DEREK PARFIT

On What Matters

VOLUME TWO

On What Matters

VOLUME TWO

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

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.

Impression: 1

First published 2011
First published in paperback 2013

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

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
Clays Ltd., St Ives plc

ISBN 978–0–19–968103–7 (Vol. 1)
978–0–19–968104–4 (Vol. 2)

Cover photograph, by the author, of the University Embankment in St. Petersburg.

On What Matters

VOLUME ONE

VOLUME TWO

This page intentionally left blank

Contents

VOLUME TWO

Preface

Though the first quarter of this volume is partly about Volume One, the remaining three quarters are entirely self-standing.

Of those who gave me comments on this volume, I was helped most by Robert Adams, Robert Audi, Selim Berker, Paul Boghossian, Laurence Bonjour, Nicholas Bostrom, Philip Bricker, John Broome, Ruth Chang, Eugene Chislenko, Roger Crisp, Garrett Cullity, Terence Cuneo, Jonathan Dancy, Cian Dorr, David Enoch, Kit Fine, Stephen Finlay, William Fitzpatrick, Alvin Goldman, Bob Hale, Michael Jubien, Shelly Kagan, Guy Kahane, Thomas Kelly, Samuel Kerstein, Patricia Kitcher, Niko Kolodny, Brian Leiter, William Lycan, Tim Maudlin, Brian McLaughlin, Charles Parsons, Ingmar Persson, Thomas Pogge, Peter Railton, Simon Rippon, Jacob Ross, Stephen Schiffer, Mark Schroeder, Russ Shafer-Landau, Peter Singer, Knut Skarsaune, Robert Stalnaker, Larry Temkin, Brian Weatherson, Ralph Wedgwood, and Timothy Williamson.

SUMMARY

PART FOUR COMMENTARIES

PART FIVE RESPONSES

CHAPTER 18 ON HIKING THE RANGE

65 *Actual and Possible Consent*

According to what I call Kant's *Consent Principle*, we ought to treat people only in ways to which they could rationally consent. Wolf suggests that, by interpreting Kant in this way, I abandon the Kantian idea of respect for autonomy, which often requires us to treat people only in ways to which they *actually* consent. But the Consent Principle does not abandon this idea, since people could seldom rationally consent to being treated in some way without their actual consent. And when such treatment would be wrong, this principle would not require such acts.

66 *Treating Someone Merely as a Means*

It is wrong to impose certain harms on people, Wolf claims, if we are treating these people merely as a means. It may be wrong, I claim, to harm people *as a means* even if we are *not* treating these people *merely* as a means. On this second view, harming people as a means would more often be wrong.

67 *Kantian Rule Consequentialism*

According to the Kantian Contractualist Formula, everyone ought to follow the principles whose universal acceptance everyone could rationally choose. This formula requires us, I argue, to follow optimific Rule Consequentialist principles. Wolf objects that everyone could rationally choose certain *non*-optimific *autonomy-protecting* principles. If everyone could rationally choose these principles, however, these principles must be optimific. But Wolf may be right to claim that everyone could rationally choose these principles.

68 *Three Traditions*

As Wolf claims, it would not be a tragedy if there is no single supreme moral principle. But it would be a tragedy if there is no single true morality.

CHAPTER 19 ON HUMANITY AS AN END IN ITSELF

69 *Kant's Formulas of Autonomy and of Universal Law*

The 'most definitive form' of Kant's supreme principle, Wood claims, is Kant's Formula of Autonomy. When revised in the way that is clearly needed, this formula becomes another version of my proposed Kantian Contractualist Formula.

70 *Rational Nature as the Supreme Value*

On Wood's interpretation of Kant's view, humanity or rational nature has the supreme value that both grounds morality and gives us our reason to obey the moral law. The supreme value of rational beings is not a kind of goodness, however, but a kind of moral status. This moral status could not be what grounds morality and gives us our reason to obey the moral law. Nor could such a ground be provided by the value of non-moral rationality. But Kant sometimes uses 'humanity' to refer to our capacity for morality and for having good wills. The supreme goodness of good wills might be the value that grounds morality. Wood's arguments against this view are not decisive.

71 *Rational Nature as the Value to be Respected*

Our acts are wrong, Wood suggests, when and because they fail to respect the value of non-moral rationality. Herman makes a similar suggestion. These suggestions seem open to strong objections. And respect for persons should be respect, not for their non-moral rationality, but for *them*.

CHAPTER 20 ON A MISMATCH OF METHODS

72 *Does Kant's Formula Need to be Revised?*

According to Kant's Formula of Universal Law, it is wrong to act on any maxim that we could not rationally will to be universal. This formula fails, I argued, because there are many maxims on which it is sometimes but not always wrong to act. Two examples are the Egoistic maxim 'Do whatever would be best for me' and the maxim 'Never lie'. We could not rationally will these maxims to be universal. But my imagined Egoist does not act wrongly when he acts on his maxim by keeping his promises, paying his debts, and saving a drowning child. Nor would it be wrong to act on the maxim 'Never lie' by telling someone the correct time.

Herman suggests that my Egoist does, in several senses, act wrongly. But Kant intends his formula to answer questions about which acts are wrong in the sense of being *contrary to duty*, and Kant would agree that my Egoist's acts are not in *this* sense wrong. And it would seldom be in this sense wrong to act on the maxim 'Never lie'. So Kant's Formula needs to be revised.

73 *A New Kantian Formula*

Kant's Formula might be claimed to tell us when acts are in certain other senses wrong. But this version of Kant's Formula would fail.

74 *Herman's Objections to Kantian Contractualism*

Herman earlier wrote that, despite a sad history of attempts, no one has been able to make Kant's Formula work. I argue that, if we revise Kant's Formula in two wholly Kantian ways, we can make this formula work.

Herman objects that, in applying both Kant's original formula and my proposed revision, I abandon one of the most distinctive parts of Kant's moral theory. I appeal to our reasons to care about our own and other people's well-being, and to the facts that give us other non-moral reasons to care about what happens. It is deeply un-Kantian, Herman suggests, to appeal to such reasons. That is not, I believe, true. And it is only by appealing to such reasons that we can make Kant's Formula work.

CHAPTER 21 HOW THE NUMBERS COUNT

75 *Scanlon's Individualist Restriction*

According to Scanlon's *Contractualist Formula*, we ought to follow the principles that no one could reasonably reject. Scanlon makes various claims about what are admissible grounds for rejecting principles. According to Scanlon's

> *Individualist Restriction*, in rejecting principles, we must appeal to their implications only for ourselves, or for other *single* people.

This restriction is given some support by Scanlon's appeal to the idea of justifiability to *each* person. But this part of Scanlon's view also has, I shall argue, some unacceptable implications.

76 *Utilitarianism, Aggregation, and Distributive Principles*

In proposing his Individualist Restriction, one of Scanlon's aims is to avoid certain Utilitarian conclusions. Utilitarians believe that it can be right to impose a great burden on one person, if we can thereby give small benefits to a large enough number of other people. Utilitarians go astray, Scanlon assumes, by adding together these people's benefits. On Scanlon's view, in such cases, the numbers don't count.

Scanlon, I suggest, misdiagnoses how Utilitarians reach such unacceptable conclusions. Their mistake is not their belief that the numbers count, but their belief that it makes no moral difference how benefits and burdens are distributed between different people. To illustrate this distinction, we should consider cases in which, if we don't intervene,

everyone will be equally badly off. In some cases of this kind, Scanlon's view would imply that we ought to benefit one of many people rather than giving to all these people a much greater total benefit that would be shared equally between them. If we are doctors, for example, we ought to lengthen a single person's life from 30 years to 70 rather than lengthening a million people's lives from 30 years to 35. That is clearly the wrong conclusion.

These cases show, I believe, that Scanlon ought to drop his Individualist Restriction. For Scanlon's Formula to apply successfully to such cases, Scanlon must allow that we can sometimes reasonably reject some principle by appealing to this principle's implications not only for us but also for the other people in some group. In the case that I have just described, each of the million people could reasonably reject any principle that did not require us to give them all five more years of life. These people could reasonably appeal to the facts that they are just as badly off as the single person, and that they together would receive a much greater total sum of benefits, which would also be more fairly shared between all these people.

Scanlon suggests that, if he gave up his Individualist Restriction, his view would cease to provide a clear alternative to Utilitarianism. That is not so. Rather than denying that the numbers count, Scanlon should return to a stronger version of one of his earlier claims, which we can call *the Contractualist Priority View*. People have stronger grounds to reject some principle, Scanlon should claim, the worse off these people are. This revised version of Scanlon's view would often conflict with Utilitarianism, and in ways that avoid implausible conclusions.

CHAPTER 22 SCANLONIAN CONTRACTUALISM

77 *Scanlon's Claims about Wrongness and the Impersonalist Restriction*

In his book, Scanlon claimed that his Contractualism gives an account of wrongness itself, or what it is for acts to be wrong. Scanlon should claim instead that, when acts are wrong in his Contractualist sense, that makes these acts wrong in other, non-Contractualist senses. He might, for example, claim that, when some act is disallowed by some

principle that no one could reasonably reject, this fact makes this act unjustifiable to others, blameworthy, and an act that gives its agent reasons for remorse, and gives others reasons for indignation. Scanlon now accepts that his Contractualist theory should take some such form.

According to Scanlon's

> *Impersonalist Restriction*: In rejecting some moral principle, we cannot appeal to claims about which outcomes would be impersonally better or worse, in the impartial reason-involving sense.

When Scanlon describes what it is for acts to be wrong in his proposed Contractualist sense, he can claim that, *by definition*, appeals to such impartial reasons are irrelevant. But if Scanlon claims that such acts are wrong in other senses, he could not defend his Impersonalist Restriction in this way. Nor could he defensibly claim that, when acts are wrong in his Contractualist sense, this fact has absolute moral priority over facts about what is impersonally better or worse. If Scanlon keeps his Impersonalist Restriction, he would have to retreat to the weaker claim that, when acts are wrong in his Contractualist sense, that makes these acts *prima facie* wrong in other senses. If Scanlon dropped this restriction, he could make the stronger claim that acts are wrong in other senses *just when*, and in part because, such acts are wrong in his Contractualist sense. If that were true, Scanlon's Contractualism would unify, and help to explain, all of the more particular ways in which some acts are wrong. That gives Scanlon a reason to make this bolder claim.

78 *The Non-Identity Problem*

Scanlon has other reasons to drop his Impersonalist Restriction. When he describes what we owe to others, Scanlon intends these *others* to include all future people. Many of our acts or policies affect the identity of future people, or *who it is* who will later live. We can often know both that

> (A) if we act in one of two ways, or follow one of two policies, we would be likely to cause some of the lives that are later lived to be less worth living,

and that

> (B) since it would be different people who would live these lives, these acts or policies would not be worse for any of these people.

We can ask whether and how (B) makes a difference. I have called this *the Non-Identity Problem*.

On one view, one of two outcomes cannot be worse, nor can one of two acts be wrong, if this outcome or act would be worse for no one. Even if such acts or policies would greatly lower the quality of future people's lives, we have no reason not to act in these ways.

According to another, better view, it would be in itself worse if some of the lives that will be lived will be less worth living, and we have reasons not to act in ways that would have such effects. If these effects would be very bad, and we knew that we could avoid them at little cost to ourselves, such acts would be wrong. This view could take two forms. According to

> *the No Difference View*: It makes no difference whether, because these future lives would be lived by the same people, these acts would be worse for these people.

According to

> *the Two-Tier View*: This fact does make a difference. Though we always have some reasons not to cause future lives to be less worth living, these reasons would be weaker if, because these lives would be lived by different people, these acts would not be worse for any of these people.

The Two-Tier View has some unacceptable implications. We ought to accept the No Difference View.

79 *Scanlonian Contractualism and Future People*

When applied to acts that affect future people, Scanlon's present view also has unacceptable implications. As before, Scanlon should drop his

Impersonalist Restriction, and allow us to appeal to impartial reasons. When our acts will affect future people, we must consider the different possible people who might later be actual. To explain why certain acts would be wrong, we must appeal to the better lives that would have been lived by the people who, if we had acted differently, *would* have later existed. We cannot defensibly claim that these acts are wrong because these people could reasonably reject any principle that permits such acts. If we acted in these ways, these people would never exist, and we cannot defensibly appeal to claims about what could be reasonably rejected by people who are merely possible. Since we cannot appeal to the *personal* reasons that are had by people who never exist, we should appeal to the *impartial* reasons that are had by people who do exist.

On this version of Scanlon's view, when we ask which are the principles that no one could reasonably reject, we would sometimes have to compare the moral weight of such conflicting personal and impartial reasons. We would have to use our judgment about which of these reasons would, in different kinds of case, provide stronger grounds for rejecting principles. As Scanlon points out, however, all claims about reasonable rejection require such comparative judgments.

Such judgments could go either way. When some act would make things go best, we would all have impartial reasons to reject principles that did not require such acts. In some cases, these impartial reasons would be decisive, and Scanlon's Formula would require us to do what would make things go best. In some other cases, some people could reasonably reject any principle that required such acts, since everyone's impartial reasons would be morally outweighed by these people's conflicting personal reasons.

There are, I have claimed, two reasons why Scanlonian Contractualism should allow us to appeal to impartial reasons. If we cannot appeal to such reasons,

> Scanlon's Formula could not be defensibly applied to many of the acts or policies with which we affect future people,

and, as I argued earlier,

> Scanlon could claim only that, when acts are wrong in his Contractualist sense, that makes these acts *prima facie* wrong in other, non-Contractualist senses.

If we can appeal to impartial reasons, Scanlon's Formula can be defensibly applied to all of our acts, and can be claimed both to tell us which acts are wrong, and to help to explain why such acts are wrong. Scanlonian Contractualism should, I believe, take this stronger form.

CHAPTER 23 THE TRIPLE THEORY

80 *The Convergence Argument*

When we apply the Kantian Contractualist Formula, I argued, it is only the optimific principles whose universal acceptance everyone could rationally choose. These principles might require us to impose a great burden on one person, for the sake of small benefits to many others. It may seem that, in some of these cases, the person who would bear this great burden could not rationally choose that everyone accepts these principles. Such cases would count against my claim that Kantian Contractualism implies Rule Consequentialism. This objection, I argue, fails.

I also argued that Kantian Rule Consequentialism could be combined with Scanlonian Contractualism. Scanlon objects that, even if the person who would be greatly burdened could rationally choose the optimific principles, this person could also reasonably reject these principles. In most cases, I believe, that is not so.

81 *The Independence of Scanlon's Theory*

In some cases, however, Scanlon's objection may succeed. Compared with Kantian Rule Consequentialism, Scanlonian Contractualism more strongly supports certain distributive principles, and may support some stronger principles. The three parts of the Triple Theory may also conflict in some other ways.

If there are such conflicts, that may seem to show that we should reject this theory. But that is not, I believe, true. All of our theories need to be revised. We are still climbing this mountain. And a team of mountaineers

may do better if they have different abilities and strengths, and they sometimes try different routes. It would be only at the mountain's peak that we, or those who follow us, would have all the same true beliefs.

PART SIX NORMATIVITY

CHAPTER 24 ANALYTICAL NATURALISM AND SUBJECTIVISM

82 *Conflicting Theories*

By asking certain questions, we can distinguish several kinds of meta-ethical view. We ought, I shall argue, to reject Non-Cognitivism and two forms of Naturalism. These views are close to Nihilism. Normativity is either an illusion, or involves irreducibly normative truths. I shall then defend one form of Non-Naturalist Cognitivism.

