we have to justify does not merely condemn the lower kind of development in comparison with the higher, but in comparison with none at all. Since we do not positively blame a man for remaining celibate (though we perhaps despise him somewhat unless the celibacy is adopted as a means to a noble end): it is difficult to show why we should condemn—-in its bearing on the individual's emotional perfection only—the imperfect development

afforded by merely sensual relations' (ME 359). After mentioning Sidgwick's use of the word 'purity', Williams calls this 'no doubt part of what Bloomsbury found oppressive and stuffy' (Williams (2003) 283). There is a strange reversal here. The Bloomsbury Group did not find Moore stuffy, though Moore refers to 'the pleasures of lust . . . of which the enjoyment is certainly an evil in itself' (Moore (1903) 209).

xl *Though Sidgwick makes mistakes, some of which I mention in a note* . . . Though Sidgwick wrote 'I will not stir a finger to compress the world into a system', he later did that. Sidgwick writes: 'If we are not to systematise human activities by taking Universal Happiness as their common end, on what other principles are we to systematise them?' (ME 406). He should not have assumed that we *are* to systematise these activities, and that we should therefore be Hedonists. Sidgwick is mistaken, I believe, to reject all significant principles of distributive justice. He should have distinguished more clearly between the concepts of what we ought morally to do, and of what we have most reason to do. And he makes some claims that are simply false, as when he writes, 'I think that a "plain man", in a modern civilized society, if his conscience were fairly brought to consider the hypothetical question, whether it would be morally right for him to seek his own happiness on any occasion if it involved a certain sacrifice of the greater happiness of some other human being—without any counterbalancing gain to anyone else—would unhesitatingly answer in the negative' (ME 382).

xlii *I assumed that there were never plain mistakes, not ones that mattered anyway*. Charmingly, Rawls adds: 'I always took for granted that the writers we were studying were much smarter than I was. If they were not, why was I wasting my time and the students' time by studying them?' (Rawls (2000) xvi–ii). Since philosophy makes progress, we can now see plain mistakes made by people who were much smarter than us.

xliii *In the Stoic's principle concerning suicide* . . . As one Stoic said, 'If your tent smokes, leave'.

xliv . . . *We should ignore such outbursts.* We should ignore the outbursts of some other great, passionate writers, such as Ruskin's contemptuous remarks about Palladio's Venetian churches. Ruskin calls the Redentore 'a mean, contemptible suburban church'. Discussing San Giorgio, he writes, 'It is impossible to conceive a design more gross, more barbarous, more childish in its conception, more servile in plagiarism, more insipid in result, more contemptible under every point of rational regard' (Ruskin (1903) 381).

*Like Sidgwick, I sometimes find him 'quite a revelation' (*HSM 151). Sidgwick's remark is about Kant's terminology. But he continues: 'we

must go back to Kant and begin again from him. Not that I feel prepared to call myself a Kantian, but I shall always look on him as one of my teachers'.

31 *We may be the only rational beings in the Universe.* If we ask 'Are we?', either answer would be astonishing. There are few such questions.

Facts give us reasons . . . Reasons can be claimed to be provided, not only by facts, but also by things in other categories, such as mental states, or properties. Some people say, for example, that our desires give us reasons, and that an act's wrongness gives us a reason not to do it. But all reasons can be redescribed as being provided by certain facts, such as facts about our desires, or about the wrongness of some act.

32 *. . . what we have most reason to do.* When we claim that we have *more reason* or *most reason* to act in some way, we use the word 'reason', not as a *count noun* — like 'tree', 'lake', and 'cow' — which refers to particular reasons, but as a *mass term* — like 'wood', 'water', and 'beef' — which refers to some reason or set of reasons without distinguishing between these reasons. Similar remarks apply to the claim that we have *sufficient reason* or *decisive reason* to act in some way.

33 *. . . what we should or ought to do*. Like the concept of *a reason*, the concept expressed by these uses of 'should' and 'ought' cannot, I believe, be helpfully defined. Some people suggest that, when we claim that we ought to do something, we mean that we have decisive reasons to do this thing. But this seems to be only part of what we mean, or imply. The word 'ought' seems to add something. Others try to define the concept of a *reason* by appealing to the concept *ought*. I doubt whether such definitions could succeed. But even if these concepts are both indefinable, they are very closely related, in ways that do something to explain them both. We can partly *identify* this version of the concept *should* and *ought* by saying that this concept applies to some act *just when, and because*, we have decisive reasons, or most reason, to act in this way.

35 *. . . if we say that false beliefs can give people reasons, we would need to add that these reasons do not have normative force . . .* Though it is best to claim that practical reasons are given only by facts, it may be on balance better to allow that some kinds of false belief can give people *epistemic* reasons.

We have a different kind of apparent reason when we believe that we have some reason. Such beliefs raise special questions, which I discuss in Section 17.

37 *. . . I shall say little about motivating reasons.* These reasons can be acceptably regarded in two ways. On the psychological account,

motivating reasons are beliefs. On the non-psychological account, these reasons are *what* we believe. When what we believe is true, the non-psychological account is more natural. If I were asked why I don't eat walnuts, I might say 'Because they would kill me'. But if I later learnt that my doctor was mistaken, since walnuts wouldn't kill me, this reply would be misleading, so I would instead say 'I used to avoid walnuts, because I believed that they would kill me.' We might also describe some motivating reason either as what we wanted to achieve, or as our desire or aim. When asked why I don't eat walnuts, I might have said either 'To stay alive' or 'Because I want to stay alive'. (This second claim, as I shall argue, need not imply that my reason is given by my desire.)

We need not choose, I believe, between the psychological and non-psychological accounts of motivating reasons, since they are compatible, and we can use them both. The acceptability of both accounts can, however, cause confusion. On one account, motivating reasons are the real or apparent normative reasons which are *what* we believe when these beliefs explain our decisions and our acts. On the other account, motivating reasons are motivating states of mind. Since motivating reasons can be regarded both as normative reasons and as motivating states, this may suggest that *normative* reasons are motivating states. That would be a great mistake.

44 *Our desires are telic. . or. . instrumental.* We have desires of other kinds, as when we want some medical test to show that we don't have some disease.

45 *Object-given reasons . . . we can call . . . value-based.* This word is slightly misleading if, as I have claimed, these reasons are really based, not on the value of these outcomes or acts, but on the facts that make them good or bad. But some Objectivists reject this claim. I shall use the word 'value-based' partly so that my claims will cover these other versions of Objectivism.

52 *. . . such events would be extrinsically good by making some longer sequence . . . intrinsically better.* On one view, for example, deserved punishment is good. Being deserved would be an *extrinsic* property, since it would depend on whether the person who is punished earlier committed some crime. But on this view, if someone commits some crime and is later punished, this longer sequence of events is *intrinsically* less bad, and in this way better, than this person's committing this crime and never getting the punishment that he or she deserves.

54 *. . . if we didn't dislike these sensations, neither they nor our conscious state would be bad.* The word 'pain' is ambiguous, since it can be used to refer to certain kinds of sensation either (1) only when they are

disliked, or (2) whether or not they are disliked. When psychologists test the pain reactions of various people, by slowly decreasing the intensity of some painful sensation, some of these people say 'It's no longer painful', but others say 'It's still painful, but I no longer dislike it'. I shall use 'pain' only in the first, dislike-implying sense. This is the sense that has normative importance. It is not bad to be having some sensation of a kind that is usually disliked, when we don't in fact dislike this sensation. (See Kahane (2009).)

60 *In some cases, however. . . . Subjectivists ought to deny that this desire gives me a reason.* There is another way to describe some of these cases. Rather than saying that we have no *reason* to fulfil such desires or aims, Subjectivists could sometimes claim that we *could not* fulfil them. If you do not deserve to suffer, my hurting you could not give you any suffering that you deserve, and if you have not injured me, I have nothing to avenge. Such telic desires, Subjectivists might say, should be regarded as implicitly taking a conditional form. What we really want is that something will happen *if* certain facts are as we believe them to be. If these facts are *not* as we believe, such desires could not be fulfilled, and this could be why they provide no reasons for acting.

70 *We can call some normative claim substantive . . .* This sense of 'substantive' contrasts with both 'conceptual' and 'trivial', rather than, as above, with 'procedural'. (Some conceptual claims, I should add, are not trivial.)

88 *There is another reason why . . . any such beginningless chain.* It might be suggested that, rather than forming such a chain, our desires might form a circle, so that we had desire-based reasons to have all of these desires. But we could not have desire-based reasons to have this circle of desires. Suppose, for example, that I want some mud so that I have something to put in a bowl, and I want a bowl so that I can put some mud in it. Though each of these desires might be claimed to give me a desire-based reason to have the other desire, this pair of desires don't give me desire-based reasons to have this pair of desires. My wanting some mud and a bowl doesn't give me a reason to want some mud and a bowl.

We can now return to Case Two . . . This case provides what we can call the *Second Agony Argument* against Subjectivism about Reasons. This argument is in one way stronger than the argument I gave in Section 11. That argument points out that, according to subjective theories, the nature of agony gives us no reason to want to avoid being in agony. This second argument points out that, according to these theories, we might have decisive reasons to cause ourselves to be in agony for its own sake. This implication of these theories is, in one way,

harder to believe. Subjectivists might however claim that, though their theory has implausible implications when applied to this imagined case, our desires do, in most cases, give us reasons. That is at least a partial reply to this Second Argument. But when they consider the First Agony Argument, Subjectivists have no reply. Even if our desires gave us reasons, that could not show that we have no reasons to have our desires. Subjectivists must admit that, on their view, what it feels like to be burnt or whipped gives us no reason to want to avoid being burnt or whipped. We don't have even the slightest reason to believe this claim.

91 *Subjectivists must reject these claims.* There are other objections to these theories. Suppose that my whimsical despot threatens that I shall be tortured unless, at noon tomorrow, I have the aim of being tortured for its own sake. According to aim-based subjective theories, since I now have the aim of *not* being tortured, I now have an aim-based reason to achieve this aim by causing myself to have the aim of *being* tortured. But if I succeed in causing myself to have *this* aim, this would give me an aim-based reason to cause myself to have the aim of *not* being tortured, and this would give me an aim-based reason to cause myself to have the aim of *being* tortured, and so on indefinitely. If I switch back and forth between these aims, I would have a chance of only one in two of having, at noon tomorrow, the aim that would save me from being tortured. On objective theories, there is no such problem, since my reason to obey the despot's threat is given, not by my aim, but by my unchanging reasons to have and try to achieve this aim.

94 *To defend this claim . . .* In some cases, however, these people could make claims of the kind I describe in my note about page 60.

96 *Though the Telic Desire Theory is not incoherent . . . And my other objections apply.* We ought, I have argued, to reject all subjective theories. We can next briefly consider a *hybrid* theory. On this view, for us to have a reason to try to fulfil some desire, we must have some value-based object-given reason to have this desire. What we want must be in some way good, or worth achieving. But when our desires are in this way rational, our having these desires would give us further reasons to try to fulfil these desires. And when we must choose between equally good possible aims, our desires or preferences can break ties, by giving us reasons to adopt one of these aims.

I believe, though not very strongly, that we ought to reject even this hybrid theory. These two kinds of theory are, I believe, like oil and water, which cannot be easily combined. When we have certain desires, this fact may make it true that we have further reasons to try to fulfil these desires. But these further reasons would be provided, not by the

fact that we would be fulfilling these desires, but by various other facts which causally depend on our having these desires. I describe some such facts near the end of Section 9. Though I believe that we should reject this hybrid theory, that may not be shown by my arguments against pure subjective theories. This question would then remain open. But this question would not, I believe, have much importance, since this hybrid theory is fundamentally objective and value-based.

100 *On what seems to be Frankfurt's real view . . . very different from the view that no ends are in themselves good.* This distinction is often overlooked. In Sartre's famous example, a young Frenchman must decide whether to care for his mother or join the Resistance. These activities would both be good, though neither would be clearly better. So this example does not, as Sartre believes, support Sartre's existentialist version of Subjectivism.

101 *According to subjective theories . . . there have been many people whose fully informed desires would not be better fulfilled when any child's life were saved.* As this remark implies, this impartial-reason-implying sense of 'best' has no connection with some 'impartial observer' accounts either of the goodness of outcomes, or of morality. These accounts define what is *best*, or *right*, as what any impartial observer *would in fact* choose, or approve. Such accounts achieve little. If we claim only that this observer has an impartial point of view, we cannot assume that all such observers would make the same choices. If we add certain psychological assumptions, we may be able to work out what such observers would choose. If such observers are benevolent, for example, they would choose what would benefit people. But such predictions would have no importance.

110 *. . . on such views . . . there aren't really any normative reasons. There are merely causes of behaviour.* These writers would reject this description, since they believe that normative reasons *are* certain causes of behaviour. Reductive views are hard to describe in a neutral way.

111 *We can next look more closely . . . not only caused by, but also justified by, my belief.* For such desires to be justified by such beliefs, they must also be caused by these beliefs in the right way. We need not here discuss what this kind of causation involves.

113 *Our having some desire is in one way rational when and because this desire itself is rational.* I am distinguishing here between some desire itself, or some type of desire, and someone's having this desire. If you and I both want Venice to be saved from the rising sea, we have the same desire, but my having this desire is not the same as your having it. In this example, we both want the same event. When we want different events, we may still have what is in a wider sense the same desire. That

would be true, for example, if we are playing a game of chess, and we both want to win. There is a similar distinction between some belief itself and someone's having this belief. The words 'desire' and 'belief' are ambiguous, since they can refer either to some desire or belief itself, or to someone's having this desire or belief. Though I shall sometimes say which I mean, these distinctions can often be ignored.

121 *. . it is not implausible to believe that we have reasons to care more about our nearer future.* This view is not implausible because certain other facts give us reasons to have such a *discount rate*, caring less about events that are more remote. Our beliefs about such events are often less likely to be true. It is often less urgent to try to produce or prevent such more remote events. There are likely to be fewer relevant psychological connections between ourselves as we are now and ourselves in the further future. And earlier benefits often produce other benefits. But these facts do not justify a *pure* time preference, or a discount rate with respect to time itself. (I discuss these questions in RP, Sections 63 to 70.)

128 *. . . the relation better than is transitive.* Some relation is *transitive* when it is true that, if A has this relation to B, and B has this relation to C, A must have this relation to C. Unlike *parent of*, for example, *ancestor of* is transitive.

132 *Two such dissimilar reasons . . . there could not be . . . any precise truth about the relative strength of these reasons.* It is worth noting here an ambiguity in claims about the intensity of pains or pleasures. One sense of 'intense' is purely psychological. When people use this sense, for example, they might claim that

> (A) compared with having one hour of intense pain, it would be less bad to have ten hours of pain that was only half as intense.

We often assume, however that

> (B) the painfulness of some ordeal depends directly on both its intensity and its duration. We have twice as much pain, for example, if our pain is twice as intense, or it lasts for twice as long.

If we assume (B), (A) implies that it would be less bad to have the longer ordeal, though this ordeal would be more painful. On this view, it would often be in itself better if there was much more suffering. These are implausible conclusions. It is better to claim that

> (C) one of two ordeals would be in itself worse when this ordeal would be more painful.

To be able to claim both (B) and (C), we must use 'intense' in a second, *evaluative* sense. We should claim, for example, that

> (D) if it would be equally bad to have one hour of intense pain, or two hours of pain that was less intense, this lesser pain would be half as intense.

And rather than claiming (A), we should claim that

> (E) if ten hours of pain would be less bad than one hour of more intense pain, this second pain would be more than ten times as intense.

As these remarks show, though we sometimes need to use 'intense' in the psychological sense, it is the evaluative sense that is more important. Many writers have made false claims because they overlook this distinction. Similar remarks apply, more obviously, to pleasure. As Sidgwick writes, 'we must be careful not to confound intensity of *pleasure* with intensity of *sensation*, as a pleasant feeling may be strong and absorbing, and yet not so pleasant as another that is more subtle and delicate' (ME 94).

133 *Sidgwick here appeals to the separateness of persons* . . . Given Sidgwick's belief that the distinction between persons is fundamental and of great normative significance, it is surprising that he gave so little weight to principles of distributive justice (ME 416–7).

135 *On Sidgwick's view, we could rationally do what we knew would be only very slightly better for ourselves* . . . Sidgwick does not consider such cases. If he had done so, he might have qualified his view.

137 *wide value-based objective views: When one of our two possible acts* . . . Similar claims apply to other kinds of case. Suppose for example that we could either (1) save some stranger from ten hours of pain, or (2) save ourselves from two hours of pain, or (3) do what would both save this stranger from five hours of pain and save ourselves from one hour of pain. Though (3) would be neither impartially best nor best for ourselves, wide value-based views would imply that, as a compromise, we could rationally do (3).

142 *The Dualism of Duty and Self-Interest* . . . According to what I earlier called Sidgwick's *Dualism of Practical Reason*, we could rationally do either what would be impartially best or what would be best for ourselves. Sidgwick does not distinguish these two versions of his Dualism, because he believes that our duty is always to do what would be impartially best. Sidgwick's other remarks about his Dualism raise several interesting questions, which I hope to discuss elsewhere.

143 *These magnificently sombre claims . . .* Sidgwick also writes: 'When a man passionately refuses to believe that the "Wages of Virtue" can "be dust", it is often less from any private reckoning about his own wages than from a disinterested aversion to a universe so fundamentally irrational that "Good for the Individual" is not ultimately identified with "Universal Good" ' (ME 1st Ed. 471–2).

151 *wrong in the evidence-relative sense . . .* This sense could have two versions, one referring to the evidence of which we are actually aware, the other to the evidence that is available in the sense that we could have made ourselves aware of it. The fact-relative sense is often called the 'objective' sense, but this word has other meanings, and the other senses are not well-called 'subjective'. There is also a *moral-belief-relative* sense, to which I shall return.

157 *This Kantian View . . . I mention in a note.* Suppose first that some man is driving very carefully, but a child runs out in front of this man's car, and is killed. Since this man is not negligent at all, he deserves no blame. But as Nagel points out (1979), we can plausibly believe that this man ought to feel appalled by the fact that he has killed a child, even in this wholly blameless way. Williams calls this response 'agent-regret' (1981, 27). It would be chilling if, because this man knew that he was blameless, he drove away, whistling a happy tune. (The term 'agent-regret', we can note in passing, seems too narrow. Suppose that some other man is blamelessly tripped by some invisible obstacle, and falls in a way that kills some child. We would expect this man to have similar regret for having caused this child's death, though this death was not the result of any *act*.)

