ON WHAT MATTERS

DEREK PARFIT

Draft of 23 January 2009

CONTENTS

VOLUME ONE 205,000 525 pages

INTRODUCTION BY SAMUEL SCHEFFLER

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

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

In his chapters on morality, Parfit aims to rechart the territory of moral philosophy. Students who take courses in the subject are usually taught that there is a fundamental disagreement between consequentialists, who believe that the rightness of an act is a function solely of its overall consequences, and Kantians, who argue – often with reference to one or another version of "the categorical imperative" -- that we have certain duties that we must fulfill whether or not doing so will produce optimal results in consequentialist terms. Although both consequentialist and Kantian views are acknowledged to admit of many variations and refinements, the division between them is assumed by most philosophers, including most consequentialists and Kantians, to be deep and fundamental.

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

Morals, are among the most widely discussed texts in the history of moral philosophy, Parfit's engagement with these texts yields a wealth of fresh observations and insights.

As is evident from his Preface, Parfit's attitude toward Kant is complex and defies easy summary. He describes him as "the greatest moral philosopher since the ancient Greeks" (146), and says that "in the cascading fireworks of a mere thirty pages, Kant gives us more new and fruitful ideas than all the philosophers of several centuries" (153). He quickly adds, however, that "[o]f all the qualities that enable Kant to achieve so much, one is inconsistency" (153-4). Whereas many commentators explicitly present themselves either as critics of Kant or as defenders of his view, Parfit's approach is different. He treats Kant's texts as a rich fund of claims, arguments, and ideas, all of which deserve to be treated with the same seriousness that one would accord the ideas of a brilliant contemporary, but many of which require clarification or revision, and some of which are simply unworkable. Parfit examines a wide range of these claims, arguments, and ideas, subjecting them to a level of scrutiny that is remarkable for its unwavering focus and analytic intensity. His primary aim is neither to defend Kant nor to criticize him, but rather to determine which of his ideas we can use to make progress in moral philosophy. At the end of the day, it is progress that is Parfit's real goal. As he says in explaining why one of Kant's formulations should be revised, "After learning from the works of great philosophers, we should try to make some more progress. By standing on the shoulders of giants, we may be able to see further than they could" (269).

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

Parfit, by contrast, sees great potential in the Formula of Universal Law. Swimming against the prevailing tide of interpretive opinion, he insists that the FUL "*can* be made to work," and he argues that when "revised in some wholly Kantian ways, this formula is … remarkably successful" (263). Indeed, he goes so far as to say that a suitably revised version of this formula "might be what Kant said he was trying to find: the supreme principle of morality" (312).

The revised version of the Formula of Universal Law that Parfit favors states that "Everyone ought to follow the principles whose universal acceptance everyone could rationally will." With its appeal to a kind of universal choice or agreement, this formulation qualifies as a form of "contractualism," and Parfit refers to it as the "Kantian

Contractualist Formula." So interpreted, the Kantian position invites comparison with contemporary versions of contractualism, especially those versions that are themselves of broadly Kantian inspiration. John Rawls's appeal to principles that would be chosen behind a veil of ignorance is one example, though Rawls applied this device almost exclusively to the choice of principles of justice for the basic structure of society. He never followed up on the idea, which he had briefly entertained in *A Theory of Justice*, that the same device might be applied to the choice of moral principles more generally. Parfit nevertheless subjects this idea to severe criticism, and concludes that it is much less promising as a general account of morality than the version of contractualism developed by Thomas Scanlon.

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

As I have noted, the opposition between the Kantian and consequentialist positions is usually taken to be deep and fundamental, and the contemporary contractualisms of both Rawls and Scanlon are motivated to a significant degree by the desire to articulate a compelling alternative to consequentialism. Yet Parfit argues that Kantian contractualism actually implies a version of "Rule Consequentialism," which holds that "everyone ought to follow the principles whose universal acceptance would make things go best." The principles whose universal acceptance everyone could rationally will, he maintains, just are these "optimific" rule-consequentialist principles. Accordingly, Kantian Contractualism and Rule Consequentialism can be combined to form a view that he calls Kantian Rule Consequentialism: "Everyone ought to follow the optimific principles, because these are the only principles that everyone could will to be universal laws" (377). Although this position is consequentialist in the content of its claims about the principles that people ought to follow, it is more Kantian than consequentialist in its account of why we should follow these principles. We should follow them because their universal acceptance is something that everyone could rationally will, and not because, as consequentialists would have it, all that ultimately matters is that things should go for the best.

Since Kantian Contractualism implies Rule Consequentialism, and since some versions of Kantian Contractualism coincide with some versions of Scanlonian Contractualism, versions of all three positions can also be combined. The resulting "Triple Theory" holds that an "act is wrong just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable (379)". The upshot

of these various possibilities of convergence, Parfit believes, is that it is a mistake to think that there are deep disagreements among Kantians, contractualists, and consequentialists. Instead, "these people are climbing the same mountain on different sides" (385).

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

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

Rejecting these "bleak" views, Parfit argues that we should instead accept an objective, value-based theory, according to which reasons for action are provided by the values that those acts would realize or fulfill (or, as he puts it, by the facts that make certain things worth doing for their own sake or make certain outcomes good or bad). Understood in this way, judgments about reasons are more fundamental than judgments about rationality, for we are rational, in Parfit's view, when we respond to reasons or apparent reasons, and our acts are rational when, if our beliefs were true, we would be doing what we had good reasons to do. This contrasts with a number of popular accounts of practical rationality, such as those that identify it with the maximization of expected utility, for example, or those that interpret practical *ir*rationality as a form of inconsistency.

As Thomas Scanlon observes in his contribution, the idea that reasons have priority over rationality also conflicts with Kant's views. For Kant, both the authority and the content of the categorical imperative are to be understood with reference to the requirements of rational agency rather than to some independent conception of the reasons that people have. As Scanlon describes the Kantian view, which he calls "Kantian constructivism about reasons": "Claims about reasons (more exactly, about what a person must see as reasons) must be grounded in claims about rational agency, claims about what attitudes a

person can take, consistent with seeing herself as a rational agent. Justification never runs in the other direction, from claims about reasons to claims about what rationality requires" (S 7).