Words, concepts, and claims may be either normative or naturalistic. Some fact is natural if such facts are investigated by people who are working in the natural or social sciences. According to *Analytical Naturalists*, all normative claims can be restated in naturalistic terms, and such claims, when they are true, state natural facts. According to *Non-Analytical Naturalists*, though some claims are irreducibly normative, such claims, when they are true, state natural facts. According to *Non-Naturalist Cognitivists*, such claims state irreducibly normative facts.

On the rule-involving conception, normativity involves rules, or requirements, which distinguish between what is or is not *allowed* or *correct*. On the reason-involving conception, normativity involves reasons or apparent reasons. On the motivational, attitudinal, and imperatival conceptions, normativity involves actual or possible motivation, or certain kinds of attitude, or commands. The reason-involving conception is, I believe, the best.

83 *Analytical Subjectivism about Reasons*

When we claim that someone has an *internal* reason to act in some way, we mean that this act would fulfil one of this person's present fully

informed telic desires, or that after informed and procedurally rational deliberation this person would be motivated or would choose to act in this way. When we claim that someone has an *external* reason to act in some way, we use a fundamental, irreducibly normative concept that cannot be helpfully explained in other terms, but can also be expressed with the phrase 'counts in favour'. Though it is clear that we often have internal reasons for acting, some people believe that there are no external reasons. If we have both kinds of reason, as I believe, it is only external reasons that are important.

84 *The Unimportance of Internal Reasons*

If we used the words 'reason', 'should', and 'ought' in their internal senses, Subjectivism about Reasons would not be a substantive normative view, but a concealed tautology. If we used such words only in their *Naturalist internal* senses, we could not even have normative beliefs. If we used such words only in their *normative internal* senses, we could have some substantive normative beliefs, but we could not have distinct normative beliefs about what we have reasons to do, or what we should or ought to do.

85 *Substantive Subjective Theories*

For Subjectivists to make substantive claims, they should use these normative words in their external, irreducibly normative senses. The concept of an *internal reason* does no useful work.

86 *Normative Beliefs*

We can defensibly assume that normative words have such external senses, and can be used to make irreducibly normative claims.

CHAPTER 25 NON-ANALYTICAL NATURALISM

87 *Moral Naturalism*

It is sometimes claimed that, if normative and naturalistic concepts necessarily apply to all and only the same things these concepts must refer to the same property. That is not so.

Some normative concepts might refer to natural properties. But this does not show, as many Naturalists assume, that some normative claims might state natural facts. Some of these people ignore the important distinction between the properties that *make* acts right and the property of *being* right.

If Naturalism were true, Sidgwick, Ross, I, and others would have wasted much of our lives.

88 *Normative Natural Facts*

Some normative fact is *natural* in the *reductive* sense if this fact could be restated by making some non-normative, naturalistic claim. Naturalists believe that all normative facts are in this sense natural. Non-Naturalist Cognitivists believe that there are some irreducibly normative facts. We can ignore the question whether such normative facts might be, in some wider sense, natural facts.

If we use 'normative' in the rule-involving sense, we can defensibly claim that certain facts are both normative and natural. We can give Naturalistic accounts, for example, of what it is for certain acts to be illegal, dishonourable, or bad etiquette, or for the uses of certain words to be incorrect. Natural facts can also be normative in motivational and attitudinal senses. But no such facts can be normative in the reason-implying sense. There is a deep distinction between all natural facts and irreducibly normative reason-involving facts.

89 *Arguments from 'Is' to 'Ought'*

Searle argues that, if we accept certain natural, institutional facts, we must accept certain normative conclusions. Such arguments cannot succeed. We can recognize rule-implying normative facts but coherently deny that these facts give us any reasons.

90 *Thick-Concept Arguments*

Some writers similarly claim that, by appealing to *thick* normative concepts, such as *chaste* or *unpatriotic*, we can give sound arguments from *facts* to *values*. On one such argument, if we admit that someone has not committed any crime, we must accept that this person's punishment

would be retributively unjust, and therefore likely to be wrong. But we can coherently deny that any way of treating people could be either retributively just or retributively unjust. These *thick-concept arguments* make a serious meta-ethical mistake. We cannot derive moral conclusions from the meanings of our words. Just as we cannot prove that God exists by appealing to what we mean by 'God', we cannot give linguistic or conceptual proofs of any positive substantive normative truth.

91 *The Normativity Objection*

Normative claims could not state natural facts because such claims are in a separate, distinctive category. This objection to Normative Naturalism would also be accepted, though for partly different reasons, by those *Metaphysical* Naturalists who are Nihilists or Non-Cognitivists.

CHAPTER 26 THE TRIVIALITY OBJECTION

92 *Normative Concepts and Natural Properties*

When irreducibly normative concepts refer to natural properties, they do that by also referring to some other, normative property, so we should not expect that we could use such concepts to make normative claims that state natural facts.

93 *The Analogies with Scientific Discoveries*

Many Naturalists appeal to analogies with scientific discoveries, such as the discovery that water is H2O or that heat is molecular kinetic energy. When looked at more closely, such analogies partly fail.

94 *The Fact Stating Argument*

According to Non-Analytical Naturalists, any true normative claim states some fact that is both normative and natural. If this fact were natural, it could also be stated by some non-normative claim. If these claims stated the same fact, they would give us the same information. Since the non-normative claim could not state a normative fact, nor

could the normative claim. So such claims could not, as these Naturalists believe, state facts that are both normative and natural.

95 *The Triviality Objection*

When we say that we ought to act in some way, we are making a substantive claim, which might state a positive substantive normative fact. If these forms of Naturalism were true, such claims would not be substantive, but would be trivial. So these forms of Naturalism cannot be true.

Naturalists claim that, when some act would have certain natural properties, this fact is the same as this act's being what we ought to do. Such claims, some Naturalists believe, might tell us what we ought to do. That is not so. And what makes such claims seem informative also ensures that they could not be true.

For such normative claims to be substantive, they cannot merely refer to the same property in two different ways, but must tell us about the relation between two or more different properties, one of which is normative.

CHAPTER 27 NATURALISM AND NIHILISM

96 *Naturalism about Reasons*

The Triviality Objection also applies to Non-Analytical Naturalism about reasons.

97 *Soft Naturalism*

According to some Naturalists, though all facts are natural, we need to make some irreducibly normative claims. This view could not be true.

98 *Hard Naturalism*

Other Naturalists believe that, since all facts are natural, we could replace our normative concepts with naturalistic substitutes. This view is close to Nihilism.

CHAPTER 28 NON-COGNITIVISM AND QUASI-REALISM

99 *Non-Cognitivism*

According to *Non-Cognitivists*, normative claims are not intended to state facts, except perhaps in some minimal sense. Morality essentially involves certain kinds of desire, or other conative attitude. According to *Expressivists*, moral claims express such attitudes.

According to the *Humean Argument for Non-Cognitivism*, if moral convictions were beliefs, we might have moral convictions that did not motivate us. Since that is inconceivable, moral convictions cannot be beliefs, but must be desires or other conative attitudes. According to the *Naturalist Argument for Non-Cognitivism*, since moral claims could not state facts, but we can justifiably make such claims, these claims are not intended to state facts. According to the *Naturalist Argument for Nihilism*, since moral claims could not state facts, as they are intended to do, these claims are all false. We can reject these arguments.

100 *Normative Disagreements*

Expressivists cannot explain how we can have moral disagreements. We cannot disagree with other people's conative attitudes, or acts. Gibbard claims that, to understand our normative concepts and beliefs, it is enough to understand what is involved in deciding what to do, and in disagreeing with our own and other people's plans. That is not so.

101 *Can Non-Cognitivists Explain Normative Mistakes?*

Blackburn argues that, though our moral judgments express desires or other conative attitudes, these judgments and attitudes can be true or false, correct or mistaken. Expressivist Non-Cognitivists can thus be *Quasi-Realists*, who can claim all or nearly all that Cognitivists or *Realists* claim.

This ambitious project does not, I believe, succeed. Non-Cognitivists cannot explain what it would be for our moral judgments and conative attitudes to be correct or mistaken. Blackburn suggests that such

attitudes might be mistaken in the sense that we would not have these attitudes if our standpoint were improved in certain ways. But to explain the sense in which this standpoint would be improved, Blackburn would have to claim that, if we had this standpoint, our attitudes would be less likely to be mistaken. This explanation would fail because it would have to use the word 'mistaken' in the sense that Blackburn is trying to explain. We might similarly claim that our headaches might be mistaken in the sense that we would not have these headaches if we had some standpoint in which our headaches would not be mistaken. That would not explain a sense in which our headaches might be mistaken.

In defending Quasi-Realism, Blackburn also claims that some apparently external meta-ethical questions are really internal moral questions. That may be so. If we ask Expressivists whether it is really true that acts of a certain kind are wrong, they can consistently answer Yes. But we are asking what it would *be* for conative attitudes and moral judgments to be true or false, correct or mistaken. This is not an internal moral question. Though Blackburn suggests that he need not answer this question, that is not so.

To defend their Non-Cognitivist Expressivism, Quasi-Realists must claim that our conative attitudes cannot be correct or mistaken. To defend their Quasi-Realism, these people must claim that these attitudes can be correct or mistaken. These people must therefore claim that these attitudes both cannot be, and can be, correct or mistaken. Since that is impossible, no such view could be true.

CHAPTER 29 NORMATIVITY AND TRUTH

102 *Expressivism*

Gibbard's Expressivist account of the concept *rational* does not achieve Gibbard's aims, since it could not help us to decide how it is rational for us to live.

103 *Hare on What Matters*

In his account of the word 'matters', Hare denies that anything could matter.

104 *The Normativity Argument*

According to a third argument for Non-Cognitivism, normative truths would not really be normative, since no truth could answer a normative question. That is not so. Only truths could answer such questions.

CHAPTER 30 NORMATIVE TRUTHS

105 *Disagreements*

When we disagree with other people, we cannot rationally keep our beliefs unless we can justifiably assume that there is some asymmetry between us and these other people, making us more likely to be right. In most of my disagreements with other people, there are, I believe, such asymmetries. My main example will be Williams, the person from whom, in several disagreements, I have learned most. If there seemed to be no asymmetries between us, I could not rationally believe that, in these disagreements, it was Williams who was less likely to be right.

106 *On How We Should Live*

Socrates asked which kind of life is intrinsically best, by being the life that we have most reason to want to live. Williams denies that some ways of living could be, in this sense, intrinsically better than others. Rather than asking Socrates' question, Williams suggests, we should ask 'What do I basically want?'

107 *Misunderstandings*

When we claim that we have a reason to want something, we are using the phrase 'a reason' in the indefinable normative sense that we can also express with the phrase 'counts in favour'. Williams believes that the phrase 'a reason' has no such intelligible purely normative sense. When Williams makes claims about reasons, these claims are about what might motivate us. That is why Williams rejects the view that some lives are intrinsically better than others. If the phrase 'a reason' can have this purely normative sense, as I believe, Williams does not fully understand the view that he rejects. When people disagree about

whether some view is true, those who fully understand this view are more likely to be right.

108 *Naturalized Normativity*

Since Williams uses the phrase 'a reason' in a motivational sense, and he assumes that normativity involves reasons, Williams's normative claims are all psychological claims, which are at most weakly normative. Suppose I say: 'I *must* keep my promise to my wife. I *cannot* let her down.' This use of 'cannot', Williams claims, is a prediction. If I later give in to temptation, and break my promise to my wife, Williams might say: 'You were mistaken. As you found out, you didn't *have* to keep your promise. You *could* let her down.' But this remark would misunderstand my earlier claim. That claim was normative, and could be true whatever I later did.

Williams's view has unwelcome implications. Most of us believe, for example, that it would be wrong for anyone to torture other people for his own amusement. On Williams's view, given some sadist's motivations, this person may have no reason to act differently. This person's torturing of other people would not then be wrong.

109 *Sidgwick's Intuitions*

On Sidgwick's view, we have equal reason to be concerned about all parts of our conscious life. We have no reason, for example, to postpone some ordeal, when we know that this postponement would only make this ordeal worse. Sidgwick also claims that, from an impartial point of view, what happens to each person is equally important. Williams misunderstood these claims.

110 *The Voyage Ahead*

When I talked to Williams, I misunderstood his claims. I failed to see that these claims were psychological. I also misunderstood Mackie's claims. When Mackie denied that there are *objectively prescriptive values*, he was not denying a normative claim. Mackie meant that there are no normative beliefs that would *necessarily motivate* us. Since I knew these people well, I am puzzled and disturbed by our failures to understand each other.

111 *Rediscovering Reasons*

Hume is often assumed to be a Subjectivist, who believes that reasons for acting are given by facts about our present desires, and that we have no reasons to have our desires. But Hume's *stated* view is not Subjectivist, since Hume never discusses whether we have reasons for acting. Nor is Hume's *real* view Subjectivist. As many of his remarks show, Hume really believed that, as well as having reasons for acting, we have value-based object-given reasons to have particular desires, preferences, and aims.

Since Hume was really an Objectivist about reasons, that might be true of some other Humeans. The way a red hot iron feels, Mackie claims, gives him a powerful reason to try to end such pain. Mackie seems to be using the phrase 'a reason' in the motivational sense that is compatible with his Metaphysical Naturalism. But if Mackie had considered some of the distinctions I have drawn, he might have moved to a different view. The way a red hot iron *would* feel, Mackie might have believed, counts in favour of his trying to avoid this future pain. In coming to have this belief, Mackie would have abandoned both Naturalism and Subjectivism.

CHAPTER 31 METAPHYSICS

112 *Ontology*

In believing that some things matter in the reason-implying sense, I am believing that there are some irreducibly normative truths. That is denied by Metaphysical Naturalists, who believe that all properties and facts must be natural properties and facts. Irreducibly normative truths, these people assume, would involve the existence of strange metaphysical entities, which are too queer to be part of the fabric of the Universe.

On one widely held view, to be or to exist is to be actual, so there cannot be anything that is merely possible. If this *Actualist* view were true, much of our thinking would be undermined. We could never choose between different possible acts, or compare their possible outcomes, nor could we ever have reason to regret having acted as we did, since

there would never be something else that we could have done instead. On the true view, which we can call *Possibilism*, there are some things that are never actual, but are merely possible. We should draw some other distinctions between the kinds of thing that do or might exist, and their ways of existing, or the senses in which they exist.

113 *Non-Metaphysical Cognitivism*

There are some abstract entities, properties, and truths that are not mind-dependent, nor created by us. Some examples are mathematical entities and truths. Some people ask

> Q2: Do numbers really exist in a fundamental, ontological sense, though they do not exist in space or time?

Platonists answer Yes. *Nominalists* answer No. According to a third view, which we can call the *No Clear Question View*, Q2 is too unclear to have an answer.

There is another kind of view, which we can call *Non-Metaphysical Cognitivism*. On such views:

> (F) There are some claims that are, in the strongest sense, true, but these truths have no ontological implications.
>
> (G) When such claims assert that there are certain things, or that these things exist, these claims do not imply that these things exist in some ontological sense.

Some examples are arithmetical truths. This view is not a form of Possibilism. Compared with actual events, merely possible events have a *lesser* ontological status. When we consider entities like numbers, this distinction does not apply. These entities have *no* ontological status. They are neither actual nor merely possible, and neither real nor unreal.

Here is one way to argue that the phrase 'there are' and the word 'exist' have an important non-ontological sense. We can claim that

> (O) it might have been true that nothing ever existed: no living beings, no stars, no atoms, not even space or time.