Consider next two men who both drive equally carelessly, one of whom kills a child. These men are both blameworthy, and should feel remorse. Many people would believe that, of these men, the one who kills a child is more blameworthy, and should feel more remorse. But those who have this belief may in part be assuming that this man ought also to experience agent-regret, feeling much worse about what he has done. The other man, who has killed no one, has no reason to feel such agent-regret. It is not clear that, as well as feeling such great agent-regret, the man who has killed this child ought also to feel greater remorse.

Consider next Bentham's plausible principle that all punishment is bad, so that we should punish people as little as we can, if the effects would be just as good. Attempted murder has the unusual feature of being a crime that no one could ever intend to commit. We cannot be attempting to kill someone unless we are intending to kill this person. Attempted murder is attempted *murder*. So if there was no punishment for attempted murder, that would not lead anyone to

choose to commit this crime. There should be no such punishment, Bentham might argue, because such punishment would not deter any crime, and would therefore do no good.

It may be objected that, though people who attempt murder cannot *intend* to fail, some of these people believe that they are fairly likely to fail. Punishing attempted murder may have some good effects, by deterring some of these people. We have another reason, though, for punishing attempts less severely than successful murders. If these punishments were the same, that might deter some first attempts. But when people have unsuccessfully attempted murder, they would have more reason to go on trying, until they succeed. That success would not increase these people's punishment, if they were caught and convicted, and would often make it less likely that they would be caught. So we have, on balance, decisive deterrent reasons to punish attempts less severely than successful murders.

In British law, we can add, though most attempted murders are crimes, it is no crime to attempt murder with some method that could not possibly succeed. A similar argument might explain why. If we all knew that such attempts would not be punished, no one would be tempted to act in this way. No one would use a method that they believed could not possibly succeed. Nor would such attempts do any harm. But those who attempt such murders may be, morally, just as blameworthy. Such people might believe, that by sticking pins into a wax dummy of their hated enemy, they are using the method that is most likely to succeed.

Return now to the moral difference between two people who attempt murder, only one of whom succeeds. Thought these people should both feel remorse, the successful murderer ought also to feel great agent-regret of a kind that the attempted murderer has no reason to feel. And the successful murderer ought, for deterrent reasons, to be more severely punished. These two claims may sufficiently describe the moral difference here. It seems doubtful that the successful murderer is more blameworthy, and that, if we believe in desert, this person deserves to be punished more. This view, as I have said, is paradoxical. Given our reasons to believe that blameworthiness and desert cannot depend entirely on luck, it is unclear how these things could even partly depend on luck.

160 *Expectabilism . . . expectably-best.* Rather than talking of the expectable goodness of these outcomes, many people talk of their *expected* goodness. But that word is misleading, since such expectable goodness is often not goodness that either is, or should be, expected.

We can similarly claim . . . the expectable goodness of some act's effects . . . Expectabilists need not assume that the expectable goodness

of outcomes depends only on the expectable *sum of benefits*. As Broome and Kamm suggest, for example, it may also matter how these benefits, or people's chances of getting these benefits, are distributed between different people. In life-saving cases that involve very many people, it might be better if everyone were given equal chances of being saved, even if slightly fewer people would then be saved. And we might have reasons to be risk-averse, giving somewhat greater weight to avoiding the worst outcomes. See 'Fairness' in Broome (1999), and Kamm (1993) Chapter 7.

166 *But Rational Egoism is best regarded, not as a moral view* . . . Though Sidgwick called Egoism one of the 'Methods of Ethics', he was using 'Ethics' in a wider sense, which covers all claims about what we have reason to do.

169 *(K) What we ought-impartially to do* . . . Rather than claiming that we ought to maximize the sum of happiness minus suffering, these Utilitarians might claim that we ought to minimize the sum of suffering minus happiness. These ways of acting are the same, just as minimizing net losses is the same as maximizing net profits. But by telling us to minimize the net sum of suffering, these Utilitarians would remind us of the most effective way of making the lives of sentient beings go better. And this statement of their view better expresses what makes it plausible. On this view's Buddhist version, the two great virtues are insight and compassion.

186 *White would not have sufficient reasons to give up her life so that I could save Grey's leg.* Things might be different if White was old and Grey was a young professional dancer. White's loss might then be no greater than Grey's. This is the kind of morally relevant further fact that, in considering my examples, we should suppose would not obtain.

209 *When applied to these two alternatives* . . . There are, of course, other alternatives. This person would have sufficient reasons to consent to my giving this money to save someone else from some similarly great harm. But the Consent Principle would then still require me to make such a gift.

212 *the Mere Means Principle* . . . Some writers argue that we can ignore Kant's claim that we must never treat people merely as a means, since it is enough to know what Kant means by treating people as ends. If we treat someone as an end, these writers claim, we shall not be treating this person merely as a means. But treating people as ends, Kant claims, consists in part in *not* treating them merely as a mean. So we should ask what that involves.

215 *Since relevance and importance are both matters of degree, it is often unclear whether (1) is true.* As this remark suggests, I am not proposing (B) as a *criterion* that might help us to decide whether someone is

treating someone else merely as a means, or is close to doing that. (B) merely describes how we might plausibly deny that some act is of this kind. No one should object to (B) by claiming that, even if (3) our treatment of someone is governed in sufficiently important ways by some relevant belief or concern, it might still be true that (4) we are treating this person merely as a means. If (4) were true, (3) would not be true, since our treatment of this person would not be governed in *sufficiently* important or relevant ways.

229 *the Harmful Means Principle*. This principle may need to be extended and revised for reasons given in Kamm (2007) Chapters 4 and 5.

241 *On Kant's view, like having a good will . . . That conclusion would be insane*. Though Ross is a pluralist Intuitionist, who is not committed to some rigid theory, he comes close to a similar conclusion. Ross considers and does not firmly reject the view that the least increase in knowledge has more value than any amount of pleasure, or, as his other claims imply, the absence of any amount of pain (Ross (1930) 149–152). It is disheartening that such a good philosopher fails to reject this appalling view. But these are early days.

242 *The word 'humanity' . . . Kant is the least exact of the great thinkers (*Kemp Smith (1962) xx). As one example, we can note how Kant misdescribes his view. Humanity, Kant claims, is an end in itself, which has dignity in the sense of supreme and unconditional value. But Kant also claims that only good wills have such supreme value. These claims do not conflict, Korsgaard suggests, because Kant uses 'humanity' to refer to 'the power of rational choice', and this power is 'fully realized' only in people whose wills are good, since these are the only people whose choices are fully rational (Korsgaard (1996) 123–4). This suggestion has some plausibility. But Kant also uses 'humanity' to refer to rational beings, which he claims to be ends-in-themselves, with supreme value. We could not similarly claim that rational beings are the same as good wills. Nor could we claim that such beings are the same as the Realm of Ends, or the Greatest Good: the world of universal virtue and deserved happiness. Though Kant claims that only good wills have dignity, we should admit that, on Kant's view, there are several kinds of thing that have such supreme or unsurpassed value. (See Dean (2006).)

247 *when we say that . . . this act is right, we mean that this act would do the most good*. In Moore's words ' "right" . . . does and can mean nothing but "cause of a good result" (Moore (1903) 196). Moore must mean 'cause of the best result'. Characteristically, Moore adds, 'it is important to insist that this fundamental point is demonstrably certain'. When Moore's clouds, for many decades, hid the light from Sidgwick's

sun, that was in part because, unlike the judicious Sidgwick, Moore writes with the extremism that makes Kant's texts so compelling. With the exception of the 'doctrine of organic unities', every interesting claim in Moore's *Principia* is either taken from Sidgwick or, like the claim just quoted, obviously false. As Williams writes: 'Moore's philosophy is marked by an affectation of modest caution, which clogged his prose with qualifications but rarely restrained him from wild error' (Williams (1985) 16). (These remarks are overstatements of the Moorean kind.)

Moore's Principia does not put forward a substantive moral view. It is surprising that Moore makes this mistake, since he devotes an entire chapter to condemning such mistakes, which he calls 'the Naturalistic Fallacy' (though it is neither naturalistic nor a fallacy). Sidgwick more accurately describes this mistake in two sentences (ME 26 note 1, and 109).

275 *The Impossibility Formula* . . . G 424. Kant also writes 'Some actions are so constituted that their maxim cannot even be thought without contradiction as a universal law . . . ' Following O'Neill, several writers call this formula the 'contradiction-in-conception test'. When we have decided what it would be for some maxim to *be* a universal law in Kant's intended sense, we may find that it would be logically impossible, and in this way a contradiction, to suppose that certain maxims are such laws. But Kant often appeals to merely empirical impossibilities, so we should do the same.

. . . *(A) would not be a useful claim.* It is too unclear in what sense it could not *be* true that we are all permitted to act on some maxim. This could not be true if such acts are wrong, but that does not help us to decide *whether* such acts are wrong. Kant also claims that it is wrong to act on any maxim if we could not *will* it to be true that we are all permitted to act upon it. That is a more plausible claim, to which I shall return.

279 . . . *the social practice of morally motivated, trust-involving promises would have ceased to exist.* We can add that, if we all believed that such lying promises were permissible, we would not even understand the concept of a moral, trust-involving promise. (There might still be a practice that was like the practice of promising, except that it took a non-moral form. Such promises would be like threats. Just as we could have reasons to fulfil our threats to preserve our reputation as a threat-fulfiller, we could have reasons to keep such promises to preserve our reputation as a promise-keeper.)

281 *As this and other such cases show, (G) is unacceptable.* These imagined cases might be claimed to be unrealistic, because in the real world the facts would not have been as simple as I have asked us to suppose.

But these cases are plausible enough to provide good tests of the acceptability of (G). Moral principles ought to succeed when applied to somewhat simplified imagined cases of this kind. And Kant's claims about his imagined lying promise, and about several other maxims, are similarly simplified.

286 *Given their similarity . . . as I explain in a note . . .* Suppose that we appealed only to the Permissibility Formula. We would then ask whether we could rationally will it to be true that everyone is morally permitted to act on some maxim, even though this would make no difference to anyone's moral beliefs, or to anyone's acts. This would not be a helpful question. First, it is hard to imagine that we could *will* it to be true that certain acts are morally permitted. In the case of most wrong acts, as Kant and many other theists claim, not even God could have willed it to be true that these acts are *not* wrong. And if the fact that certain acts are morally permitted would make no difference to what anyone believes or does, it is unclear what reasons we could have for willing that these acts be permitted, other than the fact that these acts really *are* permitted. But the Permissibility Formula could not then help us to decide *whether* such acts are permitted.

289 *Suppose that I wrongly steal . . . Kant's formulas mistakenly permit my act.* I am here assuming that, unlike Kant's Consent Principle, Kant's Formula of Universal Law is intended to be the only moral principle we need, so that when both versions of Kant's formula fail to condemn this act, this formula implies that this act is morally permitted.

292 *For examples of a different kind . . . nor could he have rationally willed . . . that everyone believes these life-saving lies to be wrong.* If Kant accepted the Whole Scheme View, as I suggest on page 255, it might not have been irrational for him to will that no one ever tells a lie. But the Whole Scheme View is false, and when we apply Kant's formula, we should ask what people could rationally will if they knew the relevant non-moral facts and had no false beliefs.

298 *When it is unclear . . . But we need not answer these questions here . . .* Such questions will also have less importance when we have revised Kant's formulas in other ways.

301 *When we apply . . . On the best version of Kant's formula, which seems to be what Kant has in mind . . .* As is suggested by Kant's remarks about his self-reliant man whose maxim is 'Don't help others, but don't cheat them either' (G 423). Kant claims that, if everyone acted on this man's maxim, such a world would be *better* than the actual world, in which some people help others but many people cheat. But Kant also claims that we could not rationally will it to be true that everyone acts on this man's maxim. So Kant's implied comparison cannot be with

the actual world, and seems to be with a world in which no one acts on this maxim.

302 *We can call such cases each-we dilemmas.* I discuss such cases in RP Part One and Parfit (1986).

303 *. . . we are doing what is certain to be worse for all of us.* There are also many *probabilistic* each-we dilemmas, which appeal to the likely effects of different acts, or to what would be expectably-best for people.

These claims are not about . . . repeated prisoner's dilemmas. In the simplest cases of this kind, (1) each of us can often either benefit himself or give a greater benefit to others, and (2) because the number of people involved is fairly small, what each person does *may affect* what, in later situations, other people do. In a two person-case, for example, if I give you the greater benefit, you may reciprocate, and give me the greater benefit. If I then switch to giving myself the lesser benefit, you may retaliate, and give yourself the lesser benefit, which will be worse for me. Though these are called 'repeated prisoner's dilemmas', they do not involve even one *true* prisoner's dilemma, or each-we dilemma. In such cases, it is not true that, if each rather than no one does what is *certain* to be better for himself, that would be worse for all of us. These cases are theoretically much less interesting, and less fundamental, since they are merely one of the many kinds of case in which it is unclear which way of acting would be best for ourselves. Such cases are also practically much less important, since they are much less common. They are, however, important to evolutionary psychologists who are trying to explain various features of animal behaviour and human psychology, and to historians who are discussing the small communities in which, in earlier centuries, most people lived.

Though each-we dilemmas are often overlooked . . . there are many cases that involve many people. It is worth mentioning another kind of case, which we can call *Samaritan's dilemmas*. Each of us can sometimes help some needy stranger, at some small but real cost or burden to ourselves. That might be true, for example, when we could help someone who has had an accident, or we could return lost property of great personal value. If all of us always gave such help to strangers, that might be better for all of us than if none of us ever gave such help. But if we live in large cities, as is now true of more than half of the world's population, it might also be better for each person if he never gave such help. This person would avoid the costs to himself. And whether he received such help would very seldom depend on whether he gave such help to others. The strangers whom each of us failed to help would hardly ever be the same people as the strangers by whom we might later be helped. So our failure to help others would hardly ever lead others, bearing a

grudge, to deny us help. But if no one helps others, though *each* of us would be doing what would be better for himself, *we* would be doing what would be worse for all of us.

These involve public goods . . . There is a further distinction between those goods which in fact benefit even those people who do not help to produce them, and those which are bound to do that, since there is no feasible way to prevent non-contributors from getting these benefits. Clean water is often in the first category, and clean air in the second.

306 *As well as conflicting . . . such principles . . . are directly self-defeating at the collective level.* There is also a way in which, in such cases, common sense morality might itself imply that we ought to cease to give priority to our M-related people. If we and the other members of the relevant group could all communicate, and we all knew each other to be trustworthy, we would all be rationally and morally required to make a joint conditional promise that we shall always act differently, by giving the greater benefits to others. If this joint promise would become binding only if everyone makes it, this fact would, when we are deciding whether to make this promise, *tie our acts together*. In making such a promise, each of us would be doing what would be best for himself or his M-related people, since he would be helping to bring it about that everyone rather than no one did what would be better for him or for his M-related people. Since this promise requires unanimity, each person would know that, if he did not make this promise, the whole scheme would fail. That is how our acts would be tied together. So common sense morality would itself tell us all both to make and to keep this promise. This solution, however, could seldom be achieved, since we are not all trustworthy, and, even if we were, it would often be too difficult to arrange and achieve such a joint conditional agreement. If we were all sufficiently conscientious Kantians, we would avoid this problem. (For a fuller discussion, see RP 100–108.)

307 *Kant must mean . . . whatever would, on the whole, best promote everyone's happiness.* In a different way, however, this solution may be *indirectly* collectively self-defeating. See page 404–5 below and RP sections 10 and 42.

These claims are true . . . and our children's children. We might, however, draw a distinction here. It is clear that, in each-we dilemmas, what we *should all ideally do* is to give the greater benefits to others. If all rather than none of us acted in these ways, that would be better for everyone. But Kant's formula requires such acts even when most other people are *not* acting in these ways. In such cases, by acting in these ways, we would lose the lesser benefits that we could give ourselves without receiving the greater benefits from others. This requirement

may sometimes be too demanding. It might also be unfair. In unsolved Parent's Dilemmas, for example, it may be unfair to our children if we give the greater benefits to other people's children, when other people are not giving such greater benefits to our children. In at least some of these cases, we might justifiably believe that it makes a moral difference how many other people are doing what we should all ideally do. We might be required to give the greater benefits to others only when *enough* other people are acting in this way. In other cases, we might be permitted, as a defensive second-best, to give the lesser benefits to ourselves, our children, or our other M-related people. (For a suggestion about what would be enough, see RP 100-1).

311 *We have not yet fully answered . . . it was permissible for him to abstain.* Curiously, as a biographer of Kant reports: 'Kant formulated the maxim: "One mustn't get married". In fact, whenever Kant wanted to indicate that a certain, very rare, exception to a maxim might be acceptable, he would say: "The rule stands: 'One shouldn't marry! But let's make an exception for this worthy pair'" (Kuehn (2001) 169).

319 *Some of these rules could take such conditional forms . . . Similar claims would apply to those versions of RC which appeal to what would happen if people accepted certain rules.* My proposed revision applies more easily to these *acceptance-versions* of Rule Consequentialism, because the optimific rules would take much simpler forms. (As Ridge points out, even if such rules took conditional forms, there may be no set of rules whose acceptance would make things go best at *each* level of acceptance. But there would be sets of rules whose acceptance at different levels would, on average, or on balance, make things go best. Ridge (2006).)

320 *. . . we* can *plausibly assume that everyone ought to have the same moral beliefs.* If people have conflicting beliefs, for example, these beliefs cannot all be true, and we can assume that everyone ought to have, or try to have, true moral beliefs . . . *it is not enough* simply *to reply 'Most people won't'*. But see RP ch. 1

321 *This rule expresses . . . three of the world's earliest civilizations.* These are the ancient Near East, India, and China. Here are some quotations:

Buddhism: Hurt not others in ways that you yourself would find hurtful (*Udana-Varga* 5:18).

Christianity: Whatever you would will that men do to you, do you to them; for this is the law and the prophets (*Matthew 7*:12).