Parfit, like Scanlon, rejects Kantian constructivism about reasons and, as Scanlon points out, all of the moral theories whose convergence Parfit seeks to demonstrate are framed in such a way as to "appeal to an idea of 'what one can rationally will' that presupposes an independently understandable notion of the reasons that a person has and their relative strength" (S 7). This distinguishes these theories from Kant's own views and also from the views of some prominent contemporary Kantians, such as Christine Korsgaard. As Parfit acknowledges, his reliance on a primitive and "indefinable" notion of "reasons," and his concomitant commitment to the existence of irreducibly normative truths, both about reasons and about morality, makes his view a version of what Korsgaard has called "dogmatic rationalism." As such, it would be resisted not only by Kantian constructivists like Korsgaard but also by proponents of some very different meta-ethical outlooks, such as various forms of naturalism and non-cognitivism.

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

Parfit realizes, of course, that many philosophers do not accept the existence of irreducibly normative truths in his sense. Nihilists and error theorists hold that all normative claims are false. Naturalists hold that normative facts can be reduced to natural facts. Non-cognitivists hold that normative claims, despite their importance in human life, do not function as statements of fact at all. Parfit offers little direct discussion of nihilism, but he provides forceful criticism of a number of influential versions of naturalism and non-cognitivism, including the views of Simon Blackburn, Richard Brandt, Allan Gibbard, Richard Hare, and Bernard Williams. None of these views, he argues, can adequately account for the normative dimension of our thought; on all such views, normativity proves to be illusory. It simply "disappears." In effect, Parfit appears to believe that all such views tend toward nihilism, and that nihilism is the only genuine alternative to the recognition of irreducibly normative truths. That may be why, more

than once, he refers to one or another version of naturalism or non-cognitivism as a "bleak" view, the same term he uses to describe the desire-based theory of reasons. Nor is he persuaded by Korsgaard's Kantian objections to "realism" about normativity. Contrary to what she maintains, he asserts, normativity does not have its source in the will, but instead consists in the existence of irreducibly normative truths about what we have reason to do, want, and believe.

As will be apparent, Parfit's aims in his discussions of reasons and normativity are very different from those he pursues in discussing substantive moral theories. In the moral case, his aim is to demonstrate that certain putatively opposing views may actually converge, so that apparent disagreement among them evaporates. But in his discussions of different views about reasons and normativity, convergence is not on the agenda. A value-based theory of reasons should be accepted, he argues, and desire-based theories should simply be rejected. Similarly, his form of Cognitivism should be accepted in preference to all forms of Naturalism and Non-Cognitivism. Parfit is clearly troubled by substantive moral disagreement, for he thinks it threatens to undermine our conviction that there is such a thing as moral truth. That is why he is so strongly driven to demonstrate the possibility of convergence. However, his response to meta-ethical or meta-normative disagreement is different. Here he simply attempts to determine which of the contending views is correct. Yet to the extent that the substantive theories whose convergence Parfit seeks to demonstrate all presuppose his views about reasons and normativity, the frankly contested character of those views may call into question the significance of the convergence he describes at the substantive moral level. Those who reject value-based theories of reasons, and those who accept one or another form of naturalism or non-cognitivism or constructivism, may be unmoved by a moral consensus that depends on accepting the very meta-ethical views that they reject. So one challenge for Parfit is to demonstrate that the significance of the convergence for which he argues is not undermined by its dependence on claims, such as those concerning reasons and normativity, about which there is no convergence.

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

The issue is a tricky one. As Scanlon notes, Parfit is forthright about his willingness, in developing a "Kantian" position, to depart from Kant's actual views whenever he thinks he can improve upon them. As Parfit says, "We are asking whether Kant's formulas can

help us to decide which acts are wrong. If we can revise these formulas in ways that improve them, we are developing a Kantian moral theory" (267). In his reply to Scanlon, he is similarly explicit about the fact that his argument for the convergence of Kantian Rule Consequentialism and Scanlonian Contractualism "does not apply to the view stated in Scanlon's book" (R 30), but rather to a version of that view that has been revised in ways that Parfit takes to strengthen it.

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

Susan Wolf appears to harbor doubts of both of these kinds about Parfit's claims of convergence. Of Parfit's ambition to reconcile the Kantian, consequentialist, and contractualist traditions, she writes: "[I]nsofar as his concluding remarks are meant to suggest that the values these different traditions emphasize can be interpreted and ordered in such a way as to eliminate the tensions among them, or that it would be in the spirit of these traditions' greatest exponents to accept revisions and qualifications to their stated views that would ultimately reconcile them with their opponents, Parfit departs from the explicit positions of any of the philosophers whose work he discusses, in a way that seems to me both interpretively implausible and normatively regrettable" (W 2-3). Wolf's view is that the Kantian, consequentialist, and contractualist traditions embody divergent evaluative perspectives, each of which has something important to contribute but which are in genuine tension with one another. These tensions reflect broader tensions within our moral thought itself. As such, she believes, they are ineliminable and not to be regretted. Any unified principle of the kind Parfit seeks will perforce be a matter of compromise rather than complete convergence, and any such principle will inevitably leave out something of value. Wolf presses this last point with special reference to Parfit's version of Kantianism, which, she argues, scants the importance of autonomy in Kant's own moral philosophy.