Someone might say: '(O) could not have been true. If it had been true that nothing ever existed, there would have been the truth that nothing existed. That is a contradiction.' We can reply: 'Truths do not have to exist, or be real, in an ontological sense. Truths need only be true. If it had been true that nothing ever existed, there would have *been* this truth, but this truth would not have existed in an ontological sense.' Similar claims apply to many other abstract entities. Even if nothing had ever existed, there would have been prime numbers greater than 100. It would also have been true that things like rocks, stars, and living beings might have existed. There would have been these possibilities.

There would also have been some irreducibly normative truths. Compared with nothing's ever existing, it would have been much better if blissfully happy beings had existed, and it would have been much worse if there had existed conscious beings whose lives involved unrelieved suffering. According to *Non-Metaphysical Non-Naturalist Normative Cognitivism*—which I shall call *Rationalism*—there are some claims that are irreducibly normative in the reason-involving sense, and are in a strong sense true. These truths have no ontological implications. For such claims to be true, it need not be true that reason-involving properties exist either as natural properties in the spatio-temporal world, or in some non-spatio-temporal part of reality.

CHAPTER 32 EPISTEMOLOGY

114 *The Causal Objection*

It is often objected that, since we could not be causally affected by irreducibly normative properties, we could not have any way of knowing about them. But we can have other ways of knowing about non-natural properties and truths. Though our computers cannot be causally affected by numbers or their properties, their internal circuitry enables them to produce true answers to mathematical questions. God might have designed our brains so that we could answer such questions, and could also respond to reasons. If God does not exist, natural selection could explain how we came to have such brains. Just as cheetahs were selected

for their speed, and giraffes were selected for their long necks, human beings were selected for their rationality, which chiefly consists in their ability to respond to reasons. By responding to epistemic reasons, our ancestors were able to form many true beliefs which helped them to survive and reproduce.

115 *The Validity Argument*

When we ask how computers work, there are two kinds of event or fact that we need to explain. At the *micro-level*, there are many physical changes in the chips, circuits, and other small components of these computers. These events can each be fully explained by the laws of physics. But the laws of physics cannot explain the higher level fact that these computers reliably produce true answers to these many mathematical questions. This fact needs to be explained, since it would otherwise involve a highly implausible coincidence. These computers have this ability only because their calculations correspond to *valid reasoning*. Similar claims apply to us. Though the laws of physics may fully explain the neurophysiological events in our brains, these laws cannot explain how we can form so many true mathematical beliefs. We can form these beliefs only *because* we reason in valid ways. Though we cannot be causally affected by the property of validity, our mental processes involve a *non-causal response* to this validity. Metaphysical Naturalists believe that all properties and facts are natural. Validity is not, in the relevant sense, a natural property. Since the explanation of these mathematical abilities must appeal to non-natural truths about validity, we should reject this form of Naturalism. And though validity is not a normative property, these facts show that we might be able to respond, in similar non-causal ways, to non-natural normative properties and truths.

116 *Epistemic Beliefs*

The words 'probable', 'likely', and 'certain' can be used in non-normative, *alethic* senses. According to *Analytical Naturalists*, epistemic normative concepts can be explained in alethic terms, and refer to alethic properties. According to *Epistemic Rationalists*, these concepts are

irreducibly normative, and refer to irreducibly normative properties. According to *Non-Analytical Naturalists*, though these concepts are irreducibly normative, they refer to alethic properties. According to Rationalists, for example, when certain facts make it likely that P is true, that makes these facts have the different property of giving us some reason to believe P. According to Non-Analytical Naturalists, when certain facts make it likely that P is true, that's *what it is* for these facts to give us such a reason.

CHAPTER 33 RATIONALISM

117 *Epistemic Reasons*

Some normative skeptics argue:

> (1) Our normative epistemic beliefs were often advantageous, by causing us to have true worldly beliefs which helped us to survive and reproduce.
>
> (2) Because these normative beliefs were advantageous, natural selection made us disposed to have them.
>
> (3) These beliefs would have had the same effects whether or not they were true.

Therefore

> (4) These beliefs would have been advantageous whether or not they were true.

Therefore

> (5) Natural selection would have disposed us to have these beliefs whether or not they were true.
>
> (6) We have no empirical evidence for the truth of these beliefs.
>
> (7) We have no other way of knowing whether these beliefs are true.

Therefore

> We cannot justifiably believe that these beliefs are true.

We can call this the *Naturalist Argument for Normative Skepticism*. When we consider normative beliefs that are grounded on alethic beliefs about what is certain or likely to be true, we should accept (3), (4), (5), and (6). But we can reject (1) and (7), as similar claims about our modal beliefs help to show.

118 *Practical Reasons*

When this skeptical argument is applied to our practical and moral beliefs, we can respond in similar ways.

119 *Evolutionary Forces*

We have many practical and moral beliefs that were not produced by natural selection, or other evolutionary forces. Though we cannot have empirical evidence for the truth of these beliefs, we do not need such evidence. We have strong reasons to believe that we can have both epistemic and practical reasons, some of which are moral reasons. In defending these claims, however, there is a further challenge that we must meet.

CHAPTER 34 REACHING AGREEMENT

120 *The Argument from Disagreement*

When people deny that there are moral truths, many appeal to the facts of widespread moral disagreement, and to the cultural origin of many moral beliefs. Similar claims apply to other normative beliefs. In response to this argument, we should ask whether we can defend the claim that, in *ideal conditions*, we would nearly all sufficiently agree. According to this

> *Convergence Claim*: If everyone knew all of the relevant non-normative facts, used the same normative concepts, understood and carefully reflected on the relevant arguments,

and was not affected by any distorting influence, we would nearly all have similar normative beliefs.

Metaphysical Naturalists believe that there could not be any irreducibly normative truths. When we consider the Convergence Claim, we should ignore such meta-ethical beliefs. We should ask what these Naturalists would believe if they believed that there could be such truths. According to Error Theorists, for example, there could not be any moral truths, not even the truth that torturing children merely for fun is wrong. But these people would agree that, if any acts could be wrong, these acts would be wrong.

121 *The Convergence Claim*

There are many ways in which, when different people seem to have conflicting normative beliefs, these cases may not involve pure normative disagreements. These people may be considering borderline cases, or have conflicting non-normative or meta-ethical beliefs, or they may not know all of the relevant facts, or they may not understand the relevant arguments, or they may be using different concepts, or be affected by some distorting influence, or they may fail to realize that many normative truths are matters of degree, that many of these truths are very imprecise, and that some normative questions may not have answers. We can also plausibly believe that we have made normative progress. These facts do not show that, in ideal conditions, we would nearly all have sufficiently similar normative beliefs. But when we consider most actual disagreements, these cases do not count strongly against this prediction. We can add that, when we consider certain important questions, we *already* have sufficiently similar normative beliefs.

122 *The Double Badness of Suffering*

Nearly everyone believes that it is in itself bad to suffer, and that it is bad when people suffer in ways that they do not deserve. Though some people have seemed to deny these beliefs, they were either not really doing that, or were under the influence of some distorting factor, or both.

CHAPTER 35 NIETZSCHE

123 *Revaluing Values*

It may seem implausible to claim that, even in ideal conditions, we and Nietzsche would have had sufficiently similar normative beliefs. In defending the Convergence Claim, we cannot ignore Nietzsche, who is the most admired and influential moral philosopher of the last two centuries. Though Nietzsche sometimes denies that suffering is bad, and that happiness is good, that is not his real view; and Nietzsche's rejection of pity depended on false beliefs. Nietzsche's thinking was often distorted in certain other ways.

124 *Good and Evil*

The German word 'sollen' can be used both to express commands, such as 'Thou shalt not kill', and to express moral claims, such as 'You ought not to kill'. Some Germans have overlooked this distinction. Nietzsche assumes that morality consists of commands, and that only God would have sufficient authority to give such commands. Since God does not exist, Nietzsche concludes, there is nothing that we ought morally to do. If we believe that moral claims are not commands, Nietzsche's claims do not straightforwardly conflict with our beliefs about what we ought to do.

Nietzsche makes some other claims which might have led him to reject our beliefs. But Nietzsche contradicts many of these claims. When Nietzsche disagrees with himself, he does not clearly disagree with us. Other conflicts are less deep than they seem.

125 *The Meaning of Life*

Nietzsche's main questions were not about what we ought to do, or what is good or bad, but about *why* humanity exists, and whether the answer can give meaning to our lives. When Nietzsche lost his belief in God, he sometimes believed that we were created by Life or Nature to achieve some purpose. When Nietzsche recognized that Life or Nature had no such purpose, he hoped that we ourselves could create new values, thereby giving our lives meaning. Since Nietzsche's normative concepts

were not reason-involving, but imperatival or command-implying, his attempt to avoid Nihilism failed.

CHAPTER 36 WHAT MATTERS MOST

126 *Has It All Been Worth It?*

The badness of suffering casts doubt on the goodness of the world. When we consider the horrors of the past, we can ask whether human history has been worth it. Some believe the answer to be No. On this view, it would have been better if no human beings had ever existed.

127 *The Future*

Even if the past has been in itself bad, the future may be good, and this goodness might outweigh the badness of the past. Human history would then be, on the whole, worth it. In deciding what we ought to do, we can ignore the badness of the past. Even if history could not be, on the whole, good, the future might be good. Since the further future might be very good, what now matters most is that we avoid ending human history, by overheating the atmosphere, or in other ways. If there are no rational beings elsewhere, it may depend on us and our successors whether it will all be worth it, because the existence of the Universe will have been on the whole good.

APPENDICES

APPENDIX D WHY ANYTHING? WHY THIS?

Why does the Universe exist? There are two questions here. First, why is there a Universe at all? It might have been true that nothing ever existed: no living beings, no stars, no atoms, not even space or time. When we think about this possibility, it can seem astonishing that anything exists. Second, why does *this* Universe exist? Things might have been, in countless ways, different. So why is the Universe as it is?

Many people have assumed that, since these questions cannot have causal answers, they cannot have any answers. Some therefore dismiss these questions, thinking them not worth considering. Others conclude that they do not make sense.

These assumptions are, I believe, mistaken. Even if these questions could not have answers, they would still make sense, and be worth considering. Nor should we assume that answers to these questions must be causal. Even if reality cannot be fully explained, we may still make progress, since what is inexplicable may become less baffling than it now seems.

APPENDIX E THE FAIR WARNING VIEW

Though punishments cannot be just or unjust in the desert-implying sense, such penalties can be fair or unfair. But when we justifiably impose fair punishments, we should greatly regret what we are doing.

APPENDIX F SOME OF KANT'S ARGUMENTS FOR HIS FORMULA OF UNIVERSAL LAW

Kant argues:

> All principles or imperatives are either *hypothetical*, requiring us to act in some way as means of achieving some end that we have willed, or *categorical*, requiring us to act in some way as an end, or for its own sake only, rather than as a means of achieving any other end.
>
> Categorical imperatives impose only a formal constraint on our maxims and our acts, since these imperatives require only conformity with the universality of a law as such.

Therefore

> There is only one categorical imperative, which requires us to act only on maxims that we could will to be universal laws.

Kant's premises are false, and, even if they were true, Kant's conclusion would not follow. Kant also argues:

> (1) When our motive in acting is to do our duty, we must be acting on some principle whose acceptance motivates us without the help of any desire for our act's effects.
>
> (2) For some principle to have such motivating force, it must be purely formal, requiring only that our acts conform with universal law.
>
> (3) Such a principle must require that we act only on maxims that we could will to be universal laws.

Therefore

> This requirement is the only moral law.

Premises (2) and (3) are false. Kant gives other arguments that seem to fail.

APPENDIX G KANT'S CLAIMS ABOUT THE GOOD

In several passages, Kant seems to overlook the sense in which happiness and suffering are non-morally good and bad, and to ignore our other non-moral reasons to care about what happens.

APPENDIX H AUTONOMY AND CATEGORICAL IMPERATIVES

According to Kant's *Autonomy Thesis*, we are subject only to principles that we give to ourselves as laws, and obligated only to act in conformity with our own will. This thesis seems to be either indefensible or trivial. In his claims about heteronomy, Kant seems to conflate two very different things: motivation by desire, and strongly categorical requirements.

APPENDIX I KANT'S MOTIVATIONAL ARGUMENT

Kant seems to argue:

> True moral laws must be both universal and normatively categorical, applying to all rational beings whatever they want or will.
>
> No principle could be such a moral law unless the acceptance of this principle would necessarily motivate all rational beings.
>
> No principle could have such necessary motivating force, and thus be able to be a true moral law, unless this principle can motivate us all by itself, without the help of any desire.
>
> Only Kant's Formal Principle has such motivating force.
>
> There must be some true moral law.

Therefore

> Kant's Formal Principle is the only true moral law, and is thus the supreme principle of morality.

This argument could not succeed.

APPENDIX J ON WHAT THERE IS

There are some things that are actual, and others that are merely possible. Some Actualists claim that, when we decide what to do, we are not choosing different possible acts, but merely choosing which way in which we shall act. But if I act in one way, by saving your life, this act would be one future event. If instead I let you die, this act would be a different event. There are here two possible events, one of which would be merely possible. Such events exist, however, in a different, ontologically thinner sense. There are also various other entities and truths that exist in a non-ontological sense. These include some irreducibly normative truths.

PART FOUR

COMMENTARIES

This page intentionally left blank

Hiking the Range

Susan Wolf

On What Matters is a *tour de force*—a fast-paced ride across the territory of philosophical ethics, filled with challenging and provocative discussions of an astonishing number of philosophical positions and problems. All of these discussions are at least loosely presented as being in the service of the search for the supreme principle of morality. To top it off, Parfit concludes the first volume of this work with what he takes to be a good candidate for such a principle—the Kantian Contractualist Formula, which tells us that

> Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose (Volume One, *342*).*

From this principle, he argues, it follows that everyone ought to follow the principles that are optimific, thus yielding the view he calls Kantian Rule Consequentialism (*411*).

One way to approach the book is to see it as displaying the thought of one philosopher picking and choosing what he takes to be the best and most insightful aspects of several different ethical theories, and putting them together to come up with a different view of his own. As such, it represents a fine way to do moral philosophy—not the only way, but a fine way—and there is much in the particular view that Parfit arrives at, as well as in the particular assessments of other views which he offers and defends along the way, that I find attractive. Another, even more ambitious way of reading the book, however, is suggested in the way Parfit presents his thought, and especially by the concluding remarks of Volume One, which give the volume's final section its

* Page numbers in italics refer to Volume One.

name. As he notes, Kantian Contractualism has a claim to being at once Kantian, contractualist and (at least one-third) consequentialist. Though these three great moral philosophical traditions are often seen as expressing deeply contrasting and mutually incompatible ethical perspectives, Parfit suggests that the plausibility of his proposed formula, in conjunction with the arguments by which he has arrived at it, gives us reason to see these traditions differently. 'It has been widely believed that there are . . . deep disagreements between Kantians, contractualists, and consequentialists,' he writes. 'That, I have argued, is not true. These people are climbing the same mountain on different sides' (*419*).

The suggestion, if I am interpreting it correctly, is 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.

It is this suggestion — or, as one might say, this ambition — with which I shall take issue in this paper. The suggestion has both a metaethical and a normative aspect. Metaethically, Parfit's work seems to embody the assumption that there are very strong reasons for wanting or hoping for there to be a single supreme, and presumably universal and timeless, principle of morality, to which all other moral principles would be subsidiary. Parfit shares this assumption with many if not all of the major figures associated with the traditions he claims to combine. However, insofar 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 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.