Confucianism: Do not do to others what you do not want them to do to you. Then there will be no resentment against you, either in the family or in the state (*Analects* 15:23). Tse-kung asked, 'Is there one word that

can serve as a principle of conduct for life?' Confucius replied, 'It is the word 'shu'—reciprocity. Do not impose on others what you yourself do not desire.' " (*Doctrine of the Mean* 13.3)

Hinduism: This is the sum of duty; Do nothing to others that you would not have them do to you (*Mahabharata* 5:1517.)

Islam: None of you is a believer until he wants for his brother that which he wants for himself (*Sunnah*). (Number 13 of *Imam Al-Nawawi's Forty Hadiths*.)

Jainism: A man should treat all creatures as he himself would be treated (*Sutrakritanga* 1.11.)

Judaism: What is hateful to you, do not do to your fellowman. This is the entire Law; all the rest is commentary (*Talmud, Shabbat* 31a).

Shinto: "The heart of the person before you is a mirror. See there your own form" (*Munetada Kurozumi*).

Taoism: Regard your neighbor's gain as your gain, and your neighbor's loss as your own loss (*Tai Shang Kan Yin P'ien*).

Zoroastrianism: That nature alone is good which refrains from doing another whatsoever is not good for itself (*Dadistan-i-dinik*, 94:5).

327 *The word 'everyone' . . . we are among the people whose well-being we ought to consider in the impartial way that this rule requires.* Kant similarly writes: 'since all others with the exception of myself would not be all, so that the maxim would not have within it the universality of a law . . . the law making benevolence a duty will include myself, as an object of benevolence, in the command of practical reason' (MM 450).

338 *None of Kant's claims . . . support Nagel's interpretation.* Kant does write 'every rational being . . . must always take his maxims from the point of view of himself, and likewise every rational being' (G 438). But this remark comes in Kant's discussion, not of his Formula of Universal Law, but of his Formula of the Realm of Ends. And if Kant had intended that we should imagine others doing to us what we do to them, he would not have so contemptuously dismissed the Golden Rule.

341 *As I explain in a note, it is enough to appeal to MB5.* If these formulas sometimes had conflicting implications, we would have to choose between them. These formulas might conflict when (1) we could not rationally will it to be true that everyone acts in some way, but (2) we *could* rationally will it to be true that everyone believes such acts to be morally permitted, because we know that, if everyone had these beliefs, there wouldn't be too many people who would choose to act in this way. If these formulas did conflict when applied to such cases, it would be MB5 whose implications were more plausible. To avoid such conflicts, we might move from LN5 to

> LN6: It is wrong to act in some way unless everyone could rationally will it to be true that everyone acts in this way, when they know that there won't be too many other people who would choose to act in this way.

But this formula is too similar to MB5 for it to be worth discussing both formulas. And MB5 is, I believe, both closer to Kant's view, and clearly better. LN6 is too simple, since it makes a difference *why* there won't be too many people who would choose to act in some way. It makes a difference, for example, whether some people are refraining from acting in some way because they believe that, given the number of people who are already acting in this way, further acts of this kind would be wrong. When that is true, those who act in this way may be unfairly benefiting from the conscientious self-restraint of others. Rather than including such details in our descriptions of how people are acting, as the Law of Nature Formula requires, we do better to include such details in the content of the beliefs to which the Moral Belief Formula refers. This formula is also in itself more plausible. As I have said, while it is often irrelevant to ask 'What if everyone did that?', it is always relevant to ask 'What if everyone thought like you?'

342 *the Kantian Contractualist Formula* . . . In restating MB5 as the Kantian Contractualist Formula, we turn from claims about people's *moral beliefs* to the *principles* that people accept. These principles can be more like the maxims to which Kant appealed. And this change has the advantage that, because we can regard these principles as like *rules* or *policies*, we can more plausibly assess these principles by appealing, in part, to the effects of their being universally accepted. If we assessed moral beliefs in this way, we would be implausibly ignoring the question whether these beliefs are true. Rules and policies can't be true or false. Kantian Contractualists can be Cognitivists, however, since the Kantian Formula states a belief which may be true.

346 . . . *Gauthier rejects appeals to such intuitive beliefs.* Gauthier also argues that, if we accept his Contractualist theory and his minimal version of morality, he can show that, even in self-interested terms, it cannot be rational to act wrongly. No other moral theory, Gauthier claims, can achieve this aim. I discuss this argument in Appendix B.

When applied to morality . . . very appealing and well supported by some of Rawls's Non-Contractualist claims and arguments. One example are Rawls's forceful claims about the arbitrariness of the natural lottery. In this assessment of Rawls, I am following Nagel (1973) and Barry (1989), (1995).

353 . . . *the Maximin Argument has implications that are much too extreme*. Even when applied to the basic structure of society, the Maximin Argument has such implications. Rawls sometimes defines the worst off group in broad terms, so that this group includes many people who are better off than some other people. On one suggestion, for example, the worst off people are those whose income is below the average income of unskilled workers (Rawls (1971) 98, RE, 84.) But if the Maximin Argument were sound, it would require a much narrower definition of this group. This argument implies that each person should try to make her own worst possible outcome as good as possible. On Rawls's suggested broader definitions, we ought to choose policies that would make the *representative* or *average* member of the worst off group better off, even when that would be *worse* for the worst off people in this group. That is precisely the policy that, when applied to society as a whole, Rawls condemns. When defending his broad definitions of the worst-off group, Rawls writes: 'we are entitled at some point to plead practical considerations, for sooner or later the capacity of philosophical or other arguments to make finer discriminations must run out' (84). But there is no difficulty in describing the worst off group as those who are equally worst-off, since these people are not better off than anyone else.

354 *Rawls merely denies . . . there is no way in which non-Utilitarian considerations could possibly enter in*. In his last book, Rawls expresses doubts about his stipulation that, behind the veil of ignorance, we would 'have no basis for estimating probabilities'. He writes 'Eventually more must be said to justify this stipulation' (Rawls (2001) 106). But nothing more is said.

Rawls adds some other stipulations which allow him to put less weight on his claims about probabilities. He tells us to suppose that, by choosing his principles of justice, we would guarantee for ourselves a level of well-being that would be 'satisfactory', so that we would 'care little' about reaching an even higher level. We should also suppose that, if we chose any other principles, we would risk being much worse off. On these assumptions, Rawls argues, it would be rational for us to choose his principles of justice. Rawls then considers the objection that, by adding these assumptions, he makes his theory coincide with one version of rule Utilitarianism, since his principles would be the ones whose acceptance would make the average person as well off as possible. Rawls replies that, on his definition, Rule Utilitarians are not Utilitarians (Rawls (1971) 181–2 and note 31, RE 158–9 note 32. This reply is disappointing. Rawls earlier described his aim as being to provide an alternative to all forms of Utilitarianism. We do not provide an alternative to some view if we accept this view, but give it a different name.

Though I have argued that Rawls's *Contractualism* fails, I should repeat that Rawls's *Theory* is a wonderful book.

360 . . . *that no one could reasonably reject.* We should not assume that, when two people disagree, at least one of these people must be being unreasonable. There can be reasonable mistakes. But if neither of two people is being unreasonable in rejecting the other's principle, there may be no relevant principle that could not be reasonably rejected, with the result that Scanlon's Formula would fail. So when Scanlon claims that no one could reasonably reject some principle, he should sometimes be taken to mean that anyone who rejected this principle would be making a moral mistake, by failing to recognize or give enough weight to other people's moral claims, even if this might be a not unreasonable mistake. Scanlon would here be using 'reasonably' in a somewhat narrower sense. It is often clear, however, whether in the ordinary sense anyone could reasonably reject some principle.

363 *We can plausibly defend this belief . . . this knowledge would make many of us anxious.* This anxiety might not be rational, but that does not undermine these claims . . . *the Anxiety and Mistrust Argument.* In giving this argument, I am ignoring one feature of Scanlon's view. Scanlon claims that, in rejecting principles, we cannot appeal to the benefits or burdens that groups of people would *together* bear. If we follow this *Individualist Restriction*, we cannot oppose the Act Utilitarian view about *Transplant* by appealing to the Anxiety and Mistrust Argument, since this argument appeals to bad effects on many people. Scanlon ought, I believe, to drop the Individualist Restriction, as I argue in Chapter 21.

368 *There is one straightforward and wholly satisfactory defence . . . the properties or facts that in this sense can make acts wrong.* I discuss this distinction further in Section 87.

374 *As I explain in a note, the rightness or wrongness of our acts cannot depend . . .* It could not be true both that

> certain acts are wrong because it would be bad if we acted in these ways,

and that

> it would be bad if we acted in these ways because such acts are wrong.

Wrong acts must have some other feature that makes them either bad or wrong. Nor could it be true that

> certain acts are wrong because such acts are disallowed by the best principles,

> such acts are disallowed by these principles because it would be worse if we acted in these ways,

and that

> it would be worse if we acted in these ways because such acts are wrong.

Just as Contractualists must claim that, when we apply their formulas, we should not appeal to the *deontic reasons* that might be provided by the wrongness of certain acts, Consequentialists must claim that, when we apply their principles, we should not appeal to the *deontic goodness* or *badness* of right or wrong acts. When Consequentialists make claims about how the rightness of our acts depends on facts about what would be best, these claims should use the word 'best' in what we can call its *deontic-value-ignoring* sense.

Similar claims apply to Non-Consequentialists. We reject Act Consequentialism if we believe that

> (A) certain acts are wrong even when it would be better if people acted in these ways.

To illustrate (A), we might claim that

> (B) it is often wrong to lie, steal, or break promises, even when these acts would do more good.

If we believe (B), would we also believe that it would be *better* if people acted wrongly in these ways? If we believe that wrong acts are in themselves bad, the answer may be No. But (B) would not then illustrate (A), since we would not believe that such acts are wrong even though it would be better if people acted in these ways. So if we reject Act Consequentialism by making claims like (A), we may need to use the word 'better' in its deontic-value-ignoring sense.

If acts can be deontically good or bad, as some Consequentialists believe, we may object that Consequentialist theories should not tell us to ignore the value of such acts. But like the Deontic Beliefs Restriction, this *Deontic Values Restriction* applies to only some of our moral thinking. Consequentialists make various claims about how the rightness of our acts depends on how it would be best for things to go. It is only *while* we apply these claims that we should not appeal to our beliefs about deontic values. At other times we can appeal to claims about what is good or bad in the ordinary, unrestricted sense. In what follows in my text, I shall sometimes use such words in their deontic-value-ignoring senses; but in most cases this distinction makes no difference.

375 *On this view . . . There could be many other forms of Indirect Consequentialism.* On one version of Motive Consequentialism, for example, the best motives for each person to have are the motives whose being had by *this* person would make things go best. What I call 'Act Consequentialism' is Direct Consequentialism applied to acts. But there could also be Act Consequentialists who were *Indirect* Consequentialists, because these people applied the Consequentialist Criterion directly to acts, but only indirectly to other things, such as rules or motives. On this view, though the best or right acts are the ones that would make things go best, the best rules are not the rules whose acceptance would make things go best, but the rule 'Always do what would make things go best', and the best motives would not be the motives whose being had would make things go best, but the motive of always trying to do what would make things go best. (These various possibilities are very well discussed in Kagan (2000) and (1998) Chs 6–7.)

393 *Of those who believe it to be wrong. . we ought to regard this fact as, in a sober way, good news.* It is sometimes claimed that we could not have impartial reasons to want anyone to act wrongly. But that is not so. Our claim should at most be that we always have impartial reasons to want *no one* to act wrongly. We might claim that, in *Lesser Evil*, it would be best if no one killed anyone as a means, even though the five would then die. But we are supposing that at least one person will act wrongly in this way. Though it would be bad if you acted wrongly, by killing me as a means, it would clearly be even worse if Grey and Green both acted wrongly, by each killing two other people as a means. If these are the only possibilities, most other people would have more reason to hope that you will act wrongly, since that would be the lesser of two evils. Fewer people would then be wrongly killed as a means. So if most other people learn that you have acted wrongly, thereby preventing the wrong acts of Grey and Green, these people should welcome this news, believing that things have gone better. This view also implies that *you* would have impartial reasons both to want yourself to act wrongly, and to act wrongly, in this way. But these impartial reasons, we could coherently believe, would be decisively outweighed by your other, *person-relative* deontic reasons *not* to act wrongly. If that were true, you would have decisive reasons, all things considered, *not* to do what you have these impartial reasons to do, and what we all have impartial reasons to want you to do.

399 *Compared with (E), this premise . . . So, as (G) claims . . .* I discuss some possible exceptions in Section 81.

400 *When combined with (H) . . . I defend these claims further in a note.* In *Essays on Derek Parfit's On What Matters (*ed. Suikannen (2009))

several people present objections to an earlier version of this argument, and to some of my other claims.

In an elegant and ingenious paper, Gideon Rosen presents various objections to the Kantian Formula. When we apply this formula, we ask which are the principles whose being universally accepted everyone could rationally will, or choose. This formula would fail, Rosen argues, if we knew that there was some evil demon, or malicious gremlin, who would cause great suffering if any principle were universally accepted. This objection is not weakened, Rosen claims, by the fantastic nature of this imagined case, since moral theories ought to apply to all possible cases.

Though this last claim is plausible, it does not apply to the *thought-experiments* to which, in their accounts of moral reasoning, some kinds of moral theory appeal. We cannot claim that any such theory ought to appeal to all possible thought-experiments. Good theories may appeal to only one such thought-experiment. To answer Rosen's objection, I suggested, we can revise the Kantian Formula. This formula can appeal to the principles that everyone could rationally choose, if each person knew that there was no malicious gremlin who would cause great suffering if any principle were universally accepted. When Rosen considers this revision, he objects that it would make the Kantian Formula less plausible. The opposite, I believe, is true. This formula would be less plausible if it allowed us to suppose that there *was* such a malicious gremlin. Similar remarks apply, I believe, to Rosen's other, similar objections.

Rosen then claims that, for some moral principle to be *supreme*, this principle must not only tell us which acts are wrong, but also describe the most fundamental property that makes these acts wrong. Kantian Contractualism, Rosen argues, does not achieve this second aim. I agree. On my account, Kantian Contractualism can at most claim to describe a higher-level wrong-making property under which all other wrong-making properties could be subsumed, or gathered. These other properties are often more important. As I wrote, 'Some acts are open to objections that are both clearer and stronger than the objections to these acts that are provided by Kant's formulas, or by any version of Contractualism, or Rule Consequentialism' (now on page 414).

Jacob Ross presents some objections to the Kantian Argument for Rule Consequentialism. In stating this argument, I had already been greatly helped by some excellent objections that Ross earlier sent me. In this published paper, Ross first points out that some optimific principles could have slightly different versions whose acceptance would be much better for different people. Suppose, for example, that we could save the lives of either of two equally large groups of people. One optimific

principle might tell us to save the people who are nearer to us, and another principle might tell us to save the people who are further away. On these assumptions, there would be no optimific principle whose acceptance could be rationally chosen by the people in both these groups. This objection is like those raised by High Stakes Egoism, and could be answered in similar ways.

Ross then imagines that one of two principles is (1) significantly non-optimific, because this principle's acceptance would make things go much worse, but that (2) this principle's acceptance would be much better for everyone who is or will be actual. On these assumptions, everyone could rationally choose this non-optimific principle, so the Kantian Formula would not here require us to be Rule Consequentialists. In one of Ross's examples, one of two principles would permit us to use cheap energy in a way that would greatly lower the quality of life in the further future. If we followed this principle, that would benefit all presently existing people by producing, in the short term, 'a much stronger global economy'. Our acts might also benefit all future people, since these people might both have lives that are worth living, and owe their existence to our acts. It might be true that, if we had followed the other, optimific principle, it would have been *different* future people who would have later lived, and had a much higher quality of life.

Ross's claims about this case could not all, I believe, be true. We would all have impartial reasons not to choose any principle whose acceptance would greatly lower the future quality of life, thereby making things go much worse. And we cannot plausibly suppose that everyone's impartial reasons would be sufficiently matched by personal reasons to choose this principle, provided by the effects of a stronger global economy. Many people's well-being does not so heavily depend on their income. I would happily give up most of my income if I could thereby bring it about that people would act in ways that greatly raised the quality of life of future generations. I could not rationally choose some other principle, thereby making things go much worse, merely for the sake of the benefits to me of having more consumer goods. Similar claims would apply, I believe, to many other people.

It is, I agree, conceivable that some principle's acceptance would both make things go, on the whole, much worse, but also be much better for all of the people who are or will be actual. That might be true, for example, if this principle's acceptance would give every presently existing person a thousand years of happy, youthful life, but would also end human history, since this longevity can only be achieved in some way that would make everyone infertile. It might then be true that every presently existing person could rationally accept this principle, despite knowing that things would later go much worse, because there

would be no future people. On these assumptions, the Kantian Formula might not require us to give up these great personal benefits, for the sake of enabling humanity to survive. The Kantian Argument for Rule Consequentialism would then need to be in one way qualified, since its conclusion would not apply to this kind of case. This argument could at most show that, in nearly all actual cases, the Kantian Formula requires us to follow optimific principles.

This qualification would, I have claimed, make little difference. Ross writes that, if I could answer the objections provided by his imagined cases, that would considerably strengthen my argument that Kantian Contractualism supports Rule Consequentialism (153). That is not, I believe, true. If my argument showed that, in nearly all actual cases, Kantian Contractualism implies Rule Consequentialism, this argument would be nearly as strong as it could possibly be. My aim is to show that there is no deep disagreement between Kantian Contractualism and Rule Consequentialism. That would be true if, in nearly all actual cases, these two kinds of theory have the same or similar implications.

In an earlier draft of his article in this volume, Michael Otsuka presented some forceful objections to some of my earlier claims about my Kantian Argument. These objections led me to drop these claims, and to add parts of Sections 59–61 and Appendix C. Since I have dropped most of the claims that Otsuka earlier criticized, I need not discuss these criticisms, and shall merely thank Otsuka for the great help that these criticisms gave me.

In his published article, Otsuka makes some remarks about claims that I have kept. According to my premise (D), it is the optimific principles that everyone would have the strongest impartial reasons to choose. Otsuka claims that, for (D) to be true, I must revise my account of what would make some principle optimific. In describing how it would be best for things to go, I must take into account the kinds of value that are to be respected rather than promoted (62–7). But this proposed revision would not, I believe, be needed. When there is some value that is to be respected, as is true, for example, of respect for persons, acts that respect this value would have the kind of value that is to be promoted. It would be better if we acted in these ways.