Barbara Herman too believes that Parfit's position departs from Kant's in fundamental ways. However, while Wolf expresses doubts about the very idea that morality rests on a unified principle of the kind that Parfit seeks, Herman is sympathetic to Kant's own unified account and believes that Parfit's theory is an unstable mixture of disparate elements. More specifically, she argues that Parfit employs a "hybrid" methodology that incorporates some Kantian features but nevertheless has "a strongly consequentialist cast." (H 1) Although Parfit's intention is to preserve what is most persuasive in Kant's view while avoiding some of the apparently unwelcome implications of that view, Herman believes that there is such a deep "mismatch" between the Kantian and consequentialist methodologies that the attempt to combine them inevitably distorts

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

Like Herman, Allen Wood also argues that Parfit's philosophical methodology departs from Kant's in important ways, although he focuses on different aspects of Parfit's approach than Herman does. Wood believes that Parfit employs a method originated by Sidgwick, which sets itself the goal of providing a "scientific" ethics. The idea is to systematize our commonsense moral opinions, correcting them when necessary, with the aim of arriving at a precise set of principles that can be used algorithmically to yield a determinate moral verdict about how one should act in any conceivable situation. Wood believes that such otherwise diverse philosophers as Kant, Bentham, and Mill employ a very different method, which he himself regards as preferable to the one he ascribes to Sidgwick and Parfit. This alternative method begins not with commonsense intuitions but rather with a fundamental principle that serves to articulate some basic value. General moral rules or duties are then derived non-deductively from the fundamental principle. These rules or duties represent an attempt to interpret the implications of the fundamental value in the conditions of human life. The rules or duties themselves admit of exceptions and require interpretation, and their application to particular cases calls for the exercise of judgment and cannot be codified in precise rules or principles. So, on the one hand, the Kantian method as Wood understands it gives less weight than the Sidgwickian method to commonsense moral intuitions; but, on the other hand, it regards as "hopeless" the aim of constructing a "scientific" ethics that can provide an algorithm for moral decision-making.

Wood believes – though Parfit's reply suggests that he would not accept this diagnosis – that the difference of method just described underlies some disagreements between Parfit and him concerning the proper interpretation of Kant's Formula of Humanity. He thinks it also underlies their sharply divergent attitudes toward one familiar type of philosophical argument. This type of argument uses our intuitive reactions to stylized and sometimes complex hypothetical examples to test candidate moral principles. Wood refers to all such examples as "trolley problems," whether or not they involve actual trolleys, in mock *hommage* to the famous case first introduced into the philosophical literature by Philippa Foot. Parfit makes frequent use of such examples in constructing his arguments. For instance, his argument for the convergence of Kantian Contractualism and Rule Consequentialism turns crucially on some claims about what a person could rationally agree to in situations where one course of action would impose a burden on the person himself and the only alternative would impose burdens on others. Parfit illustrates and defends these claims with reference to a series of hypothetical examples involving burdens of different sizes and types imposed in a range of different hypothetical circumstances. He seeks to marshal our intuitive responses in these cases to show 1) that each person could rationally will the universal acceptance of the consequentially optimific principles, even when those principles would impose some burden on the person himself, and 2) that there are no other principles whose universal acceptance everyone could rationally choose. Parfit evidently believes that the use of hypothetical examples can help to clarify the issues that are at stake in complex moral choices and enable us to make progress in moral argument. Wood, by contrast, regards "trolley problems" as "worse than useless for moral philosophy" (W 14), and the majority of his essay is given over to an extended critique of the ways in which reliance on such problems leads moral philosophers astray.

To the extent that other people share Wood's reservations about appealing to hypothetical examples in moral philosophy, Parfit's extensive reliance on such examples may be a source of resistance to his arguments. Of course, even those who do not endorse Wood's radical rejection of all such appeals may find themselves disagreeing with Parfit's reactions to some of the specific examples he discusses, although Parfit anticipates many potential disagreements and exhibits great resourcefulness in attempting to defuse them. Yet Parfit himself points out that our reactions to some of these cases may depend, for example, on whether we accept a desire-based or a value-based theory of reasons. Since he hopes to use our reactions to support his claim of convergence among different moral theories, this kind of variation represents one way in which disagreements about reasons and rationality, like metaethical disagreements about the nature of normative judgment, threaten to destabilize the moral consensus that Parfit aims to establish. As I have already said, Parfit's response to this threat is not to look for additional convergence at the level of meta-ethical theories or theories of reasons and rationality, but rather to argue that the alternatives to his Non-Naturalist Cognitivism and to the value-based theory of reasons should be decisively rejected. This is a different way

of eliminating or at least taking the sting out of disagreement: by demonstrating that there is only one position that we can reasonably accept.

The drive to eliminate disagreement – whether by establishing convergence or through a decisive demonstration of the inadequacy of competing views – is a defining feature of Parfit's work. It is sometimes marked by a sense of urgency. One place where this emerges is in his reply to Susan Wolf. Wolf takes Parfit to be trying to show "that there is a single true morality, crystallized in a single supreme principle that these different traditions may be seen to be groping towards, each in their own separate and imperfect ways" (W 2). She herself says, by contrast, that "it would not be a moral tragedy if it turned out" (W 3) that morality did not have such a unifying principle. In response, Parfit agrees that "we do not need a single supreme principle." But, he adds, "we do need a *single true morality*," for "if we cannot resolve our disagreements, that would give us reasons to doubt that there are *any* true principles. There might be nothing that morality *turns out to be*" (R 11). It is, perhaps, the spectre of this "bleak" possibility, and the even bleaker possibility that, as Parfit worries, nothing at all may matter, that is responsible for the sense of urgency with which he pursues the elimination of disagreement. Whether or not one shares his assessment of the threat posed by deep disagreement, one cannot fail to be impressed by the extraordinary ingenuity and the sheer intellectual intensity with which he pursues his goal. His rich and challenging discussion, helpfully illuminated by his exchanges with Barbara Herman, Thomas Scanlon, Susan Wolf, and Allen Wood, casts familiar debates in a fresh and unfamiliar light, and opens up many fruitful new lines of inquiry for philosophers to investigate. Nobody who is interested in the theory of morality, rationality, or normativity will want to ignore this brilliant, provocative, and tenaciously argued book.