Like Parfit, I see the Kantian, consequentialist, and contractualist traditions as each capturing profound and important insights about value. Using Parfit's metaphor, we might say that each contains, not just a grain, but rather something more like a mountain of truth. Each makes a profound contribution to our appreciation of what we have

reasons to do and to care about, and to what morality should express, protect, and promote. For Parfit, appreciation of the different evaluative perspectives poses a challenge which he aims in this book to meet: to unify, systematize, or otherwise combine the insights gleaned from these perspectives to reach a single coherent moral view that can guide our actions in a way that is free of moral remainders and normative tensions. Though I think I understand the wish to reconcile the different traditions and transform their ideas into a single unified whole, I am less gripped by it than many other moral philosophers.

Of course there are reasons for hoping that there is, or wishing that there were, a single supreme principle of morality, and if it turns out that there is such a principle, it would be good to know what it is. However, in the absence of a particular metaethical account of what morality is, there is no reason to assume that there will be such a principle, and it would not be a moral tragedy if it turned out that morality were not so cleanly structured as to have one. Moreover, on my own understanding and assessment of the contributions of the Kantian, consequentialist, and contractualist traditions, the values these different theoretical stances express continue to elude such complete unification. As it seems to me, there are fairly frequent occasions when the world presents us with choices for which there is no easy or unique moral answer: there are good moral reasons to favor one alternative and good moral reasons to favor another—and no overarching or further reason to settle the issue between these alternatives without begging the question.

There may be reasons, at the level of concrete social practice, to adopt a conventional ordering of values or a decision procedure that has the effect of a compromise between the realization and expression of competing values. Still, it seems to me important that in moral philosophical contexts, compromises and conventions be recognized as such. We should not allow our interest in reaching agreement on universal principles, much less on a single fundamental principle, to distort our understanding of the individual values on which such principles are based or to suppress our acknowledgment of the tensions that may exist among them.

In any case, it seems to me that there *are* tensions in our common moral thought at least some of which are reflected in the differences

among Kantian, contractualist, and consequentialist perspectives. (I thus share the common view, which Parfit rejects, that these views are in deep disagreement.) As Parfit critically interprets and revises Kant's theory so as to reconcile it with contractualist and consequentialist insights, some of these tensions get lost, and some of what seems to me most compelling and distinctive about Kant's own moral perspective gets diluted.

In this paper, I shall focus especially on one such tension, which is frequently associated with the difference between Kantian and consequentialist ethics, namely, that between respect for autonomy and concern for optimific results. It will be instructive to see how Parfit's transformation of Kant's theory makes this tension disappear, and what might be said in favor of a different interpretation of Kant. Following that, I will also have some things to say about tensions between contractualist and noncontractualist theories, and about the importance (or unimportance) of finding a supreme principle of morality.

Not being a Kant scholar, I do not wish to make claims about what Kant really meant or what is truly Kantian in spirit. My concern is normative rather than interpretive. Still, it seems to me there is an interpretation of Kant, or, at least, a moral perspective inspired by Kant, according to which some of Parfit's suggested revisions take us away from rather than toward a more persuasive moral theory.

Respect for Autonomy

Though Kant himself used the term 'autonomy' to refer to a metaphysical property that Parfit and probably most contemporary philosophers don't believe humans possess, there is a nonmetaphysical understanding of the term that still retains much of what Kant was concerned with. Specifically, we may understand autonomy to refer to the possession of practical reason, which gives its possessor the ability to think and decide for herself what to value, what to do, and how to live. To say that we should respect autonomy, or that we should respect people as autonomous beings, is to say that we should take this feature of persons to heart, as calling for a response, limiting our behavior toward them in certain ways, and perhaps demanding types of behavior in

others. Roughly, the idea is that respecting autonomy involves honoring people's ability to govern their own lives, refraining from interfering with their choices for themselves, and from imposing burdens on them that they would not themselves endorse. The tension between this value and concern for good results stems from the fact that people do not always know what is good, even for themselves, and they do not always know or care very much about what is good for the world at large. This tension is evident in our possibly mixed reactions to cases of paternalism, as well as in our reactions to cases like Parfit's *Bridge* (*218*) and *Means* (*201*), in which one must choose whether to impose a burden on one person (or group) in order to save another person (or group) from even greater harm. Arguably, respect for autonomy urges us to let people decide for themselves whether they want to sacrifice their own welfare for the greater good. If they do not so choose, respect for their autonomy urges us to leave them alone.

In his writings, Kant's respect for autonomy, even of this nonmetaphysical sort, is quite pronounced, and seems to many readers built into his injunction never to treat a person as a means only. It is even more obviously connected with the importance of consent in legitimating one's treatment of another human being. Yet Parfit's interpretation of Kant's Consent Principle and his interpretation of what it is to treat someone as a mere means seem to leave respect for autonomy behind. Parfit's derivation of Kantian Consequentialism from Kantian Contractualism seems also to reflect a lack of appreciation for the value of respect for autonomy. Let us see how one who is deeply impressed with that value might respond to Parfit's arguments.

Consent

We may begin with Parfit's discussion of Kant's claims about consent, which Parfit restates as '(A) It is wrong to treat people in any way to which they cannot possibly consent' (*180*). As Parfit notes, on at least one natural interpretation of (A), the claim is too strong to represent what might most charitably be understood as Kant's considered view.[1]

[1] Parfit objects, more specifically, to Korsgaard's and O'Neill's interpretation of Kant's claims, according to which '(B) It is wrong to treat people in any way

It is also too strong, we might add, to represent a reasonable view of a constraint that is meant to embody respect for autonomy. Situations may arise, for example, when one must take action but cannot obtain consent because the person is unconscious, or unable to communicate, or because there is no time to stop and ask. There may be other cases when a person explicitly refuses to consent to action because he is in the midst of a psychotic episode or is seriously misinformed. In cases like these, taking action to save someone from serious harm in the absence of consent seems neither wrong nor disrespectful. If one is reasonably assured that the person *would* consent if he were conscious, in his right mind, and so on, that would seem enough to make the action meet the standards the spirit of the consent principle demands.[2]

Parfit's own suggested redescription of Kant's claim might appear at first glance merely to be a way to build these sorts of qualifications into the statement of the position. According to Parfit, we should understand Kant's Consent Principle to say 'It is wrong to treat people in any way to which they could not *rationally* consent' (*181*). However, Parfit's version takes us much further from the original idea of consent than first meets the eye. Because Parfit employs a value-based theory in his interpretation of reasons and rationality, and because his suggested principle concerns what a person *could* rationally consent to, Parfit's version of the Consent Principle might allow us to do things to someone even if we had no reason whatsoever to suppose that the person affected by it *would* consent to it—indeed, it would allow us to do things to a person even if he explicitly refuses to consent to it under conditions of full rationality and information.[3]

to which they cannot possibly consent, because we have not given them the possibility of giving or refusing consent' (*179*).

[2] This is meant only as a rough statement of a plausible revision to the Consent Principle that would not violate the spirit of respect for autonomy. It would need to be fine-tuned, however. A Jehovah's Witness who refuses life-saving medical treatment because he believes such treatment would be against God's will, might be thought by his doctor to be seriously misinformed, yet it is arguably incompatible with respect for the patient's autonomy in this case to waive the consent condition despite the doctor's (well-grounded) belief.

[3] Parfit is careful to point out that the Consent Principle is not offered as the supreme or sole principle of morality. As he notes, 'The Consent Principle

Consider, for example, *Means*, the variant of Parfit's *Earthquake* case, in which you may save White's life, but only by moving Grey in such a way that he would lose his leg. (Both are trapped in the wreckage so that neither can move themselves.) According to Parfit's wide value-based theory of reasons, Grey could rationally choose that you move him, causing him to lose his leg in order to save White's life, but he could also rationally choose that you leave him alone, thus letting him keep his leg, but allowing Grey to die. Since Parfit's Consent Principle requires you to restrict your action to what affected parties *could* (but not necessarily would) rationally choose, that principle permits you either to move Grey or not, at least so far as Grey is concerned.

We may further imagine, however, that you happen to know Grey, and know that he is not the kind of person to voluntarily sacrifice a limb to help a stranger. Just last week, we may suppose, he refused to donate his kidney to help save his own brother. Indeed, we may imagine that Grey, though trapped in the rubble, is still alert enough to size up the situation he and White are in, and is yelling at you, 'Stay away from me, you self-righteous, do-gooding consequentialist.'

I do not want to argue one way or the other about what one *ought* to do in a situation like this. There seems to me to be something to be said for refraining from moving Grey if he refuses to consent, and something to be said for moving Grey anyway, in order to save White's life. But if one chooses the latter over Grey's protests, it seems odd to say that one has satisfied a Consent Principle.[4] It seems much more natural

does not claim that acts are wrong *only if* people could not rationally consent to them . . . This principle allows that acts can be wrong in other ways, or for other reasons.' My point is simply that Parfit's Consent Principle *itself* does not condemn or otherwise discourage treating someone in a way to which he, under conditions of full rationality and information, has explicitly refused consent.

[4] There is a way of thinking about this case in which it might satisfy a Consent Principle: if one thinks the level at which consent principles should operate is the level of general principles rather than particular actions, it is possible that under certain plausible conditions, Grey would consent to a principle that allowed you to move his leg, even though at the moment of crisis, he does not care about principles, and does not consent to the particular action. I'll discuss this very significant complication later in the paper.

to think of this as a case in which the value of restricting oneself to what someone would consent to is overridden by the value of saving a life.

Insofar as respect for autonomy—understood, as I suggested, as an injunction to try, so far as possible, to let a person decide for herself what to do—is the value motivating a principle that appeals to consent, Parfit's own Consent Principle is wholly beside the point. Respect for Grey's autonomy would require us to take Grey's values and choices into account, or, failing that, to take into account the values Grey would have and the choices Grey would make if he were in a position to consider the relevant questions, with relevant information, and so on. The fact that Grey *could* choose to give up his leg—that it would not be irrational were Grey to do so—has very little to do with Grey himself, and nothing at all to do with Grey's exercise of his own practical reason.

In his chapter on consent, Parfit considers some versions of the Consent Principle—namely, the Choice-Giving Principle and the Veto Principle—that would require a person to refrain from actions to which the affected party, under conditions of rationality and information, would not consent. He rejects these principles, at least partly because it is clear that if one were to try to restrict one's actions to ones to which all affected parties *would* consent (under conditions of full rationality and information), one would fail in one's aspirations. Frequently, we would find that one party would only consent to one action, while another party would only consent to another. Grey might not consent to losing his leg; White might not consent to losing his life. In Parfit's terms, such principles would fail to meet the Unanimity Condition (*188*).

For Parfit, searching as he is, for a supreme principle of morality, and, even short of that, for principles that will give us decisive reasons for narrowing down the range of permissible actions, the Unanimity Condition will understandably carry a lot of weight. To meet this condition, one must move beyond the interpretations of the Consent Principle that would forbid actions that would affect parties in ways to which they would rationally not consent. One way to do this, connected to philosophical positions Parfit considers later in the book, would be to 'move up a level' by asking not which particular acts a person would consent to, but rather what general principles of action would be agreed on under relevant conditions. In his discussion of the Consent

Principle, however, Parfit seems to take a different path—namely that of a restriction based on what people *could* rationally consent to, rather than on what they *would* rationally consent to.

The problem with this suggestion, as I have argued, is that it leaves what may be considered the moral point behind a consent principle behind. It leaves consent behind, and the respect for autonomy, from which the value of consent might be thought to derive. If one is concerned in the first instance not in formulating a supreme or decisive moral principle, but rather in registering and articulating important (but possibly competing) moral considerations, the need for unanimity would not be allowed to transform one's principles in this way.

Treating Someone as a Means Only

In any event, the search for a single comprehensive principle that will distinguish right from wrong action leads Parfit to dismiss even his own form of the Consent Principle, as too weak for the job (*211*). He moves on to consider the possibility of finding such a principle in the development of another aspect of Kant's Formula of Humanity. Here, too, however, as I shall argue, Parfit's interpretation fails to capture at least part of that formula's strength. The formula tells us always to treat rational agents as ends-in-themselves, and never as a means only. Tellingly, Parfit chooses to focus on the second idea, that of treating someone as a means only, rather than on the first idea, that of treating someone as an end in itself, in understanding what that principle might mean.

What does it mean to say of someone that he treats another as a means only? As Parfit shows us, if one pays special attention to the qualification 'only', and offers no context by which to interpret what that qualification might be intended to rule out, it is possible to understand treating someone as 'a means only,' or, as Parfit puts it, as 'a mere means,' as follows: You treat someone as a means only when, and only when you 'make use of a person's abilities, activities, or body, and . . . we also regard him as a mere instrument or tool: someone whose well-being and moral claims we ignore, and whom we would treat in whatever way would best achieve our aims' (*213*). By contrast, on Parfit's reading, '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' (*214*).

On this interpretation, as Parfit notes, a rabbit bred and used for experiments, a woman who is robbed of her engagement ring but not of her wedding ring, a man pushed over a bridge to prevent a greater number of deaths to other men, is not treated as a means only, so long as the treatment in question is shaped or even counterfactually constrained by restrictions on what kinds and extent of harm and suffering the agent is willing to inflict on her charge.[5]

A different way to understand the idea of treating someone as a means only might pay more attention to the formula of humanity as a whole, taking note that treating someone as a means only is contrasted

[5] As an aside, it might be noted as a point in favor of Parfit's understanding of the principle that it may be applied not only to rational agents but to nonrational animals, such as rabbits, as well. It seems to me to have broader application still, for I may also refrain from treating inanimate objects in certain ways in order to avoid damaging or destroying them. I may refrain from placing my favorite oil painting in the spot where I would get the most pleasure from it, because the sunny location would harm the painting in the long run. In similar ways, I might 'take care of' my home, my car, my breakfast dishes, and my tool kit—refraining from doing some things to them because it would damage them, and making efforts to preserve and maintain them even when, given my busy schedule, I have better things to do for myself. True, some of these activities might be justified by the fact that by keeping these objects in good shape they will be more useful to me in the long run. Insofar as this thought motivates me, I would still be treating them as means only, just being careful to consider the long view of these objects' value to me as means. But many people—and, for better or worse, I am among them—are in the habit of taking care of their possessions (and the possessions of others, too) whether it is in their interest or not. They are reluctant to destroy or damage objects of beauty or potential use, even when it is no good to them, and no known or certain good to anyone else. Though we treat these objects as means, we do not, on Parfit's interpretation, treat them as *mere* means. We would not do just anything to them as long as it suits our purposes. But this means that we do not treat even things that are first and foremost and essentially means, or tools, as mere means on Parfit's interpretation.

with treating someone as an end in itself. As I have always thought, the qualification 'only' serves as a way of recognizing that it is possible to treat people as means where this is not at all in tension with regarding them as ends-in-themselves. Indeed, we do this all the time: I treat my hairdresser as a means for securing a decent haircut; I treat my friend as a means for getting a ride to the airport; my students treat me as a means for getting training in philosophy; and my children treat me as a means for a home-cooked meal. There is nothing objectionable in any of these forms of interaction, at least in part because we offer ourselves up for such treatment. We do not treat each other in these cases as means only, or as mere means, because one of us is not using the other for his purposes *as opposed to*, or in negligence of, her own.

If we understand the Formula of Humanity along these lines, we will see it as instructing us to see rational beings, beings with purposes and plans of their own, as beings whose status forbids our using them in a way that neglects or ignores these purposes. On such an interpretation, one who pushes someone over a bridge in order to save several others from harm (assuming that he has not consented to being pushed, or shown himself about to jump anyway) is very definitely treating him as 'a means only'.[6] On this interpretation, the Formula is closely related in spirit to a principle that demands that we act only in ways to which affected parties do or would consent. Both such principles are ways of expressing the value of respect for other agents' autonomy.