Otsuka also claims that my argument 'threatens to swallow up moral theories that have been traditionally regarded as non-consequentialist' (68). One example is Kamm's view that things would go best if everyone accepted certain deontological prohibitions. On my account, if Kamm's view were true, these deontological principles would be optimific, and would be accepted by Rule Consequentialists. This fact would be no objection to my argument. These deontological principles are not *Act* Consequentialist, since they claim that certain acts would be wrong

even if we knew that these acts would make things go best. But such principles could be *Rule* Consequentialist. Such theories tell us to follow the optimific rules even if we know that we are thereby failing to do what would make things go best. Rule Consequentialists could accept such optimific deontological prohibitions. (See also Section 67.) Otsuka describes some other ways in which, if my argument is sound, what are widely held to be conflicting views would not in fact conflict. As before, that is no objection, since what I am trying to show is, in part, that apparently conflicting views do not in fact conflict.

Michael Ridge proposes a slightly revised version of my first suggested answer to the New Ideal World Objection. I have gratefully adopted this revision. Ridge then claims that, if I had stated the Kantian Contractualist Formula in a way that included my first answer, my Kantian Argument for Rule Consequentialism would have been open to certain objections. I agree. That is one reason why I did not state my argument in this way, and wrote that I would set aside, for later discussion, the complications that are raised by the New Ideal World Objection.

Ridge also asks whether, if we revise the Kantian Formula in the way that he suggests, my Kantian Argument might succeed. Disappointingly, Ridge does not discuss this question, but asks instead what is implied by a version of Rawlsian Contractualism which requires us to maximize our expected utility. So I do not yet know whether my Kantian Argument could be successfully revised in Ridge's proposed way. Ridge calls it 'peculiar' that I have not yet tried to answer this question (84). But the complications raised by the New Ideal World Objection could not, I believe, undermine any of my claims about the relations between Kantian Contractualism and Rule Consequentialism.

Seiriol Morgan's impressive paper makes some puzzling claims about my proposed revisions of Kant's Formula of Universal Law. Morgan accepts my proposal that we should replace Kant's reference to the agent's maxim with the morally relevant description of the agent's act. He even calls this revision 'wholly Kantian' (45). I also proposed that, rather than referring to what the agent could rationally will, Kant's formula should refer to what everyone could rationally will. This revision, Morgan writes, makes no difference, since 'what is practically rational will . . . be the same for every rational deliberator' (57). Though Morgan believes this revision to be unnecessary, he cannot object to my restating Kant's formula in this second way. A revision that makes no difference cannot be unacceptable. If we combine these revisions, we reach what I call the Kantian Contractualist Formula. To my surprise, Morgan writes that no genuinely Kantian agent could possibly endorse this formula, 'since to do so would be to set herself up for a contradiction in her will' (57). Since Morgan has no objection to my

proposed revisions, Morgan's remark implies that no genuine Kantian could possibly endorse this acceptable revision of Kant's formula. That could not be true.

Morgan's real objection, I assume, is to my claim that this Kantian Formula supports Rule Consequentialism. I go astray, Morgan believes, not by revising Kant's formula, but by appealing to a non-Kantian view about rationality and reasons. Rather than appealing to claims about what we have *sufficient reasons* to will, Morgan writes, we should appeal to claims about what we *could without contradiction* will (47).

My Kantian Argument could take this form. Of the principles that everyone might accept, some would be

> *Kantianly-optimific* in the sense that these are the principles whose universal acceptance would make things go in ways in which we could, without contradiction, will things to go.

My argument could become:

> (A) We ought to follow the principles whose universal acceptance we could without contradiction will.
>
> (B) There are some principles that are Kantianly-optimific.
>
> (C) We could without contradiction will that everyone accepts these principles.
>
> (D) There are no other significantly non-optimific principles whose universal acceptance we could without contradiction will.

Therefore

> We ought to follow the principles that are Kantianly-optimific.

Morgan has no objection to the Kantian Formula stated by (A). Nor, I believe, could he deny either (B) or (C). It is clear that, on Kantian assumptions, there are some principles whose universal acceptance we could without contradiction will. These claims together imply that this Kantian Formula permits us to be Kantian Rule Consequentialists, who follow these Kantianly-optimific principles. So, even on Morgan's assumptions about rationality, this Kantian Formula supports one form of Rule Consequentialism. We need not ask whether, because (D) is also true, this formula also requires us to follow these optimific principles.

It would be a catastrophe for Kantian ethics, Morgan suggests, if Kant's formulas implied some form of Consequentialism. When Morgan makes this claim, he assumes that Consequentialist theories must give supreme weight to people's well-being. Morgan writes: 'What

Kant was attempting to articulate was a moral philosophy which holds freedom to be the most important value, and certainly a value more important than happiness. Parfit's 'Kantian Consequentialism' isn't doing this at all. Rather, his moral theory appears designed to promote another value, a value which we might call "well-being impartialism". But this is a value that is inimical to Kant when elevated to supreme value status, as Parfit intends it to be' (59).

These remarks involve, I believe, some misunderstandings. My argument appeals, not to Kantian Consequentialism, but to Kantian *Contractualism*. And my Kantian Argument for Rule Consequentialism does not give supreme value to well-being. If we believe that freedom is much more important than happiness, that will affect our view about how we could rationally, or without contradiction, will things to go. We could not will that everyone accepts principles that would lead them to promote well-being in ways that, in Morgan's phrase, would 'thwart individual freedom and agency for the benefit of others' (57). Rather than denying that the Kantian Formula supports these Kantian-optimific principles, Morgan's claim should be only that these principles would embody a 'Kantian conception of value' that is very different from what Morgan calls 'well-being impartialism'.

Morgan also writes that, if Kant had been shown that his formulas support some form of Consequentialism, Kant would have considered this 'a devastating result, one that would reveal his whole outlook in moral philosophy to be in disarray. So an implication of the success of Parfit's argument is that a major philosopher was entirely confused about the nature of and implications of his own philosophy, because actually his key ideas support not his own ethical outlook but those of the kind of people he was most concerned to resist' (42).

These claims are, I believe, mistaken. Kant writes: 'Everyone ought always to strive to promote a world of universal virtue and deserved happiness.' Like many of his contemporaries, Kant assumed that we could best promote such happiness by following the various principles of common sense morality. And as I note on page 410, Kant once even proposed a hedonistic version of Rule Utilitarianism. Rather than claiming that Kant would have been devastated if his theory implied a form of Rule Consequentialism, Morgan's claim should instead be that, given Kant's assumptions about value, Kant should have proposed a version of Rule Consequentialism that is neither hedonistic nor Utilitarian. (For a discussion of similar objections, see the commentaries by Wolf and Herman in Volume Two, and my responses.)

In his vigorous defence of an expressivist, desire-based theory of reasons, James Lenman writes: 'Where there is a space of desires there is a space of reasons but only when those desires are *your* desires. From

the third person perspective where there is a space of desire there is merely a space of desire'(13). That is also how I would describe such views. Though Lenman claims that there are normative reasons, his view implies that, when we consider the world objectively, we should conclude that there are no such reasons. Lenman rightly criticizes my failure to discuss the metaphysical and epistemological objections to my belief that there *are* such reasons. I have now tried to respond to these objections by writing Chapters 31 to 34. Michael Smith's forceful paper I discuss briefly in Section 11. I shall add here that, in the last part of his paper, Smith assumes that, on my view, facts about reasons can be explained by appealing to facts about value. That is not so.

406 *This claim may seem undeniable. And if this claim were true, this version of the Kantian Formula would require us to be Act Consequentialists.* Kagan suggests a partly similar argument in Kagan (2002) 128, and 147–150. Many professional philosophers have told their students that Kant's Formula of Universal Law conflicts with Act Consequentialism. The Act Consequentialist maxim, these people assume, could not be rationally willed to be a universal law. Some students must have asked 'Why not? Why can't we rationally will that everyone does what would make things go best?' We would expect that, by now, there would be some standard answer to this question, which would be repeated in many introductory texts on ethics. Surprisingly, that is not so. Kagan is the first writer known to me to have discussed this question. (Sidgwick however writes: 'I could certainly will it to be a universal law that men should act in such a way as to promote universal happiness; in fact it was the only law that it was perfectly clear to me that I could thus decisively will, from a universal point of view' (ME xxii).)

When Kagan argues that we could rationally will that everyone follows the Act Consequentialism maxim, he appeals to claims about instrumental and self-interested reasons. Kagan notes that, if we choose that everyone becomes Act Consequentialists, we might be required to make significance sacrifices for the good of others. It would be rational to take that risk, Kagan claims, given the 'logical possibility' that we might be in anyone's position. This amounts to assuming a veil of ignorance, as in Rawls's version of Contractualism. Hare gives a similar argument in Hare (1997). These arguments differ in important ways from the Kantian arguments that I have been discussing. For another, even more different Kantian argument for Consequentialism, see Cummiskey (1996). Kant's texts are inexhaustibly fertile, provoking in different people very different thoughts.

Kagan argues that we *could* rationally will that everyone follows the Act Consequentialist maxim. So I may be the first writer who argues briefly that we could *not*.

If everyone always did whatever would make things go best, everyone's acts would, in most cases, have the best possible effects. This is not always true, as Gibbard (1965) and Regan (1980) point out. In some cases, *each* of us might be following AC though *we together* are not doing what would make things go best. It may be true of each member of some group that, if he alone had acted differently, that would have made things go worse, but that, if everyone had acted differently, things would have gone better. One such case is *Mistake*, as described on page 000 above. Each of us would have followed AC not only if we both do A, thereby saving everyone's life, but also if we both do B, thereby saving only some people. If either of us does B, it would be worse if the other did A. But if we both do B we save fewer people than we could have done. This complication does not undermine the claims in my text.

406 *As before, in losing many of our strong loves, loyalties, and personal aims, many of us would lose too much of what makes our lives worth living.* This provides a different way in which we could not always do what would make things go best. If we have the motives having which would make things go best, we shall often choose to do what would make things go worse. But if we caused ourselves to lose these motives, so that we never acted in such ways, we would thereby make things go worse. On these assumptions, we would make things go best by causing ourselves to be people who do not always do what would make things go best.

This formula does, however . . . And, compared with the UA-optimific principles, these principles are more similar to AC. That is mainly because, in asking which are the principles whose being universally followed would make things go best, we can ignore the various ways in which, when people try to make things go best, they can go astray, through miscalculation, self-deception, and the like. We can also note that, on some versions of Rule Consequentialism, we appeal to the principles that are optimific *during the period in which we are living*. Kantian Contractualism might also take this form. If we ask which principles are UF-optimific in the 21st Century, these principles would be even closer to AC than they would be in most other centuries. Given the world's extreme inequalities in wealth and power, the existence of widespread poverty, and our advances in technology, many Act Consequentialists would now be able to do much more good than other people could have done in earlier centuries. So AC might *now* be UF-Optimific. But AC was not, I believe, UF-optimific in earlier centuries. And if, as we can hope, such poverty and inequalities will be abolished, AC will cease to be optimific in future centuries.

It is worth mentioning here a revised version of Kant's Consent Principle. On what we can call

> *the Formula of Universally Willable Acts*: An act is wrong unless this act could be rationally willed by everyone.

Everyone could will some act if, were they given the choice, everyone could rationally choose that this act be done. This version of the Consent Principle may also be, in its implications, closer to AC.

409 *In asking how we could get closest to Kant's ideal, we must compare the goodness of virtue and happiness.* It is easy to go astray here. Some writers claim that, if we had to choose between doing our duty and promoting happiness, we ought always to do our duty. But we could accept this claim even if we believed that we would never have to make this choice, since our only duty is to promote happiness.

413 *We could also claim (K) . . . and (L) . . .* These claims, we can note, cannot be stated the other way round. We could not defensibly claim that, if everyone could rationally will that some principle be universally accepted, that makes this principle optimific, by making it one of the principles whose universal acceptance would make things go best. The effects of some principle's acceptance do not depend only on whether this principle's acceptance could be rationally willed. Nor could we claim that, if some principle is the only one that no one could reasonably reject, that would make it the only principle whose universal acceptance everyone could rationally will. My argument for (L) consists in claims (A) to (I) above, and there is no similar argument for this reversed version of (L).

415 *For some moral theory to succeed . . . we could then justifiably reject this theory.* In claiming that we could justifiably reject some theory, or belief, I do not imply that this theory or belief is false. We might justifiably have some false beliefs.

417 *If these claims are true . . . fit together like two pieces in a jig-saw puzzle.* Though Kantian Rule Consequentialism has different versions, which may conflict, these conflicts are not between the Kantian and Rule Consequentialist parts of this view.

417 *Rule Consequentialism may instead be founded . . .* According to Kantian Rule Consequentialists, we ought to follow the optimific principles because these are the only principles whose being universal laws everyone could rationally will. This version of Rule Consequentialism is, in this sense, founded on Kantian Contractualism. As I have also claimed, however, it is because these principles are optimific that these are the principles whose being universal laws everyone could rationally will. In this other sense, it is Rule Consequentialism that is more fundamental. But there is no contradiction here, since these views support each other in different ways.

We can also note that, though Kantian Contractualism provides this firmer foundation for Rule Consequentialism, it is only Rule Consequentialism that could be accepted on its own. As I have argued in Section 62, it is only the optimific principles that everyone could rationally will to be universal laws. So Kantian Contractualists must be Rule Consequentialists.

433 *Appendix B Rational Irrationality.* This appendix was written in 1994, in response to the text of what became Gauthier (1997). I have not tried to take into account Gauthier's later work.

In an early article, Gauthier argued . . . Gauthier (1975). This argument's fullest statement is in MA. . . . *I tried to refute that argument.* In an unpublished paper and in Sections 7–8 of RP.

Gauthier assumes . . . we can regard him as appealing to Rational Egoism. Since Gauthier means by our *utility* the fulfilment of our *present* informed considered preferences, what he appeals to is, strictly, the *Deliberative Theory*. But as Gauthier remarks (MA p. 6), most of his claims apply equally to Rational Egoism. And Gauthier often uses words, like 'benefit' and 'advantage', that refer more naturally to our interests rather than our present preferences. So we can here ignore the differences—though they are often great—between the Deliberative Theory and Rational Egoism. We can suppose that, in all of the cases we discuss, our present considered preferences would coincide with what would be in our own interests.

If we appeal to Rational Egoism . . . or be expectably-best for us. What is expectably-best may not be the same as what we can expect to be best. Some acts are expectably-best for us though we can know, for certain, that they will not actually be best for us. Trying to do what is actually best may be, given the risks, irrational.

434 *It would be even better for us, as I pointed out* . . . In RP Sections 7–8. *Gauthier replied* . . . Gauthier gave this reply in MA (especially, 173–4). In Gauthier (1997), Gauthier later gave up the claim that we could not deceive others. He suggested that, if we remained self-interested, and merely appeared to be trustworthy, that would be worse for us. Thus he writes: 'the overall benefits of being able to promise sincerely . . . may reasonably be expected to outweigh the overall costs of keeping promises when one could have gotten away with insincerity' (p. 26). But if we could get away with insincerity, what are the benefits from being able to promise sincerely? Gauthier might appeal, like Hume, to the benefits of peace of mind, and a good conscience. But that seems insufficient for his purposes. Gauthier also claims that, even if we were generally trustworthy, we would be able to make some insincere promises. But this merely limits the costs of sincerity. It does

not suggest that there is any gain. For Gauthier's distinctive argument to get off the ground, he needs, I believe, his earlier assumption that we could not rationally hope to deceive others.

435 *In his earliest statements* . . . See, for example, MA Chapter VI . . . *I challenged (A) as follows* . . . In RP Sections 7–8 . . . . *But it might not be rational for us to act upon them.* I also supposed that it might be rational to change our beliefs about rationality. This, too, was intended to help Gauthier's argument. If we did not change our beliefs, we would be doing what we believe to be irrational, and that might be enough to make our acts irrational. But we can ignore this point here.

Gauthier once accepted . . . As he wrote (like Queen Victoria), 'We are unmoved' (MA, p. 185).

436 *According to (B)* . . . Gauthier asserted (B)—which he calls his 'second level of commitment'—in Gauthier (1997) 40. I discussed a similar claim, which I called '(G1)' (RP 13). On Gauthier's second level of commitment, it is rational to act on a disposition 'so long as one reasonably expects past and prospective adherence to the disposition to be maximally beneficial'. This claim may seem to mean 'if one both reasonably believes that adherence to this disposition in the past has been beneficial, and reasonably expects that adherence to it in the future will be beneficial'. But this cannot be what Gauthier intends, since it would remove the difference between his second level of commitment and his first level (discussed below). Gauthier must mean: 'if one can reasonably believe that acquiring it was beneficial in one's life as a whole, taking the past and future together.'

Gauthier's move from (A) to (B), or from his third to his second level of commitment, hardly damages his defence of rational morality. On the view defended in MA, for morality's constraints to have rational force for us, accepting these constraints must have been expectably-best for us. On Gauthier's revised view, for these constraints to have rational force, they must also be known not to have been on the whole bad for us. Most of the constraints of Gauthier contractualist morality would meet this second requirement.

We can revise the example . . . I would have died some time ago. Perhaps I would have obeyed some order that would have proved fatal. . . . *According to (B), I should deny you the coconut, and be blown to pieces.* It might be objected that I acquired too crude a disposition. Perhaps I should have become disposed to ignore threats, except in cases in which I believed that acting in this way would be disastrous. But as Gauthier says, 'I may reasonably have believed that any qualification [to my disposition] would reduce its *ex ante* value, so that unqualified threat-ignoring offered me the best life prospects' (Gauthier (1997) 39).

We can add the assumption that only the unqualified disposition would in fact have been as good for me. (There is another reason not to allow this disposition to take this qualified form. If we did, we would have to allow similar qualifications to the disposition of trustworthiness. As we shall see, that would undermine Gauthier's argument.)

When applied to promises . . . Why should I remain a threat-ignorer? Gauthier endorses the action of a would-be deterrer who, when deterrence fails, disastrously carries out her threat. He writes 'Her reason for sticking to her guns . . . is simply that the expected utility . . . of her failed policy *depended* on her willingness to stick to her guns' (Gauthier (1984) 489.) So what? Her expectation may have depended on that willingness. But why should she remain faithful now? (We are not here discussing true *fidelity*.)