PREFACE

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

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

Kant is the greatest moral philosopher since the ancient Greeks. Sidgwick's *Methods* is, I believe, the best book on ethics ever written. There are some books that are greater achievements, such as Plato's *Republic* and Aristotle's *Ethics*. But Sidgwick's book contains the largest number of true and important claims. It is not surprising that, though a less great philosopher than Plato, Aristotle, Hume, and Kant, Sidgwick could write a better book. Sidgwick lived later. Unlike later poets or playwrights, who have no advantages over Homer or Shakespeare, later philosophers do have advantages, since philosophy makes progress. [1]

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

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

> the critical philosophy must remain confident of its irresistible propensity to satisfy the theoretical as well as the moral, practical purposes of reason, confident that no change of opinions, no touching up or reconstruction into some other form, is in store for it; the system of the *Critique* rests on a fully secured foundation, established forever; it will prove to be indispensable too for the noblest ends of mankind in all future ages; [2]

Sidgwick writes:

> The book solves nothing, but may clear up the ideas of one or two people, a little. [3]

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

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

Sidgwick also refers to his 'one damning defect of longwinded & difficult dullness.' [4]

This last phrase is too severe. Though Sidgwick's book is long, and some of its chapters can now be ignored, [5] it is not longwinded. Sidgwick seldom repeats himself, and he makes many important points concisely, and only once. Nor is Sidgwick's book difficult. Some of his claims and arguments are complicated, but they are nearly all clearly written. [6]

Sidgwick's dullness needs more discussion. Whitehead was so bored by Sidgwick's *Methods* that he never looked at another book on ethics. [7] But after reading a collection of Sidgwick's memoirs and letters, Keynes remarked, 'I have never found so dull a book so absorbing'. It is worth quoting from this book. Discussing the Church of England, Sidgwick writes:

> At Cambridge I get into the way of regarding it as something that once was alive and growing, but now exists merely because it is a pillar or buttress of uncertain value in a complicated edifice that no one wants just now to take to pieces. Here however, I feel rather as if I were contemplating a big fish out of water, propelling itself smoothly and gaily over the high road. [8]

Here are two more passages:

> There is no doubt that men in England fall in love chiefly in abnormal periods: when on a reading party, or at the seaside, or at a foreign hotel, or at Christmas, or any other occasion when something, either external circumstances or any dominant emotion, thaws the eternal ice. The misfortune is that if these casual thaws do not last long enough, all the advantage gained is lost; two lines of life that causally intersected diverge perhaps for ever, and the frost sets in with redoubled force. [9]
>
> I am bearing the burden of humanity in the lap of luxury, and in consequence not bearing it well. After all, Pascal was practically right: if one is to embrace

> infinite doubt, if it is to come into our bowels like water, and like oil into our bones, it ought to be upon sackcloth and ashes and in a bare cell, and not amid '47 port and the silvery talk of W. G. Clark. When I go to my rooms I feel strange, ghastly, that is why I write to you. But there again---if one allows this consciousness 'the time is short' to grow and get too strong, it seems to fold up all life into a feverish moment.
>
> The world shall feel my impulse or I die.
>
> Think of all the second-rate men who have said this and died---and---Who cares?
>
> Butterflies may dread extinction.
>
> This is a strange mood for me. But at Trumpington today I brushed away a spider's life and said 'This is sentience.' What am I more than elaborate sentience? [10]

Sidgwick could be amusing, and his conversation was described as 'like the sparkling of a brook whose ripples seem to give out sunshine'. But the first edition of the *Methods* contains only a few jokes, some of which Sidgwick later removed. [11] Much of the book, however, is well-written. For example:

> to suppose. . . that the ideal of 'obeying oneself alone' can be even approximately realized by Representative Democracy is even more patently absurd. For a representative assembly is normally chosen only by a part of the nation, and each law is approved by only a part of the assembly: and it would be ridiculous to say that a man has assented to a law passed by a mere majority of an assembly *against* one member of which he has voted. [12]

More soberly:

> . . . the Cosmos of Duty is thus really reduced to a Chaos, and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure. [13]

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

> I cannot fall back on the resource of thinking myself under a moral necessity to regard all my duties *as if they were* commandments of God, although not entitled to hold speculatively that any such Supreme Being really exists. I am so far from feeling bound to believe for purposes of practice what I see no ground for holding as a speculative truth, that I cannot even conceive the state of mind which these words seem to describe, except as a momentary half-witted irrationality, committed in a violent access of philosophic despair. [14]

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

> . . . . the selfish man misses the sense of elevation and enlargement given by wide interests; he misses the more secure and serene satisfaction that attends continually on activities directed towards ends more stable in prospect than an individual's happiness can be: he misses the peculiar rich sweetness, depending upon a sort of complex reverberation of sympathy, which is always found in services rendered to those whom we love and who are grateful. He is made to feel in a thousand various ways. . . the discord between the rhythms of his own life and of that larger life of which his own is but an insignificant fraction. [15]

Another passage ends:

> . . . even a man who said 'Evil be thou my good' and acted accordingly might have only an obscured consciousness of the awful irrationality of his action--- obscured by a fallacious imagination that his only chance of being in any way admirable, at the point of which he has now reached in his downward course, must lie in candid and consistent wickedness. [16]

Sidgwick warned his friends that, because his book attempts to achieve 'precision of thought', it 'cannot fail to be somewhat dry and repellent'. [17] But this precision is often finely expressed. Discussing friendship, for example, Sidgwick writes of

> the sympathy that is not quite admiration with which Common Sense regards all close and strong affections; and the regret that is not quite disapproval with which it contemplates their decay. [18]

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

> It may be said that a child owes gratitude to the authors of its existence. But life alone, apart from any provision for making life happy, seems a boon of doubtful value, and one that scarcely excites gratitude when it was not conferred from any regard for the recipient. [19]

> . . . there seems to be no justice in making A happier than B, merely because circumstances beyond his control have first made him better. [20]

> Thus the Utilitarian conclusion, carefully stated, would seem to be this: that the opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric. [21]

> . . . really penetrating criticism, especially in ethics, requires a patient effort of sympathy which Mr Bradley has never learned to make, and a tranquillity of temper which he seems incapable of maintaining.[22]

> [The book] seems smashing, but he loses by being over-controversial. There should be at least an affectation of fairness in a damaging attack of this kind. [23]