However plausible and attractive we may find such principles as capturing *a* morally important perspective, however, they are highly problematic when considered as candidates for an absolute and supreme principle of ethics. For, as we noted before, many people are relatively uninterested and unwilling to sacrifice themselves or their loved ones for the sake of strangers or the common good—nor, as Parfit agrees, need they be irrational in being so. If we must respect their own actual choices and values, at least insofar as they are rational, then we will be frequently blocked from doing things that many will think we have strong moral reasons to do. We cannot, for example, save five or perhaps even five

[6] I should have thought that this would speak in favor of the interpretation insofar as one aims to capture an ordinary sense of the phrase (see Parfit, *227*).

thousand people by sacrificing one who does not want to be sacrificed. If we remove the qualification that their choices must be rational, or interpret rationality as ranging more widely, we will be even more tightly constrained—prevented, for instance, from smashing someone's toe in order to save a child's life. With Parfit, I agree that this is an unacceptable conclusion. So strong a principle of respect for autonomy cannot be an absolute, unconditional principle of morality. What is less clear to me, however, is that this implies that we must either interpret the *idea* of treating someone as a means only (that is, as a mere means) differently or else reject the suggestion that treating someone as a means only has direct and fundamental relevance to morality. An alternative approach would reject this dilemma. Rather, it would register the thought that, other things being equal, treating someone as a means only is to be avoided, and that it is always to be regretted, while yet allowing that it may sometimes be overridden by other moral considerations.

Parfit does not choose this alternative. Instead he moves on to discuss a different formulation of the Categorical Imperative, the Formula of Universal Law, to suggest that it be revised in a way that is more explicitly contractualist than Kant's own writings are, arriving at the principle he calls Kantian Contractualism. This principle, which I mentioned at the beginning of this paper, states that' everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose' (*342*).

This formula, like Parfit's so-called Consent Principle, asks us to constrain our actions not according to what everyone (under certain ideal conditions) *would* choose, but rather to what everyone rationally *could* choose. As such, one might think that this formula is as far from embracing the Kantian value of respect for autonomy as the Consent Principle we discussed earlier. It is possible, however, for a contractualist to defend this principle against such a complaint is a way that is not open to a defender of an analogous principle (like Parfit's Consent Principle) in a noncontractualist context. Specifically, contractualists aim at finding principles that all people, if they are reasonable, can agree on. As Rawls and Scanlon have pointed out, finding any such principles requires that we imagine people deliberating under certain ideal conditions. In particular, they suggest, not implausibly, that the deliberators be thought to be

under some pressure to try to reach agreement. Because of this, a deliberator might choose principles even though they are not her favorite ones because, unlike her favorite principles, these might be chosen by everyone, and the deliberator recognizes that some principles (or, at any rate, these principles) that everyone can agree on are better than none at all.

In other words, under the conditions relevant to contractualism (in which one is looking for principles that everyone can accept), the recognition that everyone rationally *could* accept a principle may count as a reason for someone *to* accept that principle. That is, that everyone could accept a principle may contribute to its making it true that, under certain ideal conditions, everyone would accept the principle.

Kantian Contractualism

Even if the Kantian Contractualist Formula is plausibly Kantian in embodying a respect for autonomy that is one of the hallmarks of Kantian ethics, what Parfit goes on to do with this formula once again bespeaks a failure to appreciate the value of autonomy and its power to generate reasons. Specifically, Parfit argues that Kantian Contractualism should lead us to accept a version of Rule Consequentialism. That is, he thinks Kantian Contractualists should ultimately see their view as committing them to the claim that 'Everyone ought to follow the principles whose universal acceptance would make things go best' (Chapter 16). Here is perhaps the most dramatic argument for the idea that the major traditions of Kantianism, contractualism, and consequentialism can be synthesized. Here again, however, it is open to question whether a defender of the Kantian tradition, or of combined Kantian and contractualist traditions, would agree.

As the shorter form of the argument (*400*) makes especially clear, the derivation that Parfit offers is very simple. Since, on Parfit's view, everyone *could* rationally choose that everyone act on optimific principles (principles, that is, whose acceptance by everyone would make things go best), and since, as he also thinks, there are no other principles that everyone could rationally choose, Kantian Contractualists should embrace the optimific principles. But it is not clear to me that there are no other principles that everyone could rationally choose.

It will be easiest to explain my reasons for doubt by considering one of the controversial consequences that Parfit thinks his argument implies—viz., that Kantian Contractualists should support principles that would require an agent faced with *Means* (the variation of *Earthquake* referred to earlier) to sacrifice Grey's leg in order to save White's life, and that may well require an agent faced with *Bridge* to push one man over the bridge to prevent the runaway trolley from killing five others who are in the trolley's path.

Parfit realizes that insofar as one imagines oneself in the positions of Grey or the man on the bridge, one may rationally want such principles not to be followed. One may rationally want a principle that would forbid one person from deciding to sacrifice another person's life or limb without his consent for the greater good of all. However, Parfit suggests, if you imagine yourself in the positions of White or of the five people stranded on the trolley track, you cannot rationally accept such a principle, for from these points of view the principle would lead to results that are both personally and impartially worse. I am not so sure.

It seems to me that what makes people resistant to endorsing a principle that would require, or even allow, someone to push the man off the bridge in the relevant case is not just the idea that the man, who is innocently minding his own business, would lose his life.[7] After all, we can assume that the five who are stranded on the trolley tracks are innocently minding their own business, too. Rather, what is distressing has to do with the fact that someone else, a third party, another human agent, is taking it into his own hands to sacrifice this man for the greater good. Imagining oneself in the position of this man, one might want it to be the case that insofar as it is anyone's decision whether he should give up his life to save the five, it should be *his* decision. And this thought seems to me one that can be entertained and supported even if one is not in his position.

[7] Strictly speaking, the agent in Parfit's *Bridge* case is not in a position literally to *push* White off the bridge, but rather to use a remote control device to cause White to fall onto the track. This variation, constructed so as to eliminate the possibility that the agent in the case had the option of jumping from the bridge himself, does not, so far as I can tell, make a difference to the train of thought I am discussing here.

In other words, it seems to me that many people have a strong preference for being in control of their own lives—that is, for being in control of their own lives insofar as anyone is in control of it.[8] They want to be the ones calling the shots, at a fairly local level, about what happens to their bodies, not to mention their lives. Moreover, this preference does not seem to have the character of a mere preference, as opposed to a value. It may well persist even in the face of the recognition that by retaining such control, one may lower one's overall security against the loss of life and limb. Indeed, it seems to me this concern is more on the surface of people's resistance to organ-transplant schemes that would allow a doctor to secretly kill a patient whose organs could be used to save five people than any concern about the anxiety and mistrust of doctors and hospitals that such a scheme would breed (*363*).

This preference does not seem to depend on any features of the agent that are not potentially universal. It does not depend, for example, on one's social status or one's wealth or gender. It seems rather a matter of taste or temperament. If this is right, then in principle *anyone* could have such a preference. If, in addition, we allow that this preference is rational—that is, *as* rational as a preference for a principle that would permit people to intervene in one's life in (nonmedical) emergency situations where the intervention would bring about a greater impartial good—then it follows that anyone *could* rationally accept the principle that favors leaving the man on the bridge alone to the principle that favors pushing him.[9]

If it be granted, therefore, that a person may rationally prefer to maintain immediate control over his body and his life to minimizing his risk of loss of life and limb, then Parfit's argument that Kantian

[8] This last clause is meant as a preemptive response to the objection that we are not in control of whether we find ourselves in the path of a runaway trolley or pinned down by an avalanche or subject to organ failure either.

[9] Or using remote control to cause him to fall off the bridge. These remarks are suggestive of a defense of the more general principle Parfit calls the Harmful Means Principle, according to which 'It is wrong to impose a serious injury on one person as a means of benefiting others' (*361*). According to Parfit, 'the Harmful Means Principle is best defended by appealing to our intuitive beliefs about which acts are wrong (*362*). My remarks do not appeal to such intuitions, however.

Contractualists must support a form of Rule Consequentialism will not go through. Even if we grant Parfit's claim that everyone *could* rationally accept optimific principles, as I am happy to do, we would also have to admit that everyone could rationally accept nonoptimific principles, in particular principles which would more strongly protect people against interference from others in the control of their own bodies and lives.

It will by now have occurred to many readers that the preference I have been describing as competitive with a preference for welfare—the preference for control over one's own life and limbs, the preference to be calling the shots with respect to one's own life—is closely related to the value of autonomy. Indeed, it might be described as a preference for the ability to exercise one's autonomy at the level of concrete action or of direct and immediate control.

Some Kantians or Kantian Contractualists might go farther, taking the preference for principles protecting the exercise of autonomy over principles that would bring optimific results to be *uniquely* rational. For them, Kantian Contractualism not only fails to imply what Parfit calls Kantian Consequentialism, it implies principles that are very likely, if not certain, to conflict with it. My remarks are not aimed at so strong a normative conclusion, however. Rather, they are meant to suggest that in failing to notice or address the challenge to his argument that is posed by a preference for autonomy over welfare, Parfit reveals once again a failure to recognize and appreciate the value of autonomy and the point of view of someone for whom that value is irreducibly important. Insofar as the expression of that point of view and of its fundamental relevance to morality is considered a major component and contribution of the Kantian tradition, Parfit's interpretation of that tradition seems inadequate, and the suggestion that a Kantian might come to support Parfit's 'Triple Theory' without violating or abandoning the spirit that led him to be a Kantian in the first place is open to doubt. A Kantian form of contractualism does not lead so quickly or so clearly to any form of consequentialism.

Other Tensions

I began this paper by quoting some remarks from the final paragraphs of Volume One of *On What Matters*, in which Parfit questions the

widely held view that Kantians, contractualists, and consequentialists disagree in certain sorts of deep and especially recalcitrant ways. Rather, he suggests, these three types of ethical theorists are all climbing the same mountain on different sides. In supporting the widely held view that Parfit rejects, I have focused on an aspect of Kantian ethics that, it seems to me, Parfit fails to capture and address in his interpretations and suggested revisions of Kant—namely, the central role Kant and Kantians accord to the idea of respect for autonomy. As is widely recognized, this aspect of Kantian ethics is especially in tension with consequentialism. Since Parfit talks not just of two but of three traditions that he aims to integrate and synthesize, however, a full discussion of his final claim would look also at the relations between contractualist and noncontractualist theories. Are there tensions between Kantianism and contractualism and between contractualism and consequentialism as deep as the tension between Kantianism and consequentialism?

These questions are difficult, in part because of the slipperiness of the term 'contractualism', understood as a label for a type of theory, or of a moral philosophical tradition. It is not clear whether the important ethical theories that appeal in one way or another to the idea of a contract all ought to be considered part of the same ethical tradition, and even when one is focusing on a single view or closely related set of views that have been identified as contractualist, one may be uncertain about which features of these views mark them out as distinctively deserving of that label.

If we accept Scanlon's characterization of contractualism, which associates it with the view that morality is fundamentally concerned with being able to justify oneself and one's actions to others, we should not be surprised to see a kind of harmony between Kantianism and contractualism. The restriction that one's actions must be justifiable to others seems close to the idea that one must act only in ways to which affected parties would, under specified conditions, consent. As such, it might be seen as another way to capture the view that morality requires us to respect other agents' autonomy that I have been identifying as a hallmark of Kantianism. Whether there are also plausible forms of Kantianism that would oppose contractualism is an interesting question, but I shall not pursue it here.

The relations between contractualism and consequentialism seem to me more complicated, and, more specifically, asymmetrical. Even though I argued above that a Kantian Contractualist need not accept Parfit's claim that her position leads to a kind of consequentialism (and for reasons that might apply to any contractualist, Kantian or otherwise), the argument was not meant to show a tension between the very idea of contractualism and that of consequentialism. To the contrary, as I understand them, contractualists are committed to the view that the right principles of morality are *whatever* principles satisfy the condition that is identified with 'being justifiable to everyone.' If those principles turn out to be the principles whose universal acceptance would make everything go best, then contractualism and this sort of Rule Consequentialism will coincide. On the other hand, there is a powerful form of consequentialism that would reject any form of contractualism. Specifically, consequentialists like Sidgwick, Smart, and Kagan, who take the sole fundamental value in morality to be that of making the world as good a place as possible, will not acknowledge moral reasons to limit themselves to acting within the limits of principles everyone could rationally accept if contradicting such principles would make things go better from an impartial point of view. Moreover, they will not acknowledge such reasons even if the principles in question are optimific principles (principles, that is, whose universal acceptance would make everything go best).

This point has often been made in discussions of Rule Consequentialism, a view which is rationally unstable from a purely consequentialist point of view. It has often been noted that if obedience to optimific rules always produces the best outcome, then Rule Consequentialism 'collapses' into Act Consequentialism, and if such obedience doesn't always produce the best outcome, then a strict consequentialist will have reason on occasion to violate the rules. Either way, a strict consequentialist will not have reason to adopt Rule Consequentialism over Act Consequentialism. Parfit himself seems to recognize this when he acknowledges, quite sensibly, that his Triple Theory, which includes an identification of moral wrongness with a violation of optimific principles, is 'only one-third consequentialist' (*418*).

Moreover, even if one is not a consequentialist, one may well think that consequences matter morally (indeed, it is hard not to think this).

The fact that you can save more lives or alleviate more misery by taking one course of action rather than another may count morally in favor of that action even if it is does not count decisively. Though adherents of Parfit's Triple Theory will support acting always according to optimific principles, occasions will arise in which one can be reasonably confident that one can do more good — save more lives, for example — by acting in ways that these principles forbid. Why should one follow the principles in this case? Strict consequentialists will think there is no reason, thus rejecting the Triple Theory, and Rule Consequentialism, completely. But even a pluralist, who acknowledges *some* reason to follow the rules at the cost of utility, reasons having to do perhaps with being able to justify oneself to others or to act consistently with the ideal of the kingdom of ends, may question whether, and if so why, these nonconsequentialist reasons *always* trump considerations of utility.

Conclusion — Hiking the Range

An answer might be forthcoming if one holds paramount the goal of reaching agreement on a supreme principle of morality. Parfit's Triple Theory does after all recognize both consequentialist and non-consequentialist (e.g., contractualist) values and fits them together in a systematic way. If one is looking for a single principle, or even a well-ordered set of principles, that assigns some importance to considerations of overall utility as well as to considerations of making oneself justifiable to others, Parfit's Triple Theory may be the best candidate for the job.

However, the commitment to reaching agreement on a single principle and on identifying that principle with the true morality can be questioned. That commitment itself is supported by only some values among others, and the idea that it can on occasion be morally better to act in a way that would *not* be supported by principles that everyone should accept is not, at least not plainly or obviously, self-contradictory.

Insofar as we can identify individual moral theorists as exponents of distinctively Kantian, contractualist, and consequentialist traditions, we can think of them as forming so many different hiking parties hiking along different trails. Along the way, each party will come to various trail junctions, and have to decide on which branch to continue.

There will be some reasons favoring the choice of continuing along one trail, and other reasons supporting the choice of another. Making one choice will give the hikers a better chance of arriving at a theory whose principles will yield more definite results, or which will be more likely to be agreeable to a greater variety of others. The other path, however, may, have more of what attracted the hikers to that particular trail in the first place.

Some members of each party may choose the path that has the advantages of the first sort. Parfit's book gives us reasons to think and to hope that the members of each party who make this choice will indeed be climbing the same mountain and will meet at the top.