To handle this version . . . it would be rational for me to try to lose this disposition. Note that, in claiming this, we need not appeal to Rational Egoism. We need not assume that this attempt would be rational because it would be likely to be good for me. Since Gauthier rejects Rational Egoism, that would beg the question. But even on Gauthier's theory, it would be rational for me to try to lose this disposition. Suppose that I lose my dispositions whenever they become disastrous. It would be in my interests to have this meta-disposition. So, on Gauthier's theory, it would now be rational for me to act upon it.

To handle this version. . it would be irrational for me knowingly to bring about my death. Suppose first that, if I tried, I could cease to be a threat-ignorer. As I have just argued, it would then be irrational for me to keep my disposition. If Gauthier accepts this conclusion, could he still assert (B)? Could he claim that, even though it would now be irrational to *keep* my disposition, it must still be rational to act upon it?

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

If I have irrationally remained prudent, there is a different explanation of why it can be rational to act upon this disposition. Doing so will be better for me. The rationality of this act need not be defended by an appeal to the rationality of the disposition, or of my having

kept the disposition, on which I act. Things are quite different with ignoring your threat, in a way that I know will be disastrous for me. If this act is to be claimed to be rational, that can only be by an appeal to the rationality of the disposition on which I am acting. And if it is now irrational for me to keep this disposition, there seems no reason to conclude that, if I keep it, it must be rational for me to act upon it.

Suppose, next, that I could *not* lose my disposition, even if I tried. Gauthier might say that if, that is true, it is not irrational for me to keep this disposition. This is not something that I *do*. But it *would* be irrational for me to keep it, if I *could* lose it. This seems enough to undermine the claim that it must still be rational to act upon it.

437 *If he appealed to (C)* . . . (C) is one interpretation of what Gauthier calls the 'weakest' version of his view, or what he calls his first level of commitment. On this view, he writes, one should act upon some disposition, even though one's actions are 'costly . . . only so long as one reasonably expects adherence to the disposition to be prospectively maximally beneficial' (Gauthier (1997) 39).

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

When Gauthier considered this example . . . he claimed that it would be rational for me to act upon this disposition. My drug-induced insanity, Gauthier claims, is 'the rational disposition in such situations, and the actions to which it gives rise are rational actions' (Gauthier (1997) 38). Gauthier means only that it is in my interests to have this disposition *now*. He is not here concerned with a choice between two permanent dispositions. If I had to choose my disposition, not just until the police arrive, but for the rest of my life, it would be better to remain sane and give the man my gold.

438 *He does not, I believe . . . even if only for just long enough to break our promise.* Gauthier might extend his claim about translucency. He might say that we could not have reason to believe that, if we broke our promises, we could keep this fact secret. But this reply would jettison

what is novel in Gauthier's view, since it would revert to the ancient claim that honesty is always the best policy.

Gauthier might try a different reply. He might claim that . . . we would be unable to lose, or overcome, this disposition. There is one reading on which this claim must be true. It may be said that, if we are able to suspend our disposition, we were not *truly* trustworthy. But this reading is irrelevant since, for Gauthier's purposes, all that matters is whether we *appeared* trustworthy. It would be quite implausible to claim that, if we break some agreement, we cannot have earlier appeared to be trustworthy, even if, at the time, we sincerely intended to keep this agreement.

If this claim is to help Gauthier's case, he must make other revisions in his view. He writes: 'a disposition is rational if, among those humanly possible, having it will lead to one's life going as well as having any other' (Gauthier (1997) 31). This appeal to *human* possibility seems at odds with other parts of Gauthier's view. He claims elsewhere that we should not ask which dispositions are in general rational, since the answer may depend on a particular person's circumstances. Thus he writes, 'there need be no one disposition that, independently of an agent's circumstances, is sufficient to ensure that his life will go as well as possible, and thus I do not need to suppose that there need be a single supremely rational disposition' (Gauthier (1997) 31-2). A person's circumstances can surely include what is possible for *this* person.

This appeal to human possibility also raises a problem for Gauthier's argument. Trustworthiness is *not* the disposition that, among those *humanly* possible, is most advantageous. It would be more advantageous to appear to be trustworthy but to be really prudent; and that is surely possible for some human beings. If Gauthier appeals to what is humanly possible, he would have to judge trustworthiness to be an irrational disposition, even when it is had by people for whom, since they could not deceive others, it is the most advantageous possible disposition.

439 *I believe not. For reasons that I give above . . .* On pages 260–63. *. . . But it cannot show that my act is rational.* In the doctrine that 'ought' implies 'can', the sense of 'can' is compatible with determinism. If that were denied, and we assumed determinism, we would have to claim that *every* act is rational.

Reconsider Schelling's Case . . . which will be much better for me. It would of course be even better if I merely appeared to be insane. But we can suppose that this is not possible, since if I had not taken the drug, the robber would know this. (Perhaps one of the drug's effects is

a characteristic look in the eyes.) Being actually in this state is then the disposition that is best for me.

440 *Gauthier objects that my crazy acts . . .* Gauthier (1997) 37.

Though my acts . . . That is enough, on Gauthier's view, to make these acts rational. Provided, of course, that these bad effects do not outweigh the good effects of my disposition. Gauthier need not claim that, if I killed myself or my children, that would be rational.

Hume notoriously claimed . . . They could still, on Gauthier's view, be rational. It may be said that, in one respect, Gauthier's view is less extreme than Hume's. Even if my act has bad effects, these must be outweighed by the good effects of having my disposition. But we can remember here that, on Gauthier's main view, I maximize my utility if I fulfil my present considered preferences, and these need not coincide with my interests. As on Hume's view, these preferences could be as crazy as we can imagine. The difference between these views is that, on Hume's view, for my act to be rational, I must at least be trying to fulfil my aims, while on Gauthier's view, my acts need only be the side-effects of a state the having of which will achieve these aims.

Gauthier assumes that, to be rational . . . This view we can now call G. As Gauthier writes: 'Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices' (MA 187).

441 *In the cases with which we are concerned . . . which will be bad for us.* It may seem that, if that is true, breaking our promises cannot be better for us. But this may not be so. The bad effects come, not from our breaking of these promises, but from the fact that we are both translucent and disposed to break our promises whenever this will be better for us.

When we cannot maximize at both levels . . . would outweigh the bad effects of our acts. It is worth explaining why. In our assessment of the good or bad effects of our dispositions, we include the acts to which these dispositions would or might lead. If it is best for us to have some disposition, even though this will lead to acts which are bad for us, those effects must be outweighed. Since the assessment of our dispositions includes the assessment of our acts, but goes beyond it, this is the assessment that tells us what will be, on balance, best for us.

Gauthier claims that, given this fact . . . MA 170 *. . . This assumption may be incorrect . . . E agrees with G that we should try to* have *these dispositions.* It may be questioned whether G tells us, if we can, to acquire these dispositions. That does not follow from the fact that, if we do, that will be better for us. If G does not tell us to act in this way, that

would be an objection to G, and would again undermine Gauthier's argument. But Gauthier might claim that, in trying to acquire these dispositions, we would be acting on an advantageous, or maximizing, meta-disposition.

Gauthier may think that, if we accept E, we would always do what E claims to be rational. He would admit that, in practice, few of us are always rational. But he might claim that, in assessing the plausibility of these theories, we should consider what would happen if we always did what these theories told us to do. He might then claim that, if we fully followed S, we would always maximize at the level of our acts.

Gauthier may think . . . If instead we acquire these dispositions, we cannot always maximize with our other acts. It may be objected that, if we cannot always do what E claims to be rational, E cannot claim that we ought to do so. 'Ought' implies 'can'. But this confuses two questions. When I say that we cannot always do what E claims to be rational, I mean that this is not causally possible. This is the kind of possibility that is relevant when we are comparing the effects of our having different dispositions. The sense of 'can' that is implied by 'ought' does not, as Gauthier agrees, require such causal possibility, since this other sense of 'can' is compatible with determinism.

Since we cannot always do . . . what we are doing is rational. It may seem that, if we cannot always do what E tells us to do, there is no way of predicting when we shall follow S. That is not so. Suppose that we are now always disposed to do what we believe to be rational. If we know that we can acquire maximizing dispositions, we shall then do so, even though we know that this will cause us later to act irrationally. Acquiring these dispositions is, according to E, the rational thing to do. It is only *after* acquiring these dispositions that we shall start acting in ways that E claims to be irrational.

442 *When he discusses nuclear deterrence . . .* Gauthier (1984) and (1985) *. . . we would be unable to form this intention.* Gauthier (1985) 159–61.

It would be implausible to claim. . . . We might then be able simply to decide to do so. See McLennen (1988).

This may only shift the problem . . . It might then be in our interests to make ourselves acquire this belief. Such a claim is fairly plausible in the case of trustworthiness, the disposition that is Gauthier's chief concern. If we could not conceal our intentions, as he assumes, it might be better for us if we intended to keep our promises, even when this way of acting would be worse for us. Unless we have this intention, others might exclude us from advantageous agreements. And for us to be able to form this intention, we might have to believe that it is rational to keep such promises.

443 *Kagan suggested* . . . In a letter to me.

Gauthier later revised his argument . . . See MA (p. 182) and Gauthier (1997) 31). (But see also MA, pp. 170 and 158.)

444 *Gauthier, however, rightly rejects (D)* . . . Gauthier (1997) 36.

Could Gauthier withdraw this claim, and appeal to (D)? At one point, Gauthier comes close to accepting (D). He cites my book's version of (D) — -there called '(G2)' — and writes, 'to this extent I accept . . . (G2)' Gauthier (1997) 40.

This argument might show something . . . that this belief is a useful illusion. . Gauthier might reply that normative beliefs are not really *beliefs*, which might be *true*, or be *illusions*. But this would not rescue Gauthier's argument. Even on a noncognitivist view, we must give some content to the notion of a normative belief. We must be able to claim that an act *is* rational, and be able to assert or deny different theories. My remarks could be restated in these terms.

445 *Lewis rejected* . . . Lewis (1985) . . . *In his reply, Gauthier denied* . . . Gauthier (1985) 159–61.

446 *Gauthier suggests another argument* . . . Gauthier (1997) 30. . . *This unwelcome consequence, Gauthier claims* . . . Gauthier (1997) 36.

Gauthier makes one other claim . . . 'rationality' is a technical term . . . Gauthier (1997) 38.

447 *No other moral theory, Gauthier claims, can achieve this aim.* MA, 17.

450 *This act, these people believed . . . it would be wrong to save these people by killing someone else.* These may not be the best descriptions of what makes these acts permissible or wrong. For another account, see Kamm (2007) Chapters 1 to 5.

References

These notes refer to the Bibliography. Some of these notes give only the opening words of some block of text, because that is enough to make some reference clear.

xxxiii *Kant is the greatest . . . . the best book . . .* In these opinions I follow Broad (1959) 143–4. The best books on Sidgwick are Schneewind (1997) and Schultz (2004).

xxxiv *the critical philosophy . . . . Declaration concerning Fichte's Wissenschaftslehre*, Kant, *Correspondence* translated and edited by Arnulf Zweig (Cambridge University Press 1999) 560.

The book solves nothing . . . Sidgwick, HSM 284.

Always thoughtful, often subtle . . . HSM: 177.

Sidgwick also refers . . . long-winded and difficult dullness. (2006) 398. Sidgwick is referring here to another of his books, but he would have applied this claim, I believe, to his *Methods*.

Sidgwick's dullness . . . another book on ethics. Sidgwick (2000) xxviii.

xxxv *At Cambridge . . . the high road.* HSM: 396.

There is no doubt . . . redoubled force. HSM: 92.

I am bearing the burden . . . elaborate sentience. HSM: 170–1.

xxxvi *to suppose . . . he has voted.* ME 298–9 (my italics).

. . . the Cosmos of Duty . . . ME First Edition (1874) 473.

I cannot fall back . . . philosophic despair. ME 507 note.

xxxvii *. . . the selfish man . . . an insignificant fraction* (ME 501). Characteristically, Sidgwick adds: 'I do not think, however, that we are justified in stating as *universally* true what has been admitted in the previous paragraph. Some few thoroughly selfish persons appear at least to be happier than most of the unselfish; and there are other exceptional natures whose chief happiness seems to be derived from activity, disinterested indeed, but directed towards other ends than human happiness.'

. . . even a man . . . consistent wickedness. Sidgwick (2000) 118.

Sidgwick warned . . . dry and repellent. ME 295.

the sympathy . . . their decay. ME 437.

xxxviii *It may be said . . . regard for the recipient.* ME 248 note.

there seems to be no justice . . . made him better. ME 284.

Thus the Utilitarian conclusion . . . kept esoteric. ME 490.

really penetrating . . . incapable of maintaining. Mind, 1877 125–6, quoted in Schultz (2004) 349.

[The book] . . . attack of this kind. HSM: 74.

xxxix *Have been reading . . . due to it.* HSM: 421.

Sidgwick was . . . fair to his opponents.

I shall praise it . . . depressed about ethics. (1906) 411.

incessantly refines . . . or nothing. Broad (1959) 144.

xl *Criticizing himself . . . every day.* (1906) 93.

This remark . . . point of genius. Rashdall (1892). . . *extreme acuteness.* Broad (1959) 14.

xli *There are deeper . . . 'most exasperating'.* O'Neill (1989) 126.

xlii *It would also . . . quite inconclusive.* Kemp Smith (1915) 531 . . . *all the good in his power.* Kant: *R*72.

xlii *It would also . . . of the great thinkers.* Kemp Smith (1915) 527. Though this remark is about Kant's *First Critique*, it also applies, I believe, to Kant's books on ethics.

Consistency . . . greatest duty. C2 24.

If we look upon . . . his own will. G 432.

xliii *Suicide . . . the Stoic's principle . . .* LE 127,148, 369

It is the word . . . one commentator suggests . . . Korsgaard (1996) 126.

xliv *For another example . . . a mere thing.* MM 429–30.

Kant is sometimes . . . to criticism. HSM: 177.

Some of Kant's views . . . speak to us. Rawls (2000) 18.

xlv *As well as finding . . . irreducibly normative truths.* See Nagel's wonderful *The View from Nowhere* (Oxford University Press, 1986), especially Chapter VIII, and his *The Last Word* (Oxford University Press, 1997).

31 *It is hard to explain . . . counts in favour of . . .* I follow Scanlon, *What We Owe to Each Other*, (Scanlon 1998) Chapter 1.

36 *For us to be acting rationally . . . This is not, I shall argue later . . .* In Section 16.

To be fully rational . . . .. these requirements raise several interesting questions. See Kolodny (2005), Broome (forthcoming), and Scanlon (forthcoming). . . *I shall later give some arguments . . .* In Section 17.

38 *Ice formed on the butler's upper slopes.* Wodehouse (1952) 93.

39 *When something is . . . could not give us reasons.* WWO 97. Scanlon calls this *the buck-passing view*.

This view . . . such a claim. Though Scanlon claims that goodness and badness are not reason-giving properties, he sometimes mentions such derivative reasons. For example, Scanlon writes: 'There can be more than one reason to respond to a human being who is in pain: his pain is bad, and we may owe it to him to help him relieve it' (WWO, 181). The source of the first reason, Scanlon would agree, are the features that make this person's pain bad.

46 *Most things are good.. out there, shining down.* Korsgaard (1996) 225, 278.

48 *Though we can seldom . . . often give our reasons.* I follow Scanlon: WWO 18–22.

52 *Telic reasons . . . intrinsically better.* (In drawing these distinctions, I partly follow Korsgaard (1996) Chapter 8.)

54 *Since these claims are controversial . . . neither they nor our conscious state would be bad.* For a different view, see Rachels (2000).

55 *First, many people believe . . .* Korsgaard (1996) 262.

Korsgaard's remarks . . . by liking it. Korsgaard (1996) 284.

57 *No one has these attitudes . . .* I discuss these attitudes to time in Sections 62–70 of RP. (In that rather tortuous discussion I failed to make it clear that, in my view, the most rational attitude is temporal neutrality.)

61 *The Informed Desire Theory . . . in our actual uninformed state.* I follow Peter Railton, 'Moral Realism', in Darwall (1992B) 142 and and note 15.

65 *Subjectivism about Reasons . . .* Korsgaard (1996) 317. Williams (1985) 19.

We ought, I believe, to reject . . . As before, see Scanlon WWO Chapter 1. See also Raz (2000) Chapter 2.

Second . . . these desires and these beliefs. As Scanlon writes: 'if having a desire involves seeing something other than that desire as providing a reason, this may explain the plausibility of the idea that desires provide reasons. It is true that having a desire involves taking oneself to have a reason. The mistake lies in confusing the reason with one's taking it to be a reason' (Scanlon (2002) 338).

66 *Sixth . . . no reason to vote as they do.* I follow Scanlon (2003B). Similar claims apply to cost-benefit analyses. These analyses can rightly appeal to people's preferences, without thereby assuming a desire-based theory about reasons . . . . *Nozick, for example, claims* . . . Nozick (1993) 176.

70 *There is no science . . . LE*, 58–9 (27: 264–5), and Sidgwick's ME 74–5.

73 *When making . . . or other close relatives.* For discussions of such reason-giving facts, see Kolodny (2003), (2010) and (forthcoming).

77 *Subjectivists might reply . . . 'modest amount of prudence'.* Williams (2006) 111.

78 *Subjectivists cannot, however . . . We can be procedurally rational.* Williams draws this distinction in his 'Internal Reasons and the Obscurity of Blame' in Williams (1995) 36–7.

Knowing that people are . . . Rawls (1996) 49.

79 *It might be objected . . .* Michael Smith 'Desires, Values, Reasons, and the Dualism of Practical Reason', in Suikannen (2009). I here summarize the principle that Smith calls 'R' (120).

By appealing . . . future Tuesday. (op.cit. note 5.)

88 *To illustrate . . .* Smith (2004) 269–70.

94 *Nor can they coherently assert . . . more reason to try to fulfil.* For similar objections to theories of this kind, see Enoch (2005).

frees us from assessing the rationality of a choice. Korsgaard (1996) 261. If we haven't assessed the things we are choosing, it is not clear that our choices deserve to be called rational.

96 *One such person . . . what is important to us.* Frankfurt (1988) 81, and 91 note 3.