Sidgwick's irony can make him seem stuffy, when in fact he is being subversive. Bernard Williams had been misled, for example, when he wrote that Sidgwick's discussions of sexual morality, though sometimes mildly adventurous, 'make fairly uncritical use of a notion of purity'. [24] Sidgwick does ask 'What, then, is the conduct that Purity forbids?' But if we read him carefully, we find that his answer is: Nothing. In a book published in England in 1874, it was more than mildly adventurous to argue, though in guarded terms, that there is no moral objection to indulging in sexual pleasure for its own sake. [25]

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

> Have been reading Comte and Spencer, with all my old admiration for their intellectual force and industry and more than my old amazement at their fatuous self-confidence. It does not seem to me that either of them knows what self-criticism means. I wonder if this is a defect inseparable from their excellences. Certainly I find my own self-criticism an obstacle to energetic and spirited work: but on the other hand I feel that whatever value my work has is due to it. [26]

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

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

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

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

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

> incessantly refines, qualifies, raises objections, answers them, and then finds further objections to the answer. Each of these objections, rebuttals, rejoinders, and surrejoinders is in itself admirable, and does infinite credit to the acuteness and candour of the author. But the reader is apt to become impatient; to lose the thread of the argument; and to rise from his desk finding that he has read a great deal with constant admiration and now remembers little or nothing. [27]

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

Criticizing himself again, Sidgwick writes:

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

This remark is also too severe. Sidgwick is in several ways original. But that is not what makes him great. Other philosophers, like Kant and Hume, are more original, and more brilliant. These philosophers are like Newton and Einstein: geniuses of the clearest kind. Sidgwick is more like Darwin. He had what has been called 'good sense intensified almost to the point of genius'. [28] In the *Methods*, as Broad claims, 'almost all the main problems of ethics are discussed with extreme acuteness'. [29] And Sidgwick gets very many things right. He gives the best critical accounts of three of the main subjects in ancient and modern ethics: hedonism, egoism, and consequentialism. And in the longest of his book's four parts, he also gives the best critical account of pluralistic non-consequentialist common sense morality. Though Sidgwick makes mistakes, some of which I mention in a note, he does not, I believe, make many. [30] These facts make Sidgwick's *Methods* the book that it would be best for everyone interested in ethics to read, remember, and be able to assume that others have read.

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

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

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

opposed to Kant, and saying how my attitude has changed, I may perhaps persuade some other people not to ignore Kant, as I nearly did.

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

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

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

There are deeper problems. When I became obsessed with Kant, I tried to restate more clearly some of Kant's main claims and arguments, and found this task very frustrating. I couldn't fit Kant's claims together in a coherent view, and many of Kant's arguments seemed to be obviously invalid or unsound. It would have helped me to know that even some of Kant's greatest admirers have similar feelings. Onora O'Neill, for example, calls the *Groundwork* 'the most exasperating' of Kant's books. [31]

It would also have helped me to know that Kant did not have a single, coherent theory. When we ask whether Kant accepts or rejects some claim, the answer is often 'Both'. As Kemp Smith writes, 'citation of single passages is quite inconclusive'. [32] For example, though Kant writes that 'a human being's duty at each instant is to do all the good in his power', [33] he is not really, as this claim implies, an Act Consequentialist. Rawls remarks that, when he tried to understand Kant's texts, 'I assumed there were never plain mistakes, not ones that mattered anyway'. [34] But there must be mistakes, since Kant makes many conflicting claims, and such claims cannot all be true. As Kemp Smith points out, Kant often 'flatly contradicts himself' and 'there is hardly a technical term which is not employed by him in a variety of different and conflicting senses. He is the least exact of the great thinkers.' [35] (To avoid provoking Hegelians, we should perhaps say 'one of the least exact'.)

'Consistency', Kant writes, 'is a philosopher's greatest duty'. [36] That is not true. Originality and clarity are at least as important. And Kant's greatness chiefly consists in his having many original and fruitful ideas. If Kant had always been consistent, he could not have had all these ideas.

When I first re-read Kant, what I found most irritating was not Kant's obscurities and

inconsistencies, but a particular kind of overblown, false rhetoric. For example, Kant writes:

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

I didn't mind the exaggeration in the first sentence here. We can switch the volume down, turning 'all of them had to' into 'some of them did'. But since I knew that Kant believed in a Categorical Imperative, I was surprised by Kant's second sentence. I asked a Kantian, 'Does this mean that, if I don't give myself Kant's Imperative as a law, I am not subject to it?' 'No', I was told, 'you have to give yourself a law, and there's only one law.' This reply was maddening, like the propaganda of the so-called 'People's Democracies'of the old Soviet bloc, in which voting was compulsory and there was only one candidate. And when I said 'But I haven't given myself Kant's Imperative as a law', I was told 'Yes you have'. This reply was even worse. My irritation at such claims may have left some traces in this book.

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

> Suicide is the most abominable of the crimes that inspire horror and hatred. . . he who so utterly fails to respect his life . . . can in no way be restrained from the most appalling vices. . . [37]

But he also said

> In the Stoic's principle concerning suicide there lay much sublimity of soul: that we may depart from life as we leave a smoky room. [38]

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

> If freedom is the condition of life, it cannot be employed to abolish life. . . Life is supposedly being used to bring about lifelessness, but that is a self-contradiction. [39]

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

> cut off the source of the goodness of this end---it is no longer really an end at all, and it is no longer rational to pursue it. [40]

This conclusion arrives too late.

For another example, consider Kant's claim that, if we tell some lie 'even to achieve some really good end', we 'violate the dignity of humanity in our own person' and make ourselves a 'mere deceptive appearance of a human being', who has 'even less worth than if he were a mere thing'. [41] We should ignore such outbursts. On the very next page Kant suggests that, if we are asked by an author whether we like his work, we may be permitted to say what he expects.