As I have meant to show, however, others will comprehensibly choose other paths. Some Kantians will choose to forgo principles obedience to which would allow greater benefits in order to more faithfully respect autonomy—for example, they will choose principles that would forbid pushing bystanders off bridges even to save more people. Some consequentialists will sacrifice the ability to justify themselves to everyone in order to bring about a greater good—for example, they may approve of the doctor who surreptitiously kills one healthy person to use his organs to save five others. These paths will presumably take them up different mountains.

Parfit's reading of Kant makes me speculate that insofar as Parfit imagines himself to be a member of the Kantian party, his own methodological commitment to finding a supreme principle of morality illuminates one path so much more brightly than others that he fails to so much as notice some of the junctures where there may be more than one plausible way to go on. My main purpose in this paper has been to more accurately represent the landscape, so as at least to register the fact that, however good the reasons are for choosing one route, and ultimately, one mountain, over another, one who does so will inevitably miss benefits or beauties that lie along the paths not taken.

If one conceives of the enterprise of moral theorizing as the single-minded pursuit of a supreme principle of morality, then perhaps there is only one choice to make, and only one mountain worth climbing. One might instead, however, think of moral theorizing as an activity with a number of aims, including the articulation and appreciation of

the values that are fundamental to moral action and moral reasoning, and the exploration of how far these values can be jointly realized and expressed. If one does not assume that these values can be jointly realized to a maximum degree, then one will think that in order to get the most out of moral theory, one must hike the whole range.

Is there a right way to conceive of the task of moral theorizing? This is one way of asking how important it is to find, or agree on, a supreme principle (or a well-ordered set of principles) of morality. How valuable is it to find or agree on a unified set of principles that is comprehensive and that yields definite answers to questions that, at first glance, require balancing different and incommensurable values? What is to be gained by identifying such principles? What, if anything, might be lost? And what practical implications would or should the identification of such principles have?

As I mentioned at the beginning of this paper, philosophers have been searching for the supreme principle of morality since moral philosophy began. The desire for such a principle is so natural and its value so apparently obvious as to hardly call for explicit defense. Still, before concluding, I want to raise doubts about two reasons for thinking that the determination of such a principle would be as valuable and important as moral philosophers have tended to think.

One pattern of thought that makes the goal of finding a supreme principle of morality seem very desirable has to do with the ideal of social harmony, the appeal of achieving social consensus. If there is a supreme principle of morality, one might think, then everyone ought rationally to recognize and accept it, and acting according to it would be justifiable to all.[10] And wouldn't it be great to know how to live, or to act, in a way that everyone would approve?

Indeed, it would. However, there is a slide in this line of thought from the prospect of reaching the *theoretical* goal of identifying a principle that all reasonable people ought to accept and the imagined consensus of real human beings in our diverse and fractured world. While doing moral theory, we naturally take ourselves to be reasonable people, and

[10] Contractualists think the fact that a principle is justifiable to all is what makes it a supreme principle of morality; noncontractualists may think the order of explanation is reversed.

tend perhaps implicitly to assume that everyone else (everyone else in the world, that is) is equally reasonable and equally interested enough in discovering the true morality to engage in the kind of moral reflection that would be necessary for coming to see that the principle one has identified as the supreme moral principle deserves to be treated as such. But this assumption is crazy.

Even if there were a principle that it would be reasonable for everyone to accept, not everyone would accept it. Not everyone *is* reasonable, and not every reasonable person *will* accept a principle that, were they perfectly reasonable and also perfectly attentive to a set of complicated moral arguments, they should accept. The social harmony that would be achieved by identifying a supreme principle of morality and acting according to it, would, in other words, be purely hypothetical. Even if one acted according to that principle, one would be likely to find herself acting on occasion in a way to which an affected party would not consent, or in a way in which an affected party would feel himself treated unacceptably as a means, in a way that he did not regard as justifiable to him.

A second, perhaps even more powerful, reason for being deeply attracted to the goal of finding a supreme principle of morality, has to do with the desire for practical moral guidance, a wish to be given definite answers to hard moral questions. Like the desire for social consensus, this wish is reasonable, too. A lot is at stake in situations like *Earthquake*, *Means*, *Bridge*, and *Transplant*, for example, and it would be nice to have a principle to apply that would assure one of doing the right thing. To be told that there are reasons for doing one thing and reasons for doing the other, is to tell us nothing new, nothing helpful. We want more from moral theory than that.

I agree. However, it is not obvious that searching for, or even succeeding in finding, a supreme principle will give us the moral guidance we seek. The principles that Parfit defends are of less practical usefulness than might be supposed.

To be sure, these principles can be given as answers, in a sense, to any question of what to do. I find myself beside a man on a bridge, and see a runaway trolley speeding below on its way to kill five people if nothing is done to interfere. If I push the man over, he will die but halt the trolley, saving the five other people's lives. What should I do?

Kantian Contractualism has an answer of sorts: Act according to those principles whose universal acceptance everyone could rationally will, or choose. In an earlier section, I gave some reason for doubting that that principle would yield any determinate advice. Even if all rational people could accept principles whose universal acceptance would make things go best, I suggested, they might also be able to accept principles that gave higher priority to respecting autonomy.

Moreover, even if I am wrong about this and Parfit is right that Kantian Contractualism gives exclusive support to optimific principles, the question would remain which principle, in cases like this, is optimific. Parfit suggests that there is a difference between medical cases and cases that in other respects are structurally similar. But I can construct an argument concerning the *Bridge* case, too, that suggests that it would be optimific in the long run to refrain from pushing people off bridges. Between Parfit's defense of the Emergency Principle (*365–6*) and my imagined argument that suggests that the adoption of something closer to the Harmful Means Principle would lead to better results, I have no idea which argument is stronger. There is so much to consider about which it is difficult to be certain. What seems most reasonable here is to mistrust one's ability to be objective enough, imaginative enough, and thorough enough to reach reliable conclusions about such matters.

The point is that any plausible candidate for a supreme principle of morality would have to be so abstract or so complicated or both that the principle would be difficult to apply. Though such a principle may be helpful in suggesting a way to explain to ourselves why acts that we think are right really are right, or in suggesting a way to respond to concerns that some other action would be better, it is unlikely to give us practical guidance for morally difficult situations in which we don't know what to do before consulting the principle.

Although I have, in the last few paragraphs, offered reasons to question the preeminent place that Parfit and others have accorded the search for a supreme principle of morality as the aim of moral theorizing, I do not mean to suggest that the search is a worthless or a futile one. To the contrary, there is much to be gained—much indeed, that has been gained—even if we do not agree that the search has, or has yet, been entirely successful. We will gain even more if we actually

find, or, alternatively, choose to agree on, such a principle. However, I suspect that if we find or choose such a principle, acting according to it will not capture or realize all the values that are traditionally regarded as moral values without remainder. Maximizing utility does conflict sometimes with respecting autonomy, and for all I know each may conflict sometimes with obedience to principles that no one can reasonably reject. Contrary to what Parfit seems to suggest at the end of Volume One, you cannot please all the moral theorists all the time.

If that is right, then were we to find or agree on a supreme principle of morality, it would embody some degree of compromise among values, reached presumably for the sake of gaining the benefit of having some supreme principle of morality rather than none at all. In the interest of moral clarity, we ought to recognize that fact, and so acknowledge that even if an act is supported by what we have come to regard as the supreme principle, and so is, strictly speaking, morally right, that would not mean that there can be nothing to regret or to apologize for in the doing of it, and even if an act is forbidden by the supreme principle and so is, strictly speaking, morally wrong, that would not mean that there is nothing to be said in its or its agent's defense. These thoughts in turn may raise questions about what the claim that an act is morally wrong really means. Does it mean, or imply, that an agent who performs such an act ought to feel guilty, or that a third party who recognizes that the agent behaved wrongly is justified in blaming the agent? How strongly or consistently should we want people to be constrained by the principles—and in particular, by the supreme principle of morality, if there is one? How strongly should we be guided by them (or it) ourselves?

These are metaethical questions of a kind Parfit points toward in Chapter Seven, section 22. Noting that different senses of 'wrong' are associated variously with blameworthiness, with the appropriateness of reactive attitudes, and with justifiability to others, he explains that 'in the rest of this book, [he] shall use "ought morally" and "wrong" vaguely, in some combination of these senses' (*174*). 'Except in Part Six,' he continues, '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.' An

assessment of Parfit's discussion in Part Six is beyond the scope of this essay. It is both striking and impressive how well Parfit characterizes the range of these meta-ethical questions (geographical pun unintended) even here, before he subjects them to thorough examination, and much to his credit that he recognizes their significance for a satisfactory understanding of what the arguments of Volume One can be said to have accomplished. Whether they have taken us closer to a supreme principle of morality, whatever that means, is open to doubt. But even if they do not, they have surely led us on a trail worth following, full of intellectual attractions and moral philosophical insights along the way.

Humanity as End in Itself

Allen Wood

Part One: Rational Consent, Practical Reason, and Humanity as End in itself

There is a great deal in Parfit's chapters, especially in Chapters 8 to 10 (on which I am going to concentrate these comments) with which I strongly agree. I think Parfit provides a better account than O'Neill and Korsgaard do of what Kant meant in saying that for me to treat another as an end in itself, the other must be able to 'contain in himself the end of my action' (G4: 429–30),[11] and also a better account of the relation of this idea to issues surrounding hypothetical rationally consent. I also find very illuminating Parfit's remarks about the relation of possible rational consent to actual consent and how each bears on the morality of actions.

At a deeper level, too, I think I favor a reading of Kant that puts him closer to what Rawlsian style Kantians would regard as 'dogmatic rationalist' views in ethics—and I think this means closer to the position Parfit wants to defend. Thus I would accept, as good Kantianism, what Parfit calls a 'value-based' theory of reasons; Parfit's rejection of 'desire-based' theories therefore seems to me nothing but good Kantianism. I therefore also accept his thesis that 'no reasons are provided by our desires and aims.' But to this I would want to add two other things (which I don't think Parfit means to deny): first, that our desires and aims are often merely the rational expression of value-based reasons,

[11] Kant, *Groundwork for the Metaphysics of Morals*, ed. and tr. Allen W. Wood (New Haven: Yale University Press, 2002), abbreviated as 'G' and cited by volume: page number in the Akademie-Ausgabe of Kants *Schriften* (Berlin: W. de Gruyter, 1902–). Other writings of Kant will be cited by volume: page number in that edition.

and second, that our desires might constitute a crucial aspect of some of our reasons, as long as they stand in the right relation to values.

Where I think I part company with Parfit is on certain questions of method in ethical theory. He seems to prefer a method descending (as I see it) from Sidgwick — a method that involves appeal to what Sidgwick called 'the common moral opinions of mankind' (or just 'Common Sense') in the formulation and testing of moral principles. By contrast, I favor a method, which I find not only in Kant but also in utilitarians such as Bentham and Mill, that would draw the fundamental moral principle from very general and fundamental considerations about the nature of rational desire and action, and would then attempt to reconcile these principles with common moral opinions only insofar as those opinions can be seen as applications of the principles. Sidgwick seems to have thought that what he called 'primary intuitions of Reason' are to be used only to systematize and correct Common Sense,[12] which continues to exercise authority within moral theory independently of first principles, and might even help to shape the formulation of moral principles.[13]

The Kantian and Millian method that I favor, by contrast, involves a fundamental principle whose ground is independent of moral intuitions or Common Sense, and then the derivation from the fundamental principle of various moral rules or duties. Conclusions about particular cases are not inferred directly from the first principle at all, but rest on it only mediately, through what Mill calls 'secondary principles' and Kant calls 'duties' (of various kinds, of which he provides a taxonomy). The derivation of moral rules or duties from the first principle, moreover, is also not deductive. The first principle is instead fundamentally an articulation of a basic value (that of rational nature for Kant, that of happiness for Mill). The rules or duties represent an interpretation of the normative principles applying that basic value under the conditions of human life. In their application, moreover, the rules or duties themselves require interpretation, and admit of exceptions, by reference to the first

[12] Henry Sidgwick, *The Methods of Ethics* (Indianapolis: Hackett, 1981), 373–4.

[13] In this respect, Rawls's method of 'reflective equilibrium' owes more to Sidgwick than it does to Kant.

principle.[14] More recent (Sidgwickian) theory sets itself the goal of providing a precise principle or set of principles which, along with a set of facts, enable one to deduce the 'right' conclusion about what to do under any conceivable situation. That's what it is for Sidgwick to make ethics 'scientific'.[15] For Kantian or Millian theory, as I understand them, this is such a hopeless goal that it would be wrongheaded to orient your theoretical method to it.

[14] This interpretation of Mill might be controversial, but I would defend it based on the following things: (1) the account he gives of the relation of the rules of morality to the principle of utility, as social 'direction-posts,' giving us some guidance regarding the social pursuit of the general happiness, which he regards as a standard exercising only a very general (and even largely unacknowledged) influence on the content of such rules (Mill, *Utilitarianism*, ed. G. Sher, 2nd edn. (Indianapolis: Hackett, 2001), 24–6); (2) Mill responds to the charge that there is not enough time prior to each action to weigh all the utilities on every side by comparing the application of the principle of utility to the application, by Christian ethics, of the Old and New Testaments—which would involve the *interpretation* of the scriptures in the light of human experience—so likewise, I suggest, Mill regards moral rules as resulting from the interpretation of the principle of utility in the light of experience (p. 23); and (3) the fact that Mill's formulation of the first principle itself—that 'actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness' (p. 7)—is a rather loose one, not a formulation from which anyone could justifiably think that we could directly determine what to do in particular cases. It may also be controversial (though it should not be) that Kantian duties always in principle admit of exceptions. 'Exceptivae' constitutes one of the twelve basic 'categories of freedom' Kant presents (analogously to the twelve theoretical categories) in the *Critique of Practical Reason* (5: 66). Most of the twenty-odd 'casuistical questions' Kant discusses in the Doctrine of Virtue concern possible exceptions to the duty in question. The general purpose of these discussions is described by Kant as 'a practice in how to *seek* truth' regarding 'questions that call for judgment'–and judgment (the correct application of a rule to particular circumstances) is something Kant insists can never be reduced to maxims, rules or principles since 'one can always ask for yet another principle for applying this maxim to cases that may arise' (6: 411). Thus casuistry, the interpretation and application of moral rules or duties to particular cases, always involves a distinct stage of thinking that cannot be made a matter of rules or principles.

[15] Sidgwick, *Methods of Ethics*, 359–61.