97 *Someone genuinely does not care.* Frankfurt (2004) 22.

98 *It is not the factual question.* Frankfurt (2004) 28

Love is itself . . . Frankfurt (2004) 37.

99 *When some end . . .* Frankfurt (2004) 56. Frankfurt uses the word 'obligation', but he is not discussing morality, so this obligation can at most be a decisive reason.

103 *a person's good . . .* Rawls (TJ) 395. Rawls's *thick* theory of the good is surprisingly similar.

would adopt if . . . Rawls (TJ) 417.

Though it is a normative question. . it is a psychological question what, after such deliberation, someone would in fact choose. As Sidgwick notes in ME 112. Rawls claims that, in giving this definition, he is following Sidgwick. But though Sidgwick suggests a similar definition, and claims that it has some merits, Sidgwick rejects this definition, in part because

it isn't normative. Sidgwick then defines his good as 'what I should practically desire if my desires were in harmony with reason, assuming my own existence alone to be considered'(ME 109–113). (In an earlier edition, Sidgwick refers to 'the ultimate end or ends *prescribed* by reason as what *ought* to be sought or aimed at' (ME 5th edition 112) my italics.)

which would be chosen. . Rawls (TJ) 408 (my italics).

We can be . . . would want, or approve. Rawls (TJ) 184–5.

104 *This example . . . no way to get beyond deliberative rationality.* Rawls (TJ) 560.

104 *Rawls intends . . . everything we might want to say.* Rawls (TJ) 401. See also 111 and 451.

107 *These bleak views . . . Most Subjectivists take it for granted . . .* That is true even of some writers who claim to be questioning desire-based subjective theories. Nozick, for example, makes twenty three proposals about how we should go beyond a purely instrumental, desire-based account of rationality (Nozick (1993) Chapter V. None of these proposals include the idea that we might have reasons to have desires that are given by the intrinsic features of their objects, or what we want.

108 *Some Subjectivists argue . . . desire-based.* This argument is suggested, for example, by Williams's remarks in 'Internal and External Reasons' (Williams (1981) 102 and 106–7, and in 'Internal Reasons and the Obscurity of Blame'(Williams (1995) 39. For a longer discussion of such arguments, see my 'Reasons and Motivation' Parfit (1997).

110 *For the philosophical naturalist . . .* Darwall (1992) 168.

There seems nothing for value to be . . . Stephen Darwall, Allan Gibbard, and Peter Railton, in Darwall (1992B) and Darwall (1996) 176–7.

Such Naturalist accounts . . . must be false. In these remarks, I follow Nagel (1986) and (1997).

113 *Many people . . . Hume suggests . . .* Hume writes that though desires cannot be strictly 'contrary to reason', they are, in a loose sense, 'unreasonable' when they are 'founded on false suppositions'. Hume's *Treatise*, Book II, Part III, Section III. I discuss Hume's view further in Section 111.

113 *To be fully rational. . these requirements here.* See Kolodny (2005), Scanlon (forthcoming) and Broome (forthcoming).

118 *Trying to reach the truth . . . practical and epistemic reasons support answers to different questions.* See Kelly (2003).

119 *As before, however. . As Scanlon notes . . .* WWO Chapter 1.

121 *When Scanlon discusses . . .* WWO 29–31.

123 *I have rejected . . . At one point, Scanlon suggests . . .* WWO 25–30.

124 *On one common view . . . rational irrationality.* For some examples, see Appendices A and B.

126 *According to some other, similar . . . Brandt suggests.* In Brandt (1979) and (1992).

128 *Things are different.. theoretically very interesting.* For some fascinating arguments, see Temkin (1987), (1996), (2009), and (forthcoming), and Rachels (1998). If some of these arguments succeed, they would have great practical importance.

130 *Objective theories . . . That, I shall argue, is not true.* In Chapter 22.

131 *In his great, drab book . . .* Though Sidgwick calls Egoism one of 'the *Methods of Ethics*', he is discussing a view about what he calls 'the *rational* end of conduct for each individual' (ME xxviii, my italics).

The Dualism . . . ME, Concluding Chapter. This is only part of Sidgwick's view. Sidgwick makes other claims, to which I shall turn in Section 20.

133 *Sidgwick's defence . . . life as a whole.* In Sidgwick's words, 'It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently "I" am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual' (ME 498). . . *the fundamental fact for ethics.* Findlay (1961) p 294. Compare Rawls's claim: 'Utilitarianism does not take seriously the distinction between persons' (Rawls (TJ) 27).

Sidgwick's Dualism . . . what Nagel calls . . . In Nagel (1986) especially chapters VIII and IX, and Nagel (1991) Chapter 2. Sidgwick writes of 'the inevitable twofold conception of a human individual as a whole in himself, and a part of a larger whole. There is something that it is reasonable for him to desire, when he considers himself as an independent unit, and something again which he must recognize as reasonably to be desired, when he takes the point of view of a larger whole' (Third Edition of ME, p 402, quoted in Schneewind (1977) 369.) Sidgwick also writes: 'the good of any one individual is of no more importance, from the point of view . . . of the Universe, than the good of any other . . . And . . . as a rational being I am bound to aim at good generally . . . not merely at a particular part of it' (ME 382).

(Nagel calls 'the transcendence of one's own point of view . . . the most important creative force in ethics' (Nagel (1986), 8).)

133 *Suppose next . . . no guidance.* ME 508.

134 *We can call this the Two Viewpoints Argument.* Sidgwick does not explicitly assert (D) and (E), but his reasoning seems to need these premises.

136 *I shall soon turn . . . . their well-being.* For a discussion of these reasons, see Kolodny (2003).

Suppose I have been . . . Nagel (1986) 160.

138 *the pain can be detached . . .* Nagel (1986) 161.

142 *Sidgwick doubted . . .* ME 386 note 4.

When they consider . . . or a fool. Reid (1983) 598. Reid may not be committed to this view, since he believed that we did not face this dilemma.

143 *the whole system . . .* ME First Edition (1874) 473. Since Sidgwick cut this passage from later editions, it is worth quoting in full: 'But the fundamental opposition between the principle of Rational Egoism and that on which such a system of duty is constructed, only comes out more sharp and clear after the reconciliation between the other methods. The old immoral paradox, 'that my performance of Social Duty is good not for me but for others', cannot be completely refuted by empirical arguments: nay, the more we study these arguments the more we are forced to admit that, if we have these alone to rely on, there must be some cases in which the paradox is true. And yet we cannot but admit with Butler that it is ultimately reasonable to seek one's own happiness. Hence the whole system of our beliefs as to the intrinsic reasonableness of conduct must fall, without a hypothesis unverifiable by experience reconciling the Individual with the Universal Reason, without a belief, in some form or other, that the moral order which we see imperfectly realized in this actual world is yet actually perfect. If we reject this belief, we may perhaps still find in the non-moral universe an adequate object for the Speculative Reason, capable of being in some sense ultimately understood. But the Cosmos of Duty is thus really reduced to a Chaos: and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure'.

144 *There is a third . . . what justice requires.* Rawls (TJ) 575.

146 *We can next note . . . less important.* This is forcefully argued, for example, by Kolodny (2005), Scanlon (2007), and Broome (forthcoming).

156 *This view is in itself . . . plausible implications.* See 'Moral Luck' in Nagel (1979).

158 *Of the facts . . . slave to escape.* See Bennett (1974) 123–134.

161 *There is another reason. . as Sidgwick points out.* ME 207–8.

168 *The good of any one individual . . .* ME 382–3.

168 *This kind of . . . 'ignoble end'.* ME 200, 403.

169 *These people may believe . . . rival to morality.* Sidgwick, for example, writes that a Utilitarian morality may be accepted by people 'who choose "general good" as their ultimate end, whether they do so on religious grounds, or through the predominance in their minds of impartial sympathy, or because their conscience acts in harmony with Utilitarian principles, or for any combination of these or any other reasons.. ' (Sidgwick (2000) 607). Sidgwick's distinction between the second and third groups seems to suggest Impartial-Reason Consequentialism.

173 *Suppose first . . . what Scanlon calls . . .* WWO, 97.

174 *There are some deep . . .* Rawls (TJ) 52, Nagel (1995) 182.

Rather than proposing . . . 'the supreme principle of morality.'. The Groundwork, henceforth *G*, 392. Page references are to the page numbers of the Prussian Academy edition, which are given in most English translations.

177 *the Formula of Humanity . . .* In Kant's words: 'the human being and in general every rational being exists as an end in itself, not merely as a means to be used by this or that will at its discretion; instead he must in all his actions, whether directed to himself or also other rational beings, always be regarded at the same time as an end' (*G*428–9).

he whom I want to use . . . G 430.

Korsgaard comments . . . Korsgaard (1996) 139.

O'Neill similarly writes. O'Neill (1989) 111.

178 *Korsgaard concludes . . .* Korsgaard (1996) 140.

Kant's claim . . . deception is always wrong. I here follow Korsgaard (1996) 295–6. (Korsgaard does not herself believe that deception is always wrong.)

179 *Return now to Kant's claim . . .* After saying that the person whom he deceives 'cannot possibly consent to my way of treating him', Kant refers to this remark as having introduced what he calls 'the principle of other human beings' (G 430). (A) is the simplest statement of this principle.

O'Neill similarly writes . . . O'Neill (1989) 110.

180 *the Choice-Giving Principle* . . . Korsgaard writes: 'the other person is unable to hold the end of the very same action because the way you act prevents her from choosing whether to contribute to the realization of that end' (Korsgaard (1996) 138–9).

181 *I shall call this the Consent Principle.* Other writers have assumed or claimed that this is what Kant means. See, for example, Hill (1992) 45.

We have several reasons . . . could not rationally do this thing. That seems often true, for example, when Kant claims that we could not will that some maxim be a universal law.

could not possibly agree . . . G 429–30, my italics.

182 *Whether we could.. Rawls suggests* Rawls (2000) 100–91. A similar claim is made in Hill (1992) 45.

183 *To support this suggestion . . . 'precisely the same law'.* G 436. *Rawls therefore proposes..* Rawls (2000) 191 and 182–3.

Kant is a greater . . . even one new principle. C2, note on p.8. *what Herman calls.* Herman (1993) vii.

198 *As I have said, though our own preferences* . . . See 000 above.

210 *Having the resources . . . Metaphysics of Morals*, henceforth *MM* 454. See Wood (1999) 5–8, from whom I take this and the next quotation.

one can participate . . . LE 179 (27: 416).

In applying this version . . . difficult question. This may be the most important moral question that most rich people face. For four excellent discussions, see Murphy (2000), Mulgan (2001) Cullity (2004) and Pogge (2002). And see, and (I hope) respond to: www.givingwhatwecan.org

211 *The Consent Principle cannot, however, be . . . the supreme principle.* G 392.

212 *the Mere Means Principle* . . . Kant writes, 'all rational beings stand under the law that each of them is to treat himself and all others never merely as means but always at the same time as ends-in-themselves' (G 433).

213 *Kamm rejects* . . . Kamm gave me this objection in discussion. In Kamm (2007) 12–13 and her notes to these pages, Kamm gives an account of treating merely as a means which is very different from mine. On Kamm's account, whether we are treating someone merely as a means does not depend on our attitude to this person. And we might be treating someone *merely* as a means even if we are not treating this person *as a means*, or are sacrificing our life for this person's

sake. Though Kamm makes several plausible moral claims, she is not, I believe, describing the ordinary meaning of the phrase 'treat merely as a means'.

it is wrong for the rich . . . G 423. (Kant discusses someone for whom 'things are going well', and who 'contributes nothing' to those who are in need.)

For some moral belief . . . *Kant's view that cruelty.*. MM 443. But Kant also praises Leibniz for taking the trouble to place a worm back on its leaf after examining it under a microscope (C2 5:160).

216 *he who intends* . . . G 429.

217 *If we cannot appeal to (3)* . . . *our claim should be only* . . . For a further defence of these claims, see Section 42 below.

the Third Mere Means Principle. This is claimed, for example, by Nozick (1974) 31.

218 *We can next connect* . . . *We treat people as ends, Kant claims* . . . For example, 'rational beings . . . are always to be valued at the same time as ends, that is, only as beings who must be able to contain in themselves the end of the very same action' (G 429–30, my italics).

219 It might be claimed that, in *Bridge*, you would not really be *killing* me as a means . . . This objection is given by Bennett (1995) Chapter 11.

224 *If we coerce or deceive others* . . . O'Neill (1989) 111 and 114.

Korsgaard similarly writes . . . Korsgaard (1996) 347. Korsgaard may be intending only to describe Kant's view.

225 *To treat something as a mere means* . . . O'Neill (1989) 138.

On Kant's view, Korsgaard also writes . . . Korsgaard (1996) 142.

This claim implies . . . *Korsgaard also writes.*. Korsgaard (1996) 93.

226 *We are sometimes justified* . . . *For example, Rawls suggests* Rawls (TJ), 111. and 184. *Rawls also suggests.* Rawls (1999) 355 . . . . *That would make it trivial to claim* . . . Since Rawls makes no use of these proposed senses of 'right' and 'true', my remarks are no objection to his moral theory

227 *The phrase 'merely as a means'* . . . *in a special sense.* As when he claims that, if someone kills himself to avoid suffering, or gives himself sexual pleasure, this person thereby treat himself merely as a means (G 429, and MM, 425).

229 *This principle* . . . *According to Thomson.* . Thomson (1990) 166–168. Thomson adds: 'Where the numbers get very large, however, some people start to feel nervous. Hundreds! Billions! The whole population of Asia!'

231 *Turn next . . . On Thomson's view. .* Thomson (1990) 153. Thomson's claim is about an act that would save four people's lives; but she would apply it, I believe, to the saving of a single life.

233 *Kant writes, 'Every rational being . . . '* G 428.

Allen Wood suggests . . . Wood (1999) 152–5.

We can treat people rightly . . . Wood suggests. Wood (1999) 117.

234 *Some wrong acts . . . arrogance or contempt.* MM 462–8.

Wood calls (C). . . Wood (2002) 172 . . .. *provides, Wood claims . . .* Wood (1999) 155. *Wood points out* Wood (1999) 139. *last and longest book.* This book is *the Metaphysics of Morals*.

Kant's remarks . . . Wood himself concedes. Wood (2006) 346. *Kant writes that our duty.* MM 444 and 392 . . . . *define humanity. .* MM 423–5. *a mere thing.* MM 429–30.

235 *Wood suggests that . . .* Wood (1999) 154, and 371, note 32.

When Kant explains . . . Rawls calls it. Rawls (TJ) 31, note 16. *Wood in contrast..* Wood (1999) 141. *Herman suggests* Herman (1993) 208, 153.

Before we consider . . . in certain ways. I here follow Scanlon WWO Chapters 1 and 2.

237 *when we assert . . .* Moore (1903) 171. (At the end of this paragraph Moore seems to contradict this claim.)

Understanding something's value . . . WWO 99.

is primarily a matter . . . WWO 104.

239 *Scanlon rejects . . .* WWO 105.

We can next consider . . . what Kant calls dignity . . . value or worth. G 435–6.

240 *One such end . . . consists in good willing.* Herman writes, 'the domain of the good is rational activity and agency, that is willing' (Herman (1993) 213). . . . *the true vocation.* G 396–7.

Another end-to-be-produced . . . G 433 and 438.

A third such end . . . make them deserve. For references, see the notes about the start of Section 34 . . . . *everything possible to its production.* C2 119. . Kant's phrase is 'das höchste Gut', which literally means 'the Highest Good'. But Kant's phrase is misleading. As Kant himself points out, what he calls 'das höchste Gut' does not have a goodness that is *higher* than the goodness of a good will, but only the goodness that is most complete (C2, 111). My translation 'the Greatest Good' better suggests what Kant means, since this good is the greatest, not by being the highest, but by being the most complete.

There may be a fourth such end . . . G 428. *The Critique of Judgment* 442–3.

We can now return to. . Wood (1999) 133. Herman (1993) 238. Wood writes: 'Kant, however, proposes to ground categorical imperatives on the worth of any being having humanity, that is, the capacity to set ends from reason, irrespective of whether its will is good or evil' (Wood (1999) 120–1). Kant sometimes remarks that, by acting wrongly in certain ways, we would throw away our dignity, so that we had even less worth than a mere thing. But that is not really Kant's view.

241 *I have been ignoring . . . Herman writes.* Herman (1993) 213. *Wood even calls . . .* Wood (1999) 121. Thomas Hill similarly writes that, when Kant claims that persons are ends-in-themselves, that is a short way of saying that rationality in persons is such an end ((1992) 392).

On Kant's view . . . which he calls price. G 435 . . . *Cardinal Newman claims* Newman (1901) Vol I, 204.

242 *It might next be claimed . . .* Hill (1992) 50–57.

morality, and humanity. . G 435.

The word 'humanity' . . . Kant does condemn.. MM 427.

243 *In his claims about value. .* Herman (1993) 215, 210.

There is another. . what Herman calls 'moral standing'. Herman (1993) 129.

244 *The Highest or Greatest Good, Kant claims . . .* For example, Kant writes 'the greatest good of the world, the *Summum Bonum*, or morality coupled with happiness to the maximum possible degree' (LE 440 (27: 717). (See a note on page 240 above on why I translate such claims with the word 'greatest'.)

Everyone ought to strive . . . C2 125 (when Kant writes 'we', he means 'all of us'). He also writes, 'The production of the Greatest Good in the world is the necessary object of a will determinable by the moral law' (C2 122), and 'it is our duty to realize the Greatest Good to the utmost of our capacity' (C2 143 note).

the moral law commands me . . . C2 129.

245 *This ideal world . . . to Kant's ideal.* I am here following Kant, who writes, 'By this they meant the highest good attainable in the world, to which we must nevertheless approach, even if we cannot reach it, and must therefore approximate to it by fulfilment of the means' (LE 253 (27:482). He also writes: 'This Summum Bonum I call an ideal, that is, the maximum case conceivable, whereby everything is determined and measure. In all instances we must first conceive a pattern by which everything can be judged ' (LE 44 (27:247).

It would be best.. Some writers suggest . . . For example, Stephen Engstrom writes that, on Kant's view, the achievement of such proportionality would be 'the next best thing' (Engstrom (1992) 769).