Kant is sometimes thought of as a cold, dry, rationalist. But he is really an emotional extremist. As Sidgwick writes, 'Oh, how I sympathize with Kant! with his passionate yearning for synthesis and condemned by his reason to criticism. . .' [42] Kant seldom uses words like 'most', 'many', 'several', or 'some', preferring to write only 'all' or 'none'. Kant uses 'good', he says, to mean 'practically necessary'. And he seldom uses the concept of a reason: a fact that merely *counts in favour* of some act, since his preferred normative concepts are *required, permitted,* and *forbidden*. Temperamentally, I am an extremist too, who has to struggle to be more like Sidgwick.

Oxford University once had a useful marking grade: *Alpha Gamma*. As everyone should agree, Kant's books are pure Alpha Gamma, containing nothing that is *Beta*, or mediocre. Our disagreement should be only about how much of what Kant wrote is Alpha, and how much is Gamma. And if we have understood what is Alpha, we can ignore what is Gamma. [43]

I still believe that Kant is too close to Hume, being a more dangerous Anti-Rationalist because, unlike Hume, he seems to be exalting what he calls *Pure Reason*. And Kant's influence has been, I believe, in some other ways bad. But he is very great, and his influence has been, in other and less obvious ways, good. Though Kant makes many claims that are false, and many of his arguments fail, he also gives us some profound truths. Like Sidgwick, I sometimes find him 'quite a revelation.' [44] Kant's books are very thought-provoking, containing many remarks that suggest a whole new line of thought. As Rawls writes 'Part of the wonderful character of the works we study is the depth and variety of ways they can speak to us.' [45]

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

I turn now to the other people from whom I have learned most. When I was young,

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

As well as finding, in the long-dead Sidgwick, someone who had greater hopes for practical and moral philosophy, I was encouraged to find some living philosophers who had such hopes. I was encouraged most by Thomas Nagel, and in particular by Nagel's claims about reasons, and about irreducibly normative truths. [46] I have also learnt a great deal from Tim Scanlon. I often cannot remember whether some thought was mine or his. I dedicate this book to these two people.

Many other people have helped me to write this book. I am grateful to Christine Korsgaard, whose impressive books led me to reread Kant, and whose critique of what she calls 'dogmatic rationalism' helped to rouse me from my undogmatic slumbers. I have also been greatly helped by the remarkable recent series of other books and articles on or inspired by Kant, by such writers as Barbara Herman, Allen Wood, Thomas Hill, Onora O'Neill, Paul Guyer, Henry Allison, Thomas Pogge, and Samuel Kerstein.

Of the many people who have commented on drafts of this book, I must thank first . . .

SUMMARY

VOLUME ONE

PART ONE REASONS

CHAPTER 1 NORMATIVE CONCEPTS

1 Sufficient and Decisive Reasons

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

2 Reason-Involving Goodness

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

CHAPTER 2 OBJECTIVE THEORIES

3 Two Kinds of Theory

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

4 Responding to Reasons

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

5 State-given Reasons

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

6 Hedonic Reasons

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

When we are in pain, what is bad is not our sensation but our conscious state of having a sensation that we dislike. It is similarly good to have sensations that we like. Such *hedonic likings* or *disliking* cannot be rational or irrational, since we have no reasons to like or dislike these sensations. We also have *meta-hedonic desires* about our own and other people's pleasures and pains. Such desires or preferences *can* be rational or irrational, since we have strong reasons to have them. It is our hedonic likings and dislikings, not our meta-hedonic desires, that make these conscious states good or bad; so the examples of pleasure and pain do not support the view that our desires can give us reasons, and can make their objects good.

7 Irrational Preferences

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

CHAPTER 3 SUBJECTIVE THEORIES

8 Subjectivism about Reasons

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

9 Why People Accept Subjective Theories

Since so many people believe that *all* practical reasons are desire-based, aim-based, or choice-based, how could it be true that, as objective theories claim, there are *no* such reasons? How could all these people be so mistaken? There are several possible explanations, since there are several ways in which our desires or aims may seem to give us reasons.

10 Analytical Subjectivism

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

11 The Agony Argument

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

CHAPTER 4 FURTHER ARGUMENTS

12 The All or Nothing Argument

Subjective theories could also imply that we have decisive reasons to cause ourselves to be in agony for its own sake, to waste our lives, and to try to achieve other bad or worthless aims. In response to this objection, Subjectivists might claim that, for some desire or aim to give us a reason, we must have some reason to have this desire

or aim. But these people cannot defensibly make this claim. On subjective theories, all that matters is *whether* some act would fulfil our present fully informed desires or aims. It is irrelevant *what* we want, or are trying to achieve. Either *all* of these desires give us reasons for acting, or *none* of them do. Since it is clear that some of these desires could not give us reasons, we should conclude that none of them do.

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

13 The Incoherence Argument

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

14 Reasons, Motives, and Well-Being

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

15 Arguments for Subjectivism

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

CHAPTER 5 RATIONALITY

16 Practical and Epistemic Rationality

We are rational insofar as we respond well to reasons or apparent reasons. We have some *apparent* reason when we have false beliefs about the relevant facts whose truth would give us some reason. Our desires and acts are rational when, if our

beliefs were true, we would have sufficient reasons to have these desires, and to act in these ways. Some people add that, for our desires or acts to be rational, they must depend on rational beliefs. This claim is misleading, and not worth making.

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

17 Beliefs about Reasons

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

18 Other Views about Rationality

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

CHAPTER 6 MORALITY

19 Sidgwick's Dualism

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

We should reject Sidgwick's argument. We ought to assess the strength of all our reasons from our actual, personal point of view, and we do not need a neutral point of view. We should also revise Sidgwick's conclusion. We have personal and partial reasons to be specially concerned, not only about our own well-being, but also about the well-being of certain other people, such as our close relatives and those we love. These are the people to whom we have *close ties*. We also have impartial reasons to care about anyone's well-being, whatever that person's relation

to us. Though there are truths about the relative strengths of these two kinds of reason, Sidgwick's view is partly right, since these comparisons are, even in principle, very imprecise. As *wide value-based objective* theories claim, when one of two possible acts would be impartially better, but the other act would be better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either way.