The system of moral philosophy, following the Kantian conception, consists of three different things: first, a fundamental principle or value (which Kant thought was *a priori*); second, a body of empirical information and theory about human beings and their situation (which in the Groundwork Kant called 'practical anthropology' (G4:388) and later described as 'empirical principles of application' for the moral principles (MS 6:217)); and finally a set of rules, duties, or other moral conclusions resulting from the interpretation of the former principle or value in light of the latter information. This third part of Kantian ethical theory is the taxonomy or system of duties expounded in the *Metaphysics of Morals* (the *ethical* part in the Doctrine of Virtue). It corresponds roughly to the set of moral rules that Mill regards as involved in every case of moral obligation, and relates only loosely to the principle of utility, which he does not regard as imposing on us any obligations directly, and from which Mill immediately derives (even together with facts about the consequences of actions) no substantive conclusions about what to do in particular cases.[16]

I think this way of conceiving of moral theory, and the fact that Parfit favors a different theoretical method, accounts for some of the ways Parfit disagrees with my interpretation of Kant at the beginning of Chapter 10. He quotes me interpreting Kant's Formula of Humanity as End in Itself (FH) as saying that 'we must always treat people in ways that express respect for them' and then objects that 'most wrong acts do not treat people in disrespectful ways.' The remark he quotes here

[16] Thus Mill is neither an 'act utilitarian' nor any member of the large species of 'rule utilitarian' whose procedure takes the form of stating a utilitarian principle from which, along with a set of facts, conclusions about what to do could be drawn. For Mill, the main functions of the first principle seem to be three: (*a*) to provide the basic value-orientation of ethics, whose interpretation provides the basis for accepted moral rules; (*b*) to provide a standard through which the accepted moral rules can be corrected and improved, and (*c*) to provide a ground on which exceptions to these rules may be admitted. None of these functions, however, takes the form of a decision procedure through which specific rules or the making of exceptions to them is to be arrived at by deductive inferences. In this way, Mill seems to me the most sensible (and incidentally, despite the gross misunderstandings of Kant displayed in *Utilitarianism*, also the most Kantian) of the great historical utilitarians.

occurs in the context of a more systematic exposition of Kant's theory, which, as I read it, is what Parfit would call a 'narrow' or 'monistic' value-based theory. For this theory, all reasons are grounded, directly or indirectly, on the single value of rational nature, which Kant expresses in two ways: as the objective worth of humanity as end in itself, and the dignity of personality as universally legislative.

Respect, as I understand it, is first of all a feeling or emotion. Contrary to the Stoics (and to some grossly mistaken misinterpretations of Kantian ethics), Kant thought it impossible for a finite rational being to act rationally at all without having certain feelings and emotions and manifesting them in its actions. In the *Metaphysics of Morals*, Kant specifies four such feelings (moral feeling, conscience, love of human beings, and respect). These feelings are rational rather than empirical in origin, and susceptibility to them is a condition for being a moral agent at all (MS 6:400). I would describe *respect* in general as the feeling appropriate to the rational recognition of objective value.[17]

Respect is something we not only feel but also show in actions that express it. It is the active expression of respect rather than the mere feeling that matters for moral conduct. On Kant's monistic value-based theory of practical reasons, all reasons for action are based directly or indirectly on the objective value of rational nature, and this is especially true of moral reasons that take the form of categorical imperatives. Obedience to every categorical imperative thus involves showing respect for the objective value of rational nature. In that sense, what morality demands most fundamentally is that we show respect for that value, and violations of morality all involve treating that value—often, the value of rational nature in the person of rational beings—with disrespect. Many morally wrong actions do not 'display disrespect for people' in any conventional sense of that phrase, but if Kant's theory is correct,

[17] From this observation about respect I immediately infer that all metaethical antirealists, who deny there is such a thing as objective value, are either radically defective specimens of humanity who are incapable of feeling respect for anyone or anything, or else every time they do feel it they commit themselves to contradict their own metaethical theories—theories which are often ravishingly subtle and sophisticated in execution, but must nevertheless be recognized from the start by all rational agents as obviously and brutally false.

the moral wrongness of these actions always consists fundamentally in the way they show disrespect for the objective value of rational nature.

Parfit recognizes the Kantian distinction between values to be respected and values to be promoted. But he is worried that the claim that dignity is a value above all price may commit Kantians to the view that rational nature as a value to be promoted must take absolute priority over other values to be promoted. This is, for instance, the way Parfit reads the following statement by Thomas Hill: 'Kant's view implies that pleasure and the alleviation of pain, even gross misery, have mere price, never to be placed above the value of rationality in persons.'[18] That fear seems to me based on a misunderstanding. Promoting rational nature (as one value that can be promoted) is grounded in respect for rational nature (as the basic value to be respected). It is the latter value that has a dignity that is beyond all price, and it must be given priority over all competing values. But equally, concern for the alleviation of human suffering (as a value to be promoted) is grounded in this same fundamental value. But this implies no absolute priority of the value of developing rational nature (as one of the values to be promoted) over other values to be promoted that are also grounded in respect for rational nature. If the above quotation from Hill is correctly read as asserting *that* priority, then his position is not a correct interpretation of Kantian doctrines.

In Kant's view, the objective value of rational nature grounds two general kinds of ends which are duties: our own perfection and the happiness of others. (The value of our own happiness, except as an indirect duty, is for Kant an object of prudential rather than moral reason; and the perfection of others is a duty for us only insofar as we contribute to perfections they want to acquire, and therefore falls under the heading of their happiness.) Perfection prominently includes our rational nature (both moral and nonmoral) as a value to be promoted. Both kinds of duty are wide or imperfect. Thus for Kant there is no systematic priority of perfection over happiness as ends or values to be promoted.

[18] Thomas E. Hill, *Dignity and Practical Reason* (New York: Cornell University Press, 1992), 56–7.

Parfit is also in danger of misunderstanding Kant when he says that the 'humanity' which has dignity cannot refer to non-moral rationality. Kant says that humanity, as the capacity to set ends according to reason, is an end in itself and that humanity insofar as it is capable of morality has dignity. As I interpret him, Kant holds that it is our *humanity* that is an end in itself—where 'humanity' has a technical sense, referring to our capacity to set ends (which includes both instrumental rationality and prudential rationality—the capacity to frame a concept of happiness and to give our happiness priority over more limited aims of inclination). We should therefore include the permissible ends of others, especially their happiness (as the general and comprehensive conception of those ends), among our ends as well (though there are no strict rules in general regarding the priority we must give all these ends among one another). *Dignity*—by which Kant means that supreme worth which must never be sacrificed or traded away—belongs to rational nature not in its capacity to set ends, but only in its capacity of giving (and obeying) moral laws (G 4:435).

It is the *capacity* for morality, however, not its successful exercise, that has dignity.[19] Thus I agree with Parfit when he interprets Kant as saying that even the morally worst people have dignity, and in that sense they have exactly the same worth as even the morally best people. I also agree with Parfit when he says that this view of Kant's expresses a 'profound truth.' Parfit is further correct to point out that none of this implies that my having dignity as a human being makes me a *good human being*. Not everything having value is thereby something *good*, especially good of its kind. For Kant, the *good* is that which is recognized as practically necessary independently of inclination (G 4:412). Having a character like that of a bad person is the direct reverse of what is practically necessary, though it is also practically necessary to treat even

[19] Parfit concludes that Kant's uses of 'humanity' are 'shifting and vague'. I think this is right insofar as he speaks of the 'dignity of humanity', whereas, to be strictly accurate, it is personality (the capacity to give universal law and obey it) rather than humanity (the capacity to set ends according to reason) that has dignity. But if, as I believe, Kant does hold (and must hold) that humanity and personality in these senses are necessarily coextensive, then no serious error is involved in his use of the phrase 'dignity of humanity'.

the worst person with the respect due to the dignity of rational nature, and so it is that treatment of the bad person, and not the bad person, that is good.

Parfit denies that FH—the principle that we should always respect humanity as an end in itself—is a practically useful principle. In response to my claims that it provides us with the right value-basis for settling difficult issues and that on many difficult issues, it is an advantage of FH that different sides can use it to articulate their strongest arguments, Parfit asserts that on a wide range of disputed issues appeals to FH do not in fact constitute the strongest arguments of each side. I think we may be talking past each other here, because we are beginning from different assumptions (which I have tried to clarify above) about the aims and structure of moral theory and the relation of a theory's basic principle to conclusions about what to do. Kantian theory is grounded on a supreme principle, which is then applied interpretively to a body of empirical information and theory about human nature and human life, yielding a set of moral rules or duties. These in turn are applied to particular circumstances, through practical judgment, in determining what to do.

FH is one of Kant's formulations of the supreme principle, the one he uses most often in deriving his system of duties in the *Metaphysics of Morals*. That is the role FH is playing when I make the claims about which Parfit is skeptical. I suspect that Parfit, on the other hand, thinks of moral theory as the attempt to formulate precise principles from which we can rigorously derive a set of conclusions about what to do in all actual or imaginary cases. The acceptability of these principles, for Parfit, depends on how the conclusions derivable from them match up with Sidgwick's 'Common Sense' or 'common moral opinions of mankind'. Principles well-grounded might in difficult cases give us reasons for revising our conclusion about particular cases, but flagrant and systematic conflict of a candidate principle with our intuitions is regarded as invalidating that principle. Parfit is treating FH as a principle to be evaluated by these criteria, and he is rejecting it as too indeterminate to yield the specific conclusions such a principle is supposed to yield, and hence also incapable of providing adequate arguments on different sides of a moral controversy that would be required by this conception of moral theory.

When FH is regarded in this way, I think Parfit is right, but not when it is regarded in the way I regard it—which is also the way I think Kant regarded it. (My way of reading Kant obviously involves reading his four famous illustrations of the Formula of Universal Law in quite a different way from that in which they are customarily read—including, I think, the way Parfit chooses to read them in Chapters 12 and beyond. But that difference will not be pursued further in these comments.)

Part Two: 'Trolley Problems'

The rest of my comments here will contain some general reflections on some of the examples Parfit uses, especially in Chapters 8 and 9. I think these comments are relevant to the theoretical differences I have tried to sketch above, for they concern one now fashionable way of executing the methodological strategy I have suggested that Parfit draws broadly from Sidgwick. I don't think the following remarks do anything at all to discredit the Sidgwickian program broadly conceived. Like many ambitious philosophical projects, it is too formidable in its conception ever to be refuted by a few clever arguments or examples. But I do intend to challenge some fashionable ways of carrying out such a program. My comments also relate to FH, in that they help to illustrate the way in which I think it can figure productively in moral reasoning. I should also frankly admit that these comments give me the opportunity to get off my chest some complaints about what many moral philosophers do nowadays.

In May of 2001, the Tanner lecturer at Stanford University was Dorothy Allison, author of the novel *Bastard Out of Carolina*. Allison didn't talk much about moral philosophy as such, but she did discuss a 'lifeboat problem' that she had heard about from a philosopher. Her reaction was to *reject* the problem—to refuse to answer it at all—on the ground that we should refuse on principle to choose between one life and five lives. Even to pose the question in those terms, she said, is already immoral. The only real moral issue raised by such examples, she thought, is why provision had not been made for more or larger lifeboats. To many philosophers her remarks would no doubt seem naïve or even unreasonable. Yet I think Allison's reaction to the lifeboat

problem is far more sensible and right-minded than what we usually get from most of the philosophers who make use of such examples.

I am going to refer to these kinds of examples not as 'lifeboat problems' but as 'trolley problems'. (None of Parfit's examples are actually about trolleys, though two of them are about trains.) They are all examples where the main point is that you must choose between saving more people from death and saving fewer. Since we think a human death is in general something very bad, it is natural also to think that the option involving fewer deaths must be preferable to the one involving more deaths. The examples gain their poignancy from the fact that this apparently obvious point suddenly begins to seem questionable or even counterintuitive when the fewer deaths are *caused* in the wrong way. The intent of the examples is usually to incite us to formulate principles that correspond to, or even justify, our moral intuitions (or deliverances of Sidgwickian 'Common Sense') about the difficult or problematic cases presented in the examples. The hope is apparently that principles arrived at in this way will help us decide difficult cases in real life with Sidgwickian scientific precision.

Some might think that if FH regards every rational being as having dignity (or worth that cannot be rationally traded away to get anything else), then it might very well not only support Allison's judgments about the lifeboat problem, but also entail that there could be no rational way of choosing between one life and five lives, or if it comes to that, five billion lives. If so, then FH would appear to have consequences that seem plainly unacceptable according to our intuitions. We apparently could never permit even a single death, not even to save the whole human race.

No doubt the fact that rational nature has dignity or incomparable worth *does* mean that the lives of beings having rational nature are valuable and important. But merely from the fact that the value of *rational nature* cannot be rationally sacrificed or traded away, it clearly *does not* follow that the *lives* of rational beings can never be rationally sacrificed. If a person heroically sacrifices her life to save others, or to uphold some important moral principle, that is not a case of undervaluing her own rational nature. Depending on the circumstances and the principle involved, it might even be a case of *preferring* the value of her *rational nature* to the value of her *life*, and Kantian ethics

might even require it. Nor does FH lend unambiguous support to the vague idea of the 'sanctity of human life'—an idea that, in its popular and political application, usually involves a lot of self-deceptive rhetorical posturing, and is sometimes put in the service of some of the most pernicious moral superstitions currently on sale in the marketplace of moral ideas (for instance, dreadful superstitions about the unexceptionable wrongness of euthanasia, or the right to life of human embryos and fetuses). I strongly caution against associating FH with morally obscene popular prejudices such as these.

The bearing of FH on trolley problems is therefore also not entirely clear. One thing I hope is clear by now is that for Kantian ethics, the point of a moral principle such as FH is not directly to tell us what we should *do*. It is rather to ground a set of rules or duties, and more generally to orient us as to how we should and should not *think* about what we should do. We would be right to conclude from FH, for instance, that we should be reluctant to treat human lives as having the sort of value that can be measured and reckoned up. That is what I think Dorothy Allison was getting right. It would follow that answers to problems like Parfit's *Lifeboat*, *Tunnel* and *Bridge*, therefore, can never be as clear (or as trivial) as the arithmetical fact that five is greater than one. The tendency of some moral philosophers to draw such inferences is due to their bad habit of thinking that the canonical form of every moral principle must consist in the scientifically precise way it preferentially ranks states of affairs (as the outcomes of actions). But what FH tells us is that the fundamental bearers of value are not states of affairs at all, but persons and the humanity or rational nature in persons. This is not a kind of value that translates easily into preferential rankings of states of affairs.

FH does not imply that it is always immoral to choose five lives instead of one, but I think it does imply that we should be reluctant to think about such choices in those terms, or indeed in terms of any preferential rankings of states of affairs. FH rather implies that we ought to arrange things in the world so that agents are not faced with choices of that kind. Of course this means arranging things, as far as possible, so that one life need not be sacrificed to save five. But it also means arranging things—including our moral deliberations—so that when numbers of lives are at stake, the choices dictated by our moral principles are not

based merely on the numbers, as trolley problems—in the very way they are posed, through the careful selection of information included in and excluded from them—often suggest they have to be.

I have long thought that trolley problems provide misleading ways of thinking about moral philosophy. Part of these misgivings is the doubt that the so-called 'intuitions' they evoke even constitute trustworthy data for moral philosophy. As Sidgwick was fully aware, regarded as indicators of which moral principles are acceptable or unacceptable, our intuitions are worth taking seriously only if they represent reflective reactions to situations to which our moral education and experience might provide us with some reliable guide.[20] Poll-takers are well aware that the way a question is framed often determines the answer most people will give to it. What might seem to us genuine intuitions are unreliable or even treacherous if they have been elicited in ways that lead us to ignore factors we should not, or that smuggle in theoretical commitments that would seem doubtful to us if we were to examine them explicitly.

Most of the situations described in trolley problems are highly unlikely to occur in real life and the situations are described in ways that are so impoverished as to be downright cartoonish. (In imagining *Bridge*, for instance, I can't help casting my favorite cartoon superhero, Wile E. Coyote, in the role of the hapless single person who may be toppled onto the track.) But this by itself is surely not a problem. It is extremely rare for a man to lure teenage boys into his apartment, then kill, dismember and eat them; and at this writing, at any rate, it remains an utterly unique occurrence for a group of terrorists to hijack airliners and crash them into skyscrapers filled with innocent people going about their daily lives. But the rarity of such cases does not lead us to mistrust our moral intuitions about these cases. Nor do we mistrust our moral reactions to the absurdly fantastic villainy sometimes depicted in comic books and action movies.[21]

[20] Sidgwick, *Methods of Ethics*, 96–103, 374, 421–2.