Kant claims . . . As Kant assumes, we often have such beliefs . . . Kant for example writes that 'a rational and impartial spectator can never be pleased' at the sight of the happiness of a will lacking any trace of virtue, and that when such happiness is removed 'everyone approves and considers it as good in itself'. And he writes,, 'if someone who likes to vex and disturb peace-loving people finally gets a sound thrashing for one of his provocations . . . everyone would approve of it and take it as good in itself even if nothing further resulted from it' (C2 61).

247 *the concepts of good* . . . C2, 63–4.

All imperatives are . . . G 413.

248 *Kant may here seem . . . practically necessary, that is, good.* G 412.

K1: Good wills . . . In Kant's words, 'It is impossible to think of anything at all in the world . . . that could be considered good without limitation except a good will.' G 393. He goes on to say that this goodness is unsurpassed, and absolute.

The ancient Greeks, Kant claims.. C2 64.

249 *This may seem . . . his duty, at each instant* . . . R 72.

250 *the laws of morality* . . . LE 440–1 (27:717). This 'highest end' is the Greatest Good.

the strictest observance. . R 6: 7–8 (long note).

251 *This assumption is also fairly plausible, as Sidgwick later argued.* ME Book IV, Chapters III to V.

252 *Moore came close to accepting (C) . . . . we ought always to try to follow* . . . Provided, Moore adds, that these rules are both 'generally useful and generally practiced' (Moore (1903) 211–13). Moore denied that it would be best if there was most happiness; but this point is irrelevant here.

254 *When Hume discusses. . the whole scheme. . Enquiry* Appendix III, 256 (my emphasis). He also writes 'The result of the individual acts is here, in many instances, directly opposite to that of the whole system of actions; and the former may be extremely hurtful, while the latter is, to the highest degree, advantageous.' In the *Treatise* Hume writes: 'however single acts of justice may be contrary, either to public or private interest, 'tis certain, that the whole plan or scheme is highly conducive, or indeed absolutely requisite, both to the support of society, and to the well-being of every individual. 'Tis impossible to separate the good from the ill'. Book III, Section 2, 497 in Selby-Bigge.

254 *When Kant defends . . . I bring it about.* . SRL (8: 425–30).

255 *Thus a lie . . .* SRL 8: 426.

to promote the happiness . . . LE 388 (27:651).

If there is to be . . . Metaphysik L1, 28:337 From lectures given around 1778, cited in Guyer (2000) 94.

257 *Though we might try . . . Kant claims elsewhere that we have two ends . . .* MM 385–388. Our duty to promote our own virtue is the most important part of a wider duty to promote our own perfection, which includes our other abilities as rational beings. *On Kant's view . . .* See, for example, the quotations in a note on page 245 above. *These claims cannot be . . .* As Rawls writes: 'There is nothing in the CI-procedure that can generate precepts requiring us to proportion happiness to virtue' (Rawls (2000) 316.)

Reason does not approve . . . C1 640. He also writes: 'there is in the idea of a practical reason something further that accompanies the transgression of a moral law, namely its deserving punishment' (C2 37).

258 *According to determinists . . . Kant claims . . .* In Kant's words, 'he must also assume freedom of the will in acting, without which there would be no morals.' REV 8: 14

261 *Kant calls this compatibilist view 'a wretched subterfuge'.* C2 5: 96.

262 *unless we think . . .* REV 8: 13.

The practical concept . . . REV 8:13

265 *tantamount to handing it over to blind chance.* C2 5:95.

266 *To avoid this argument . . . are* not *events.* For a brilliant discussion of these questions, see Nagel (1987) Chapter VII.

267 *The human being must make . . .* R 6: 44.

Aristotle similarly writes . . . Nicomachean Ethics 1114a19; cf.1114b30 *seq.*

268 *(J) If our acts . . . could make us deserve to suffer . . .* See Nagel (1986) Chapter 7. In my statement of this argument, I partly follow Galen Strawson, who gives excellent versions of this argument in Strawson (1994) and Strawson (1998).

272 *We can deserve . . . But no one could ever deserve . . .* For discussions of the many questions raised by the belief that no one can deserve to suffer, see Sidgwick, ME, Chapter V, especially section 4, and Pereboom (2001) Chapters 5 to 7.

Though Kant makes . . . this alternative to Consequentialist reasoning'. Herman (1993) vii.

275 *Whether our acts are . . . Some of Kant's examples are . . .* C2 27. C2, 19. G 423. C2 34.

a universal permissive law. MM 453. Kant also refers to the universality of a law that everyone *could* act in certain ways (G 422, my emphasis).

277 *Some writers suggest that Kant means (D)*. . . See, for example, O'Neill (1989), 157. (O'Neill's view has since changed. See, for example O'Neill (1996) 59.)

Kant's actual formula is (F). . . Kant appeals to (F) when he discusses lying promises (G: 422), and in many other passages, some of which I mention below. I should admit that, when Kant discusses the maxim 'Kill myself to avoid trouble' (also G: 422), he does not appeal to (F). Kant does not claim that, if we all accepted this maxim, or believed such suicides to be permissible, that would make them impossible. That would be obviously false. Kant claims instead that, since the motive of such suicides is concern for our own well-being, and nature gave us this motive for the purpose of preserving our lives, there could not be a natural law by which this motive led us to destroy ourselves. If there were such a law, he writes, nature would 'contradict itself'. These claims do not support Kant's formula. Kant elsewhere claims that if everyone 'could end his life at will, such an arrangement would not be an enduring natural order.' This claim assumes that, if we all believed that suicide was permissible, humanity could not survive, since too many people would kill themselves before they were old enough to have children. That is most unlikely to be true.

Consider first.. As Herman points out . . . Herman (1993) 118–119.

278 *seems adequate* . . . Herman (1993) 119.

lies are usually . . . Korsgaard (1996) 136.

Were it to be a general rule . . . *LE* 232–3 (29:609).

279 *When Kant discusses* . . . *would not 'harmonize with itself'*. C2, 19.

These failures . . . *'Make lying promises* . . . *'* G 402–3, and 422.

the universality of a law . . . G 422.

In assessing this claim, as Rawls suggests . . . Rawls (2000) 169.

281 *Kant might have accepted* . . . *given his claim.* . SRL **8**: 425–30.

282 *Korsgaard partly answers* . . . Korsgaard (1996) 95.

283 *O'Neill proposes a weaker version* . . . O'Neill (1989), 133 and 215 and elsewhere.

This formula condemns . . . *O'Neill claims* . . . O'Neill (1989) 138–9 . . . *we 'undercut their agency'*. . . O'Neill (1989) 215–6.

(I), moreover . . . *competitive games with the aim of winning.* O'Neill (1989) 102–3.

284 *Korsgaard proposes* . . . Korsgaard (1996) 92–3 . . .. *to use tennis courts when they are least crowded.* Herman (1993) 138–9. . *pay the debts on our credit cards* . . . Blackburn (1998) 218 . . . *or give surprise parties.* Herman (1993) 141.

285 *Formula of Universal Law* . . . Of Kant's many versions of this formula, most take the form of commands, so that they could not be either true or false. But when Kant first proposes this formula, he writes 'I ought never to act except in such a way that I could also will that my maxim would become a universal law' (G 402).

Kant remarks . . . *As Herman points out* . . . Herman (1993) 123.

285 *In willing that some maxim* . . . *Kant sometimes claims* . . . He writes, for example, 'Maxims must be chosen as if they were to hold as universal laws of nature' (G 436). See also G 421, and C2 69–70.

286 *the Permissibility Formula* . . . For example, Kant writes 'could I indeed say to myself that everyone *may* make a false promise when he finds himself in a difficulty?' (G 403), and he refers to 'the universality of a law that everyone . . . *could* promise whatever he pleases with the intention of not keeping it' (G422,). Similarly Kant refers elsewhere to 'the law that everyone *may* deny a deposit which no one can prove has been made' (C2 27). And as I have said, Kant writes of a maxim's being 'a universal permissive law' (MM 453). (In all these quotations the emphases are mine.) This permissibility version of Kant's formula was suggested by Scanlon in unpublished lectures in 1983. See also Pogge (1998) Wood (1999) 80, and Herman (1993) 120–1.

the Moral Belief Formula . . . Kant does not explicitly appeal to this formula. But he is reported to have said, in lectures, 'you are so to act that the maxim of your action shall become a universal law, i.e. would have to be universally *acknowledged* as such' (*LE* 264 (27: 495–6). And Kant also writes: 'if everyone. . *considered* himself authorized to shorten his life as soon as he was thoroughly weary of it' (C2 69). (As before, the emphases are mine.)

Kant remarks that he is proposing . . . G 403.

288 *As before, Kant does not* . . . *As Rawls and Herman suggest* . . . Rawls (2000) 167 note 3.

289 *This objection can be partly answered* . . . O'Neill (1989) 85.

290 *These remarks do not fully* . . . See the discussion in Wood (1999) 103–5.

290 *It might be claimed* . . . *Rawls claims* . . . Rawls (2000) 187 *Kant similarly claims* . . . MM 455–7.

291 *Some writers suggest* . . . *But Kant often discusses..* C2 34.

293 *After considering this and other..* Wood (2006) 345, and Wood (2002) 172. Herman (1993) 104, 132. O'Neill (1975) 129, 125. See also O'Neill (1989) 130. Hill (2002) 122.

Because these people . . . Herman (1993) 117. O'Neill (1989) 86, 98, 103.

294 *to inform myself in the shortest and yet infallible way . . .* G 403.

someone feels sick of life . . . G 422.

he still has enough conscience . . . G 423.

he asks himself whether his maxim . . . G 421–2.

determines quite precisely . . . C2, 8 note. Kant also writes: 'all imperatives of duty can be derived from this single imperative', and 'These are a few of the many actual duties . . . whose derivation from the one principle is clear' G 424.

common human reason, with this compass in hand . . . G 404.

295 *Some writers suggest . . . (LN2). . .* O'Neill, Herman, Pogge, and Kagan all make or discuss proposals of this kind (O'Neill (1989) 87, 130–1; Herman (1993) 147–8; Pogge (2004) 56–58; and Kagan (2002) 122–127.

297 *In many cases, to give the morally relevant description . . . To give Sidgwick's example . . .* ME 202 note. Sidgwick claims that, though this revolutionary's intention was to kill the Czar, it would be false to say that he did not intend to kill the other people. It is better, I believe, that what this man was intentionally doing was acting in a way that he knew would kill many people.

298 *On my proposed revisions . . . This maxim is not a policy . . .* In Kant's longer statement, this maxim is: 'from self-love I make it my principle to shorten my life when its longer duration threatens more troubles than it promises agreeableness' (G 422). This maxim might be a policy, since we can often act in ways that shorten our lives. Smokers might do that every time they smoke. But Kant is here discussing a single act of suicide.

299 *Some people might question . . .* O'Neill (1975) 112 . . . . *But as Kant points out. .* G 424. O'Neill herself later writes 'this is not to say that in the actual world there is some contradiction in the thinker of each deceiver' (O'Neill (1989) 132).

O'Neill also suggests . . . O'Neill (1989) 87 . . .. *and O'Neill elsewhere claims . . .* O'Neill (1975) 112–117, and 124–143, and O'Neill (1989) 130. Herman makes similar claims in Herman (1993) Chapters 4 and 10.

307 *a maxim of promoting . . .* MM 393.

308 *One example is the maxim 'Have no children . . .* I take this example from Pogge (1998) 190.

309 *Pogge suggests* . . . Pogge (1998) 190. Pogge is here following an unpublished lecture given by Scanlon in 1983.

313 *is designed for* . . . Korsgaard (1996) 149. Korsgaard makes this claim not about Kant's Law of Nature Formula but about his Formula of Humanity. But this difference is irrelevant here.

314 *the problem is* . . . Hill (2000) 66.

317 *Since R2 is* also . . . *such as those theories* . . . For the best recent statement and defence of Rule Consequentialism that is known to me, see Hooker (2000).

318 *For some maxim to pass . . . who* don't *act on this maxim,* everyone else *does*. I am partly following some of Kagan's suggestions in Kagan (2002), and Kagan (1998) 231–5.

319 *This revision . . . to give a great deal.* See Hooker's discussion of this question Hooker (2000).

In revising Kant's . . . 'Most people won't'. As Herman notes, Herman (1993) Chapter 7.

321 *I want everyone else* . . . MM 451. I have changed 'benevolent' to 'beneficent', since that must be what Kant means.

It cannot be a universal law . . . G 430 note.

322 *As before . . . No one could will such a law, Kant claims* . . . G. 423.

325 *On Nagel's proposal* . . . Nagel (1970) 000, Hare (1963) Chapter 6, Harsanyi (1955), Rawls (TJ), *passim*.

326 *Return now to Kant's claim . . . judges could reject this argument.* As Leibniz pointed out, in Leibniz (1988) 56.

According to Kant's remaining objection . . . As Kant elsewhere suggests, however . . . . MM 450–1.

328 *Kant's Formula . . . the 'intuitive idea'* . . . O'Neill (1989) 94.

329 *Another objection to the Golden Rule* . . . Rawls (TJ), section 30.

334 *This question has some value . . . as Kant points out* . . . See, for example, G422.

335 *If this man made these claims . . . that would condemn such racist attitudes* . . . See Wood (1999) 3 and 7.

When Kant illustrates . . . Kant assumes . . . isn't such a law. See, for example, G422.

337 *the kind of case* . . . Korsgaard (1996) 101.

338 *It might be claimed . . . Nagel suggests* . . . Nagel (1991) 42–3.

many cases could occur . . . G 423 (my emphases).

339 *Rawls proposes* . . . Rawls writes: 'I believe that Kant may have assumed

that [our] decision . . . is subject to at least two kinds of limit on information. That some limits are necessary seems evident . . . ' (Rawls (2000) 175.) Of the two passages that Rawls cites, one is about the Formula of the Realm of Ends, and the other (discussed on p 176) seems to give no support to Rawls's reading.

On a third interpretation . . . Williams (1968) 123–131.

Scanlon proposes . . . WWO 170–1, and in unpublished summaries of lectures.

I ought never to act . . . G 402.

340 *the idea of the will* . . . G 432. And he refers to 'the concept of every rational being as one who must regard himself as giving universal law . . . ' But Kant never explicitly appeals to what everyone could rationally will. The phrase just quoted, for example, ends 'through all the maxims of his will' (G 434). If each person regards himself as giving laws through the maxims of *his* will, he is not asking which laws everyone could will. At several other points, when Kant seems about to appeal to what everyone could will, he returns to his Formula of Universal Law, telling us to appeal to the laws that we ourselves could will.

341 *In Scanlon's words* . . . WWO 171.

After considering some similar objections . . . Wood (1999) 172, Herman (1993) 104,132, O'Neill (1975) 125, 129.

344 *To justify certain principles . . . in a single round of concessions'.* Gauthier (1986) 133) . . . . . *there would be no set of principles that everyone ought rationally to choose.* See, for example, Sugden (1990).

345 *Some writers accept . . . to kill the native inhabitants . . . who are congenitally handicapped.* Gauthier (1986) 18 note 30, and 268. *But Gauthier rejects appeals* . . . Gauthier (1986) 269.

346 *I have rejected . . . 'a great many plain duties'.* Quoted in Gauthier (1986) 17. . And, as Rawls writes, 'to each according to his threat advantage is not a conception of justice' (TJ 134.)

Though Rawls also appeals . . . which he calls rightness as fairness. Rawls (TJ) sections 18–9. Though Rawls says little about this part of his view, he clearly regarded it as of great importance. He writes, for example: 'Perhaps I can best express my aim in this book as follows. During much of modern moral philosophy the predominant systematic theory has been some form of utilitarianism . . . they failed, I believe, to construct a workable and systematic moral conception to oppose it' (xvii). It is only in his contractualist account of morality that Rawls proposes an alternative to utilitarianism.

347 *If a person wants* . . . Rawls (TJ) 569.

Since Rawls's theory . . . would not have sufficient reasons to do what justice requires. Rawls (TJ) 575.

knowing that people are rational . . . Rawls (1996) 49.

348 *Since this definition . . .* Rawls (TJ) 184–5. Compare his claim 'in order that the parties can choose at all, they are assumed to have a desire for primary goods' in Rawls (1999) 266.

As a Contractualist, Rawls appeals . . . he adds a motivational assumption. In appealing to his formula, Rawls writes, 'we have substituted for an ethical judgment a judgment about rational prudence' (Rawls (TJ) 44). When we are behind the veil of ignorance, we are 'assumed to take no interest in one another's interests' (Rawls (TJ) 147. The people behind the veil of ignorance, he also writes, 'are prompted by their rational assessment of which alternative is most likely to advance their interests' (1999) 312). Rawls does *not* assume that, in the actual world, everyone is self-interested.

In choosing between principles . . . Rawls (TJ) 142.

349 *First, if everyone . . . would be hopelessly complicated'.* Rawls (TJ) 140.

Second, as Rawls points out . . . As Rawls writes, 'The combination of mutual disinterest and the veil of ignorance achieves the same purpose as benevolence. For this combination of conditions forces each person in the original position to take the good of others into account' (Rawls (TJ) 148). Rawls's comparison here is with *impartial* benevolence, and, as he points out, the veil of ignorance makes *partiality* impossible.

One of Rawls's main aims . . . Rawls (TJ) 22 . . . *As Rawls himself points out . . .* He writes, for example, 'the Utilitarian extends to society the principle of choice for one man'(Rawls (TJ) 28).

350 *Rawls is aware of this problem . . . On that assumption, Rawls claims . . .* Rawls (TJ) 165–6, RE.

there seem to be no objective grounds . . . Rawls (TJ) 168.

This remark treats . . . This formula, he writes . . . Rawls (TJ) 122 and 121, RE.

351 *aims for the thickest possible veil of ignorance . . .* Rawls (1999) 335–6. See also Rawls (TJ) Section 40.

This defence . . . . we could know enough to make a rational decision. As Rawls claims (TJ) 397 . . . *determination of alien causes'.* Rawls (1999) 265.

352 *Remember next that, as Rawls claims . . .* Rawls (TJ) 166 . . .. *does not, as he hopes, provide an argument . . . This objection to Rawls's argument I take from Nagel (1973).*

354 *decide solely on the basis . . .* Rawls (TJ) 584.

It is perfectly possible . . . Rawls (TJ) 29.

355 *for the contract view* . . . Rawls (1999) 174.