20 The Profoundest Problem

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

Though reasons are more fundamental, the rest of these chapters are about morality. But these chapters also discuss reasons. Several moral principles and theories appeal to claims about what, in actual or imagined situations, we would have sufficient reason or most reason to consent to, or agree to, or to want, or choose, or do.

CHAPTER 7 MORAL CONCEPTS

21 Acting in Ignorance or with False Beliefs

By distinguishing several senses of 'ought morally' and 'wrong', we can recognize some important truths and avoid some unnecessary disagreements. Acts can be wrong in *fact-relative*, *evidence-relative*, *belief-relative*, and *moral-belief-relative* senses. Facts about these kinds of wrongness provide answers to different questions. When what we ought to do depends on the goodness of our act's effects, we ought to try to do, not what would in fact be best, but what would be *expectably-best*.

22 Other Kinds of Wrongness

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

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

PART TWO PRINCIPLES

CHAPTER 8 POSSIBLE CONSENT

23 Coercion and Deception

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

24 The Consent Principle

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

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

25 Reasons to Give Consent

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

26 A Superfluous Principle?

According to some writers, even if the Consent Principle is true, this principle adds nothing to our moral thinking. What is morally important is not the fact that people could not rationally consent to certain acts, but the various facts that give these people decisive reasons to refuse consent. When applied to acts that affect

only one person, this objection has some force. But when our acts would affect many people, if there is only one possible act to which everyone could rationally consent, this fact would give us a strong reason to act in this way, and would help to explain why the other possible acts would be wrong. It is also worth asking whether we could achieve Kant's ideal.

27 Actual Consent

It is wrong to treat people in certain ways if these people either do not, or would not, actually consent to these acts. Such acts are wrong even if these people could have rationally given their consent. That is no objection to the Consent Principle, which claims to describe only one of the facts that can make acts wrong.

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

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

28 Deontic Beliefs

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

29 Extreme Demands

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

CHAPTER 9 MERELY AS A MEANS

30 The Mere Means Principle

It is wrong, Kant claims, to treat any rational being merely as a means. We treat people in this way when we both use these people and regard them as mere tools, whom we would treat in whatever way would best achieve our aims. On a

stronger version of Kant's principle, it is wrong to treat people merely as a means, or to *come close* to doing that.

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

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

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

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

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

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

32 Harming as a Means

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

CHAPTER 10 RESPECT AND VALUE

33 Respect for Persons

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

34 Two Kinds of Value

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

35 Kantian Dignity

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

Some of Kant's remarks suggest that non-moral rationality has supreme value. But Kant's main claims do not commit him to this implausible view. Kant fails to distinguish between being supremely good and having a kind of moral status that is compatible with being very bad. But we can add this distinction to Kant's view.

36 The Right and the Good

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

37 Promoting the Good

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

CHAPTER 11 FREE WILL AND DESERT

38 The Freedom that Morality Requires

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

39 Why We Cannot Deserve to Suffer

According to another of Kant's arguments, if our acts were merely events in time, we could never be responsible for these acts in some way that could make us deserve to suffer. Since we *can* be responsible for our acts in this desert-involving way, our acts are not merely such events. Though this argument is valid, it is not sound. We ought to accept Kant's claim that, if our acts were merely such events, we could not deserve to suffer. But since we ought to reject this argument's conclusion, we ought to reject Kant's other premise. Our acts *are* merely events in time. So we cannot deserve to suffer.

PART THREE THEORIES

CHAPTER 12 UNIVERSAL LAWS

40 The Impossibility Formula

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

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

good or morally required acts.

41 The Law of Nature and Moral Belief Formulas

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

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

According to

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

According to

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

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

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

42 The Agent's Maxim

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

warmer clothing, and saves some drowning child in the hope of getting some reward. These implications are clearly false. When this Egoist acts in these ways, his acts do not have what Kant calls *moral worth*, but they are not wrong.

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

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

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

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

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

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

The Moral Belief Formula could become:

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

These formulas will need some further revisions.

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

CHAPTER 13 WHAT IF EVERYONE DID THAT?

43 Each-We Dilemmas

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

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

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

44 The Threshold Objection

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

45 The Ideal World Objections

Kant's Law of Nature Formula, it is often claimed, requires us to act as if we were living in an ideal world, even when in the real world such acts would have predictably disastrous effects and be clearly wrong. We are required, for example, never to use violence even in self-defence, and required to act in various ways that

mistakenly ignore what other people will in fact do. This *Ideal World Objection* can be answered. Kant's formula does not require such acts.

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

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

CHAPTER 14 IMPARTIALITY

46 The Golden Rule

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

47 The Rarity and High Stakes Objections

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

48 The Non-Reversibility Objection

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

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

49 A Kantian Solution

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

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

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

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

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

CHAPTER 15 CONTRACTUALISM

50 The Rational Agreement Formula

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

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

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

51 Rawlsian Contractualism

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

> *Rawls's Formula*: Everyone ought to follow the principles that it would be rational in self-interested terms for everyone to choose, if everyone had to make this choice without knowing any particular facts about themselves or their circumstances.

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

52 Kantian Contractualism

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

53 Scanlonian Contractualism

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

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

54 The Deontic Beliefs Restriction

When we apply any Contractualist formula, Contractualists must claim, we cannot appeal to our intuitive beliefs about which acts are wrong. If we appealed to such *deontic* beliefs, these formulas would achieve nothing. Some Contractualists claim that we should never appeal to such intuitive deontic beliefs, which involve mere prejudice, or cultural conditioning. We should reject this claim. When we are

trying to decide which acts are wrong, we must appeal to these intuitive beliefs. Contractualists should claim instead that, though we cannot appeal to such beliefs *while* we are working out what their formula implies, we *can* appeal to these beliefs when we later try to decide whether, given these implications, we ought to accept this formula.

CHAPTER 16 CONSEQUENTIALISM

55 Consequentialist Theories

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

56 Consequentialist Maxims

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

57 to 62 The Kantian Argument

According to one version of

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

Such principles we can call *optimific*.

Kantians could argue:

> Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.