[21] We ought, however, to mistrust its dramatic purpose, which is typically to render morally acceptable to us the fantastic brutality and violence practiced by the heroes of such stories. It seems to me by no means implausible to think that the currency of such dramatic situations has helped create a climate in

The deceptiveness in trolley problems is indirectly related to their cartoonishness, however, in that it consists at least partly in the fact that we are usually deprived of morally relevant facts that we would often have in real life, and often just as significantly, that we are required to stipulate that we are certain about some matters which in real life could never be certain. The result is that we are subtly encouraged to ignore some moral principles (as irrelevant or inoperative, since their applicability has been stipulated away). And in their place, we are incited to invoke (or even invent) quite other principles, and even to regard these principles as morally fundamental, when in real life such principles could seldom come into play, or even if they did, they would never seem to us as compelling as they do in the situation described in the trolley problem.

Trolley problems focus primary attention on the value or disvalue of certain consequences or states of affairs (usually, more human deaths or fewer). But trolley problem philosophers are by no means all consequentialists. Trolley problems are quite frequently used, in fact, to support anti-consequentialist conclusions in moral philosophy, and many of them appear to do so. But in these problems, attention is directed exclusively to the consequences of certain actions for the weal or woe of individuals and also the way those actions relate causally to those consequences. Typically, the circumstantial rights, claims and entitlements people would have in real life situations are put entirely out of action (ignored or stipulated away). In the process, an important range of considerations that are, should be, and in real life would be absolutely decisive in our moral thinking about these cases in the real world is systematically abstracted out. The philosophical consequences of doing this seem to me utterly disastrous, and to render trolley problems far worse than useless for moral philosophy. I would like to illustrate these general points by briefly discussing three problems used by Parfit in Chapters 8 and 9.

which a great many people can find morally acceptable the monstrous conduct, domestic as well as foreign, of the utterly evil regime that ruled the U.S. from 2001 to 2009.

1 *Lifeboat*

It seems to me that when faced with a situation like *Lifeboat*, there is only one morally defensible policy: You must seek to rescue all six people as quickly and efficiently as possible. It might very well be true that, following this policy, you should first set about rescuing the five and only then try to rescue the single person, because in that way you will go farther, faster and with greater certainty toward achieving your only legitimate goal (which is rescuing all six). But if you thought you could go farther faster and with greater certainty toward the goal of saving all six by rescuing the single person first (say, because this person's rock is right on your way to the rock with the other five on it), then you obviously should do that.

It is relevant here — even decisive — that in the real world, if both rocks are in imminent danger of being swept under the water, then you would very likely not know for certain that you must choose between saving the single person and saving the five. (The stipulation that you are certain about this ruins the real moral issue just as certainly as it would ruin some issue in rational choice theory to stipulate that you are sure which box being offered you contains the larger amount of money.) Rather, in real life there would always be some chance that you would save all six, and if both rocks were about to go under there would also probably be a significant chance that no matter what you did, all six people would drown. When a philosopher simply stipulates that we are certain you can save all and only the inhabitants of exactly one rock, then we should be clear that he is posing a problem so different from otherwise similar moral problems you might face in real life that any 'intuitions' we have in response to the philosopher's problem should be suspect.

There is one intuition about a situation such as *Lifeboat* that is perfectly clear and not the least suspect. It is this: if any of the six drown, the result is tragic — it is unacceptable. You will regard yourself as having failed significantly in your rescue efforts no matter what you did, even if you know your failure was inevitable and not your fault. Another vivid and reliable intuition is that all concerned have an urgent obligation to call to account whoever is to blame for the fact that there were not enough lifeboats. They should try to find out why this happened, and take steps to minimize the chances of its happening ever

again. We saw this point illustrated dramatically several years ago in the universal reaction to the utter incompetence of federal authorities to hurricane Katrina.

These intuitions are at least as strong and certain as any intuition we might have about what you should actually do about the single person and the five. To many trolley problems, as they are posed,[22] I think the right reaction is to regard it as simply indeterminate what the agent should do, and the only real moral issue raised by the problem is (as Dorothy Allison rightly said), how the situation in question was permitted to arise in the first place. The fact that lives are at stake is intended to compel us to reject this correct reaction, and make us feel that we simply must decide to do *something*—hence to decide that something is morally right and something else is morally wrong.

Yet trolley problem philosophers would regard us as missing the whole point of the problem if we even bothered to express any of the moral intuitions that don't directly involve saying what the agent should do. These philosophers are focusing our attention shortsightedly, even compulsively, solely on the question about what you should do in the immediate situation, as if that were the only thing moral philosophy has any reason to care about. In the context of the moral epistemology that goes with Sidgwickian style moral theory, the reasons for this restriction of attention are clear enough. But the fact that the clearer and more compelling intuitions about such a case are irrelevant to what interests them ought all by itself to make us distrust the philosophical value of the questions these philosophers are posing.[23]

[22] Here the qualification 'as they are posed' is also important, since I will be arguing that in the real world there would *always* be other facts that the philosopher is not permitting us to consider, and these would frequently determine what should be done. Often enough, these facts would dictate an answer directly contrary to the one the philosopher thinks our intuitions would dictate to the problem as he has posed it.

[23] This is a problem with much of moral philosophy generally, which behaves as if every moral problem must have a single right answer and as if it is moral philosophy's only job to say what it is. In real life, if a friend of yours faced a serious moral dilemma—for instance, whether to turn a guilty child in to the police or to lie to the police and let the child escape—I think most of us

2 *Why trolley problems mislead*

In real life, people go to a lot of trouble to arrange things so that no one will ever be placed in the position that, for example, the bystander in the train examples is placed. There are sound *moral* reasons why this is so, reasons that could be derived from FH and that are closely connected to Dorothy Allison's reaction that it is already immoral to ask anyone to decide between one person's life and five people's lives. The way I would put the point is to say that even if some choices do inevitably have the consequence that either one will die or five will die, there is nearly always something wrong with looking at the choice only in that way. But trolley problems are posed so that you know from the start that you are not supposed to look at them in any other way. You are given virtually no facts about the choice facing you except how many people will die if we choose each option and how you will bring about these deaths. Sometimes you even have it *stipulated* for you that there *are* no other relevant facts.

Such a stipulation cannot be regarded as either theoretically neutral or morally innocent. Suppose a moral philosopher posed for you the following problem: 'A group of white people are stranded on one rock and a group of black people are stranded on another. Before the rising tide covers both rocks, we could use a lifeboat to save either the white people or the black people. It is stipulated that there are no other relevant facts. Which group should we save?' Since the philosopher has told you nothing about how many people are in each group, nor even anything else about them except their skin color, I would hope that you would resist giving any answer at all to the philosopher's question. If you did have the 'intuition' that you should save the group whose skin color is the same as your own, then I would hope that you would resist answering on the basis of that 'intuition', and also that you would

would respect whatever choice the friend made, as long as we were sure that the friend had thought about the situation the right way, weighing appropriately both society's and their own child's moral claims on them. Any moral principle that dictated a single, unambiguous answer to the question what such a parent should do would be unacceptable simply because it did so. This is the valid point Sartre is making in his famous example of his student who had to choose between staying with his mother and joining the Resistance.

be heartily ashamed of yourself for having had that 'intuition' at all. Certainly you should not think that agreement with such an 'intuition' ought to serve as a test all moral principles ought to pass.

What is most objectionable here is the conversational implicature of the philosopher's question itself, in light of his outrageous stipulation that there are no other relevant facts. The question implies, namely, that you have been given enough information to answer the question as posed, or at least enough to have some 'intuition' worth reflecting on about what the answer should be. In this example, that implicature is morally offensive all by itself in a very obvious way. But most trolley problems differ from that example in that in them we have been given information about the situation that is at least *prima facie* morally relevant: the number of people on each rock is at least not so obviously and offensively irrelevant. Yet it may still be true that in trolley problems we have typically not been given enough information or the right information, to evoke intuitions that are worth anything. In the cases of *Tunnel* and *Bridge*, for example, in the real world there would simply *have* to be relevant facts about the situation beyond those we have been given, and in the real world what we should do would turn far more on those facts than they do on the facts we have been given. So the stipulation that these are the only relevant facts is not one we should accept at face value.

3 *Tunnel*

Here's what I mean: Trains and trolley cars are either the responsibility of public agencies or private companies that ought to be, and usually are, carefully regulated by the state with a view to ensuring public safety and avoiding loss of life. There ought to be, and usually are, provisions for physically preventing anyone from being in places where they might be killed or injured by a runaway train or trolley. If either the five or the single person in *Tunnel* are disobeying such rules by entering such dangerous areas, then they are behaving recklessly and are present there entirely at their own risk. Their claim to protection from harm is obviously far less than that of anyone who is in a permitted area. The claim of interlopers to protection in comparison to the claim of people in permitted areas is not increased proportionately (I submit it is not

increased at all) just because there are more of the interlopers. Further, mere bystanders ought to be, and usually are, physically prevented from getting at the switching points of a train or trolley. They would be strictly forbidden by law from meddling with such equipment for any reason, and they would be held criminally responsible for any death or injury they cause through such meddling.

These facts, if we were allowed to take account of them, would be decisive in a case like *Tunnel*: As mere bystanders, we would be forbidden by law to touch the switching points. (Unless railway officials have been criminally derelict in their duty, we would probably also be physically prevented from touching them.) In the real world there are not only good reasons for the existence of such laws, but in the real world there would also always be overwhelmingly good reasons for us to obey them. In real life, we would most likely not be sure we know how to operate the mechanism properly. For all we could know, our attempt to save the five might result in wrecking the runaway train and killing dozens of people on board. Further, if in real life we see five people in one tunnel and one person in another tunnel, we would have no way of knowing whether just a bit farther down the track from the one there are not many more people we would also be killing by switching the points. For all a mere bystander could know, the five people are interlopers, present on the track illegally and entirely at their own risk, while the single person is an employee of the railway who is there on the job. In the real world, these uncertainties would always be present, and the likelihood of their applying would never be merely negligible. That is an important reason why bystanders would be, and why they always should be, strictly forbidden by law from meddling with switching mechanisms.

Of course if in the situation as just described I were the bystander who correctly did nothing, I might nevertheless second-guess myself in my nightmares for years afterward, tormenting myself with the thought that there might have been something I could have done to save the five. This would be a natural human reaction to the horrible scene I had witnessed. But my feelings of guilt and self-reproach, though perhaps understandable, would be irrational. Far worse, however, and far more

irrational, would be the truly monstrous state of mind of the bystander who switched the points, killing the single person but saving the five, and then thought for the rest of his life that he had been treated unjustly when he was sentenced to prison for manslaughter—as he obviously should be.

4 *Bridge*

Many of the same observations apply here as apply to *Tunnel*, except that here the criminal wrongdoing of the bystander who acts to save the five is obviously far graver. For here the bystander surely must suppose that the single person, in walking on the *Bridge* over the train, is in a place where people have a perfect right to walk and to regard themselves as free from risk of harm from the deeds either of railway employees or meddling bystanders. The five, however, can be presumed to have entered a forbidden zone at their own risk. To kill the single person to save the five would in this case not be merely manslaughter but murder. The meddling bystander, sitting in his cell during the long years of his prison sentence, might have the consolation that many prestigious professors of moral philosophy at the world's leading universities think it worthwhile to reflect on the moral intuitions that put him where he is. I hope I may be forgiven for the ungenerous wish to deprive him of this one last consolation.

If a case such as *Tunnel* or *Bridge* were to occur in the real world, there would surely be an enraged public outcry against the railway system. The question whether one died or five died would be (and should be) of far less importance to the protesters than the fact that a runaway train had caused death. If it were further to come to light that the choice of who died had been at the mercy of a mere bystander, acting solely on his or her moral intuitions, this would only be further ground for public outrage. Relatively little attention would (or should) be paid to whether the bystander had chosen the death of one or the death of five. The protesters, in other words, would—and rightly so—care far less about the question that obsessively concerns the trolley problem philosophers than about relevant facts that these philosophers have lightheartedly stipulated away.

5 *Rights and entitlements*

Trolley problem philosophers seldom consider the kinds of entitlements to protection the people on the tracks might have, or might have forfeited, nor do they ever worry about our claim to be entitled, as mere bystanders, to choose who is to live and who is to die based only on our moral intuitions.[24]

Do they think the people on the tracks all necessarily have the same right to protection from harm, no matter how they came to be where they are? Are they supposing that the switches ought to be conveniently located where the general public can get at them, so as to have maximal opportunity to act on their moral intuitions in cases of emergency? Or, on the other hand, are they supposing instead that we know we are behaving both recklessly and illegally by touching the switches, but assuming that we would be justified nonetheless arrogating to ourselves the decision who should live and who should die (even when we can't be sure we aren't killing many others besides those we intend to kill)? In that case, the moral assumptions they are tacitly taking for granted are surely far more doubtful than any moral intuitions they could possibly hope to evoke in us.

One reason some philosophers might wish to abstract from every consideration of people's claims to protection from harm or entitlement to operate the switching mechanism is that they are tacitly assuming as a fundamental moral principle that all rights and claims must be derivative from the very moral principles they intend to use trolley problems to test. In that way, trolley problems seem theory-driven to the extent that they appear to assume that the basic subject matter of normative ethics consists solely in reckoning up the goodness and badness of states of affairs for particular people—though they also take into account the various causal relations human actions may have to those states of affairs. Some trolley problems seem little more than vehicles for representing certain abstract moral principles that are based

[24] A notable exception is Judith Thomson, *The Realm of Rights* (Cambridge: Gauthier, 1986; Harvard University Press, 1990), ch. 7, who does discuss the relevance of the question whether the people on the track are entitled to be there or have ignored some notice telling them to keep off the track. I thank Parfit for bringing this reference to my attention.

on that unargued assumption.[25] But the assumption is never stated, and one suspects that one aim of trolley problems might be to sneak the assumption past people's critical faculties as though *it* were simply given along with our moral intuitions about the problems themselves.

Clearly, however, it is defensible to hold that the value we attach to states of affairs is derivative from other values (such as the dignity of rational nature) which may also place significant constraints on when we value states of affairs and also the ways we compare and rank the value of states of affairs. For example, at least part of the value of the state of affairs consisting in a promise being kept is derivative from the obligatoriness of the principle that promises should be kept. The value of the state of affairs of the single person's being protected from harm by others is likewise derivative from this person's right to such protection, which (for someone who grounds rights on FH) is in turn derivative from the dignity of this person's humanity as an end in itself. It is so far from being true that all rights and entitlements are based on calculations about welfare that one excellent reason for arranging things so that people have rights and entitlements is simply to *make it false* that moral issues can ever be reduced to such calculations. FH is one moral principle, though by no means the only principle, that could provide such a reason.

Some people mistrust rights not based on welfare considerations because they think that such rights are typically appealed to only by privileged minorities (such as wealthy property owners) to justify prevailing social systems (such as those involving manifestly unequal distribution). These people may think that the assumptions built into trolley problems are right-headed, and my rejection of them is necessarily pernicious. But it would be naïve to think that this is the only meaning such rights could have. In the real world, policies favoring the welfare of

[25] It is true that the philosophers who use trolley problems do not necessarily accept this assumption, and some, such as Thomson and Philippa Foot, explicitly reject the idea that it is necessarily worse if more people die. As I have already mentioned, trolley problems sometimes seem to be designed to make the point that whether an action is morally right depends not only on the value of the states of affairs it produces, but also on the causal process through which it produces them. Still, the problems seem to assume a theory in which those two factors are the only relevant ones.