356 *does not allow the sacrifices* . . . Rawls (TJ) 4.

According to several writers.. See especially 'Equality' in Nagel (1979).

360 *Scanlon's Formula* . . . WWO 4–5 and Chapter 5.

This objection overlooks the fact . . . Scanlon appeals to this *Deontic Beliefs Restriction* (though not with this name) in WWO 4–5, 194, and 213–6.

361 *It would be unreasonable* . . . Scanlon (1997) 272.

362 *Like Rawls, Scanlon proposes his Contractualism* . . . WWO 215.

the implications of Act Utilitarianism . . . Scanlon (1997) 267.

367 *Another reply appeals* . . . *According to Intuitionists, Rawls writes* . . . Rawls (1999) *344* . . . *According to a different view, which Rawls calls Constructivism* . . . Rawls writes: 'the idea of approximating to moral truth has no place in a constructivist doctrine: . . . there are no such moral facts to which the principles adopted could approximate' (1999, 353.) *In Rawls's phrase, it's for us to decide* . . . Rawls (1999) Essays? 351

This method is, properly understood,. . . Scanlon (2003) 149.

370 *Scanlon now accepts* . . . See note 000 below.

371 *Pain is bad* . . . *Hume often uses* . . . As when he writes, 'Besides good and evil, or in other words, pain and pleasure . . . ' *Treatise* Book II, Section 19.

good or evil is, strictly speaking . . . C2, 60. Kant also claims that the principle of prudence, or self-love, is a hypothetical imperative, which applies to us only because we want future happiness. This claim assumes a desire-based view, ignoring our reasons to want our future happiness.

371 *When Kant claims* . . . *the reason-implying sense* . . . On one interpretation, the Stoics were making the interesting claim that pain is not bad even in this non-moral sense. See for example, Irwin (1996) 80. According to some other writers, the Stoics *were* merely claiming, like Kant, that pain is not morally bad.

372 *So does Ross* . . . Ross (1939) 272–284. (Though Ross makes these claims about pleasure, he intends them to apply to pain.)

As well as being bad for . . . *In Nagel's words* . . . Nagel (1986) 161 . . . *Many people believe* . . . Thomson, for example, writes: 'Suppose someone asks whether [something] would be a good event. We should reply 'How do you mean? Do you mean 'Would it be good *for* somebody?". We had better be told whether that is what is meant, or whether something else is meant . . . Consequentialism, then, has to

go' (Thomson (2003) 19). In making this last claim, Thomson assumes too quickly that her question can't be answered.

374 *(E) everyone ought always to do . . . what would make things go expectably-best.* In the sense explained in Section 21 above.

376 *When we apply Kant's formula . . . what would best fulfil our true needs . . .* See, for example, Rawls (2000) 173 – 6 and 232 – 4.

381 *the Nearness Principle . . .* For a partial defence of such a principle, see Kamm (2007) Chapters 11 and 12.

387 *deep attachments to other persons . . .* Williams (1981) 18.

389 *On some value-based objective theories . . . Scanlon's examples . . .* WWO 125.

404 *As Sidgwick argued, the answer is No.* ME, Book IV, Chapters III to V.

407 *Since these different versions . . . which we can use in different parts of our moral theory.* I discuss some of these questions in Sections 37 to 43 of RP. And see Kagan (2000) and (1998) Chapters 6 and 7.

410 *the supreme end.. The First Critique*, A 851 B 879.

If we conduct ourselves . . . Metaphysik L1,28:337, cited in Guyer (2000) 94.

414 *For the reasons that I earlier gave . . .* On pages 368 – 70.

415 *Remember next . . . One aim of such a theory, as Scanlon writes . . .* WWO 11.

420 *Case One, some whimsical Despot.* Credit for such examples may be due to Kavka (1986).

426 *Since Q1 and Q3 are different questions . . .* For a similar appeal to the difference between such questions, see Hieronymi, (2005) and (2006). Hieronymi does not, however, conclude that there are no state-given reasons.

The references for Appendix B are included in my notes on this appendix.

Bibliography

Barry, Brian (1989) *Theories of Justice, Volume 1* (Harvester-Wheatsheaf).

——(1995) *Justice as Impartiality* (Oxford University Press).

Bennett, Jonathan (1974) 'The Conscience of Huckleberry Finn', *Philosophy*, Vol.49, No 188 (April).

——(1995) *The Act Itself* (Oxford University Press).

Blackburn, Simon (1998) *Ruling Passions* (Oxford University Press).

Brandt, Richard (1979) *A Theory of the Good and the Right* (Oxford University Press).

——(1992) *Morality, Utilitarianism, and Rights* (Cambridge University Press).

Broad, C. D. (1959) *Five Types of Ethical Theory* (Littlefield, Adams, and Co).

Broome, John (1999) *Ethics out of Economics* (Cambridge University Press)

——(forthcoming) *Rationality Through Reasoning* (Oxford University Press).

Cullity, Garrett (2004) *The Moral Demands of Affluence* (Oxford University Press).

Cummiskey, David (1996) *Kantian Consequentialism* (Oxford University Press).

Darwall, Stephen (1983) *Impartial Reason* (Cornell University Press).

——(1992) 'Internalism and Agency', *Philosophical Perspectives, Vol. 6. Ethics*.

——(1992B) Allan Gibbard, and Peter Railton 'Toward Fin de Siecle Ethics: Some Trends', *The Philosophical Review* (January).

——(1996) *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (Oxford University Press).

Dean, Richard (2006) *The Value of Humanity in Kant's Moral Theory* (Oxford University Press).

Engstrom, Stephen (1992) 'The Concept of the Highest Good in Kant's Moral Theory', *Philosophy and Phenomenological Research*.

Enoch, David (2005) 'Why Idealize?', *Ethics*, 115 (July).

Findlay, John (1961) *Values and Intentions* (George Allen and Unwin).

Frankfurt, Harry (1988) *The Importance of What We Care About* (Cambridge University Press).

——(2004) *The Reasons of Love* (Princeton).

Gauthier, David: MA: *Morals by Agreement* (Oxford University Press, 1986).

——(1975) 'Reason and Maximization', *Canadian Journal of Philosophy*.

——(1984) 'Deterrence, Maximization, and Rationality', *Ethics*, 94.

——(1985) 'Afterthoughts', in *The Security Gamble*, edited by Douglas MacLean (Rowman & Allanheld).

——(1997) 'Rationality and the Rational Aim' in *Reading Parfit*, edited by Jonathan Dancy, (Blackwell).

Gibbard, Allan (1965), 'Rule Utilitarianism: Merely an Illusory Alternative?', *Australasian Journal of Philosophy*, 43.

Guyer, Paul (2000) *Kant on Freedom, Law, and Happiness* (Cambridge University Press).

Hare, Richard (1963) *Freedom and Reason* (Oxford University Press).

——(1997) 'Could Kant Have Been a Utilitarian?', in R. M. Hare *Sorting Out Ethics* (Oxford University Press).

Harsanyi, John (1955) 'Cardinal Utility, Individualistic Ethics, and Interpersonal Comparisons of Utility', *Journal of Political Economics*, 63.

Herman, Barbara (1993) *The Practice of Moral Judgment* (Harvard University Press).

Hieronymi, Pamela (2005) 'The Wrong Kind of Reason', *The Journal of Philosophy*, 102 no 9 (September).

——(2006) 'Controlling Attitudes', *Pacific Philosophical Quarterly*, 87, no 1 (March).

Hill, Thomas E. (1992) *Dignity and Practical Reason* (Cornell University Press).

——(2000) *Respect, Pluralism, and Justice* (Oxford University Press).

——(2002) *Human Welfare and Moral Worth* (Oxford University Press).

Hooker, Bradford (2000) *Ideal Code, Real World* (Oxford University Press).

Hume, David: *A Treatise of Human Nature*.

Kagan, Shelly (1998) *Normative Ethics* (Westview Press).

——(2000) 'Evaluative Focal Points', in *Morality, Rules and Consequences*, edited by Brad Hooker, Elinor Mason, and Dale E. Miller (Edinburgh University Press).

——(2002) 'Kantianism for Consequentialists', in *Groundwork for the Metaphysics of Morals*, Immanuel Kant, edited and translated by Allen Wood (Yale University Press).

Kahane, Guy (2004) *The Sovereignty of Suffering, Reflections on Pain's Badness* (Oxford University D. Phil. Thesis)

——(2009) 'Pain, Dislike, and Experience' *Utilitas*, Vol 21, No 3.

Kamm, Frances (1993) *Morality, Mortality, Volume One* (Oxford University Press).

——(1996) *Morality, Mortality, Volume Two* (Oxford University Press).

——(2000) 'Famine Ethics: the Problem of Moral Distance and Singer's Ethical Theory' in *Singer and his Critics*, edited by D. Jamieson (Blackwell).

——(2004) 'The new problem of distance in morality', in *The Ethics of Assistance*, edited by Deen K. Chatterjee (Cambridge University Press).

——(2007) *Intricate Ethics* (Oxford University Press).

Kant, Immanuel: Most of my references give the page numbers of the Prussian Academy edition, which are included in the margins in most English editions. I use the following abbreviations:

C1: *Critique of Pure Reason*.

C2: *Critique of Practical Reason*.

G: *The Groundwork of the Metaphysics of Morals*.

LE: *Lectures on Ethics*.

MM: *The Metaphysics of Morals*.

R: *Religion within the Limits of Reason Alone*.

REV: *Review of Schulz, Attempt at an Introduction:* 8: 10 – 14.

SRL: *On a Supposed Right to Lie from Altruistic Motives* 8: 425 – 30.

Kant: *Correspondence*, Translated and edited by Arnulf Zweig (Cambridge University Press, 1999).

Kavka, Gregory (1986) 'The Toxin Puzzle', *Analysis*, 43.

Kelly, Thomas (2003) 'Epistemic Rationality and Instrumental Rationality: A Critique', *Philosophy and Phenomenological Research*, Vol. LXVI, No.3, (May).

Kemp Smith, Norman (1915) 'Kant's Method of Composing the Critique of Pure Reason', *The Philosophical Review*.

Kerstein, Samuel (2002) *Kant's Search for the Supreme Principle of Morality* (Cambridge University Press).

Kolodny, Niko (2003) 'Love as Valuing a Relationship', *The Philosophical Review*.

——(2005) 'Why be Rational?', Mind, Volume 114.

——(2010) 'Which Relationships Justify Partiality? The Case of Parents and Children', *Philosophy & Public Affairs*, Volume 38 Number 1.

——(forthcoming) 'Which Relationships Justify Partiality? General Considerations and Problem Cases', in *Partiality and Impartiality*, edited by Brian Feltham and John Cottingham (Oxford University Press).

Korsgaard, Christine (1996) *Creating the Kingdom of Ends* (Cambridge University Press).

Kuehn, Manfred (2001) *Kant* (Cambridge University Press).

Leibniz (1988) *Political Writings* 2nd edition translated by Patrick Riley (Cambridge University Press, 1988).

Lewis, David (1985) 'Devil's Bargains and the Real World', in *The Security Gamble*, edited by Douglas MacLean (Rowman & Allanheld).

McClennen, Edward (1988) 'Constrained Maximization and Resolute Choice', *Social Philosophy and Public Policy*, *5*.

Moore, G. E. (1903) *Principia Ethica* (Cambridge University Press).

Mulgan, Timothy (2001) *The Demands of Consequentialism* (Oxford University Press).

Murphy, Liam (2000) *Moral Demands in Nonideal Theory* (Oxford University Press).

Nagel, Thomas (1970) *The Possibility of Altruism* (Oxford University Press).

——(1973) in his 'Rawls on Justice', *Philosophical Review* April, reprinted in Norman Daniels, *Reading Rawls* (Blackwell, 1975).

——(1979) *Mortal Questions* (Cambridge University Press).

——(1986) *The View from Nowhere* (Oxford University Press).

——(1991) *Equality and Partiality* (Oxford University Press).

——(1997) *The Last Word* (Oxford University Press).

Newman, Cardinal John Henry (1901) *Certain Difficulties Felt by Anglicans in Catholic Teaching* (Longman).

Nozick, Robert (1974) *Anarchy, State and Utopia* (Blackwell).

——(1993) *The Nature of Rationality* (Princeton).

O'Neill, Onora (1975) *Acting on Principle* (Columbia University Press).

——(1989) *Constructions of Reason* (Cambridge University Press).

——(1996) *Towards Justice and Virtue*, (Cambridge University Press),

Parfit, Derek: RP: *Reasons and Persons* (Oxford University Press, 1984, reprinted with some corrections in 1987).

——(1986) 'Comments' in *Ethics*, (Summer).

——(1997) 'Reasons and Motivation', *Proceedings of the Aristotelian Society, Supplementary Volume*.

Pereboom, Derk (2001) *Living Without Free Will* (Cambridge University Press).

Pogge, Thomas (1998) 'The Categorical Imperative', in *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, edited by Paul Guyer (Rowman and Littlefield).

Pogge, Thomas (2002) *World Poverty and Human Rights* (Polity).

—— (2004) 'Parfit on What's Wrong', the *Harvard Review of Philosophy* (Spring).

Rachels, Stuart (1998), 'Counterexamples to the Transitivity of Better Than', *Australian Journal of Philosophy*, 76, No.1 (March).

—— (2000) 'Is Unpleasantness Intrinsic to Unpleasant Experiences?' *Philosophical Studies*, Vol. 99, No. 2.

Rashdall, Hastings (1892) a review of Sidgwick's *Elements of Politics*, in the *Economic Review* 2.

Rawls, John: TJ: *A Theory of Justice* (Harvard University Press, 1971).

—— (1996) *Political Liberalism* (Columbia University Press).

—— (2000) *Lectures on the History of Moral Philosophy*, edited by Barbara Herman (Harvard University Press).

—— (2001) *Justice as Fairness* (Harvard University Press).

—— (2001B) *Collected Papers* edited by Samuel Freeman (Harvard University Press).

Raz, Joseph (2000) *Engaging Reason* (Oxford University Press).

Regan, Donald (1980) *Utilitarianism and Cooperation* (Oxford University Press).

Reid, Thomas (1983) *The Works of Thomas Reid* (Georg Olms Verlag).

Ridge, Michael (2006) 'Introducing Variable Rate Rule-Utilitarianism', *The Philosophical Quarterly* (April).

Ross, Sir David (1930) *The Right and the Good* (Oxford University Press).

—— (1939) *Foundations of Ethics* (Oxford University Press).

Ruskin, John (1903) *The Works*, edited by E.T.Cook and Alexander Wedderburn, Volume XI (London).

Scanlon, T.M. WWO: *What We Owe to Each Other* (Harvard University Press, 1998).

—— (1997) 'Contractualism and Utilitarianism', *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (Oxford University Press).

—— (2002) 'Replies', in *Social Theory and Practice* Vol. 28.

—— (2003) 'Rawls on Justification', in *The Cambridge Companion to Rawls*, edited by Samuel Freeman (Cambridge University Press).

—— (2003B) 'Value, Desire, and the Quality of Life', in *The Difficulty of Tolerance* (Cambridge University Press).

—— (2007) 'Structural Rationality', in *Common Minds*, edited by Geoffrey Brennan, Robert Goodin, and Michael Smith (Oxford University Press).

Scanlon, T.M. (forthcoming) *Being Realistic about Reasons* (Oxford University Press)

Schneewind, Jerome (1977) *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford University Press).

Schultz, Bart (2004) *Henry Sidgwick: Eye of the Universe, an Intellectual Biography* (Cambridge University Press).

Sidgwick, Henry: ME: *The Methods of Ethics* (Macmillan and Hackett various dates).

—— HSM: *Henry Sidgwick: A Memoir*, by A.S. and E. M. S (Macmillan).

—— (2000) *Essays on Ethics and Method* edited by Marcus Singer (Oxford University Press).

Smith, Michael (1994) *The Moral Problem* (Blackwell).

—— (2004) *Ethics and the A Priori* (Cambridge University Press).

Stratton-Lake, Philip (2004) *On What We Owe To Each Other* (Blackwell).

Strawson, Galen (1994) 'The Impossibility of Moral Responsibility', *Philosophical Studies* 75.

—— (1998) 'Free Will' in the *Routledge Encyclopaedia of Philosophy*, edited by E.Craig (Routledge).

Sugden, Robert (1990) 'Contractarianism and Norms', *Ethics* 100.

Suikannen, Jussi (2009) *Essays on Derek Parfit's On What Matters*, edited by Jussi Suikannen and John Cottingham (Blackwell).

Temkin, Larry (1987) 'Intransitivity and the Mere Addition Paradox, *Philosophy & Public Affairs*, 16, no. 2.

—— (forthcoming) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning* (Oxford University Press).

Thomson, Judith (1990) *The Realm of Rights* (Harvard University Press).

—— (2003) *Goodness and Advice* (Princeton University Press).

Williams, Bernard (1981) *Moral Luck* (Cambridge University Press).

—— (1985) *Ethics and the Limits of Philosophy* (Fontana).

—— (1995) 'Internal Reasons and the Obscurity of Blame', in his *Making Sense of Humanity* (Cambridge University Press).

—— (1995B) *Making Sense of Humanity* (Cambridge University Press).

—— (2003) *The Sense of the Past* (Princeton University Press).

—— (2006) *Philosophy as Humanistic Discipline* (Princeton University Press).

Williams, T. C. (1968) *The Concept of the Categorical Imperative* (Oxford University Press).

Wodehouse, P.G. (1952) *Pigs Have Wings* (Ballentine Books).

Wood, Allen (1999) *Kant's Ethical Thought* (Cambridge University Press).

——(2002) *Groundwork for the Metaphysics of Morals*, translated by Allen Wood, (Yale University Press).

——(2006) 'The Supreme Principle of Morality', in *The Cambridge Companion to Kant and Modern Philosophy*, edited by Paul Guyer (Cambridge University Press).

This page intentionally left blank

INDEX

Since pages 1 to 28 contain summaries of my main claims and arguments, I shall not here repeat some of the information in those summaries. This index gives page numbers for (1) my main discussions of various subjects, with numbers below 28 referring to summaries, (2) my scattered remarks elsewhere about these subjects, (3) my brief remarks about some other subjects, and (4) some other people's claims. Some of these entries overlap, either because their subjects overlap, or to reduce the number of entries that merely tell you to see some other entry.