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

Therefore

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

Therefore

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

Therefore

> Everyone ought to follow these principles.

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

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

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

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

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

CHAPTER 17 CONCLUSIONS

63 Kantian Consequentialism

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

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

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

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

64 Climbing the Mountain

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

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

If we accept this theory, we should admit that acts can have other properties that make them wrong. The Triple Theory should claim to describe a single complex higher-level property under which all other wrong-making properties can be subsumed. If this theory succeeds, it would describe what these other properties have in common.

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

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

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

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

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

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

VOLUME TWO

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 view, harming people as a means is more often 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 statement 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. Barbara 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. Nor would it always be in this sense wrong to act on the maxim 'Never lie'. So Kant's formula does need to be revised.

73 A New Kantian Formula

Kant's Formula might be claimed to tell us when acts are in 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, 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 one of many people's lives from 30 years to 70 rather than lengthening all these 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, the many 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 a version of an earlier claim. 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 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* 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.

81 The Independence of Scanlon's Theory

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. 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 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 believe, to accept some non-Platonic form of Non-Naturalist Cognitivism. I shall argue that we ought to reject both Non-Cognitivism and two forms of Naturalism. These views, I believe, are close to Nihilism. Normativity is either an illusion, or involves irreducibly normative truths.

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 Naturalist*s, 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 and attitudinal conceptions, normativity involves actual or possible

motivation, or certain kinds of attitude. 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 desires, or that after informed and procedurally rational deliberation this person would be motivated 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 which cannot be helpfully explained in other terms. 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

Some Naturalists claim that, if normative and naturalistic concepts necessarily apply to all and only the same acts, these concepts must refer to the same property. That is not so.

Certain irreducibly normative concepts might refer to natural properties. But this does not show, as many Naturalists assume, that irreducibly 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 Reductive Naturalism

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.

89 Rules, Reasons, Concepts and Substantive Truths

If we use 'normative' in the rule-involving sense, we can claim that certain facts are both normative and natural. We can give Naturalistic accounts, for example, of what it is for acts to be illegal, dishonourable, or bad etiquette, or for the uses of words to be incorrect. If we use 'normative' in the better, reason-implying sense, we cannot give such accounts. There are no valid arguments with wholly naturalistic premises and normative conclusions. And like truths about what exists, no substantive normative truths could follow from our concepts or the meanings of our words.

90 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 different reasons, by those *Metaphysical* Naturalists who are Nihilists or Non-Cognitivists.

CHAPTER 26 THE TRIVIALITY OBJECTION

91 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.

92 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.

93 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.

These Naturalists claim that, when some act would have certain natural properties, that 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. What makes such claims seem informative also ensures that they could not be true.

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 fail. For 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 different properties. No such claim could refer only to natural properties.

94 Naturalism about Reasons

Similar objections apply to Non-Analytical Naturalism about reasons.

95 Soft Naturalism

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

96 Hard Naturalism

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

CHAPTER 27 NON-COGNITIVISM AND QUASI-REALISM

97 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 Non-Cognitivist *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 a similar 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.

98 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.

99 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 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. That 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 28 NORMATIVITY AND TRUTH

100 Expressivism

Gibbard's Expressivist account of the concept *rational* does not achieve Gibbard's aims.

101 Hare on What Matters

In his account of what matters, Hare denies that anything could matter.

102 Normative Questions

According to *the Normativity Argument for Non-Cognitivism*, normative truths would not really be normative, since truths cannot answer normative questions. That is not so. Only truths could answer such questions.

If there were no such truths, we would have no reason to try to decide how to live. These decisions would be arbitrary, since there would not be any better or worse ways to live. We would not be the animals that can understand and respond to reasons. In a world without reasons, we would act only on our instincts and desires, living as other animals live. The Universe could not contain rational beings.

CHAPTER 29 NON-NATURALIST METAPHYSICS AND EPISTEMOLOGY

103 Metaphysical Objections

104 Epistemological Objections

CHAPTER 30 IRREDUCIBLY NORMATIVE TRUTHS

PART SEVEN APPENDICES

APPENDIX A 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 B STATE-GIVEN REASONS

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

APPENDIX C RATIONAL IRRATIONALITY AND GAUTHIER'S THEORY

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

APPENDIX D DEONTIC REASONS

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

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

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

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

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

APPENDIX E 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 F 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 G 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 H 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.

PART ONE REASONS

CHAPTER 1 NORMATIVE CONCEPTS

(My endnotes are best ignored, unless they are attached to claims that seem false, or whose meaning is unclear. Several notes need to be added, some acknowledging my debts to others.)

1 Sufficient and Decisive Reasons

We are the animals that can both understand and respond to reasons. These abilities have given us great knowledge, and power to control the future of life on Earth. Though there may be life elsewhere, there may be no other animals like us. We may be the only rational beings in the Universe. [47]

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

It is hard to explain the *concept* of a reason, or what the phrase 'a reason' means. Facts give us reasons, we might say, when they count in favour of our having some attitude, or our acting in some way. But 'counts in favour of' means roughly 'gives a reason for'. Like some other fundamental concepts, such as those involved in our thoughts about time, consciousness, and possibility, the concept of a reason is indefinable in the sense that it cannot be helpfully explained merely by using words. [48] We have to explain such concepts in a different way, by getting people to think thoughts that use these concepts. One example is the thought that we always have a reason to want to avoid being in agony.

We can have reasons, I shall say, of which we are unaware. Suppose that I ask my doctor, 'Since I'm allergic to apples, do I have any reason not to eat any other kind of food?' If my doctor knows that walnuts would kill me, her answer should be Yes. This fact gives me a reason.

Rather than saying that certain facts *give* us reasons, some people say that these facts *are* reasons for us. And some people say that, to *have* some reason, we must be

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

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

If our reasons to act in some way are stronger than our reasons to act in any of the other possible ways, these reasons are *decisive*, and acting in this way is what we have *most reason* to do. [49] If such reasons are much stronger than any set of conflicting reasons, we can call them *strongly* decisive. Though most kinds of reason are decisive only in certain cases, there may be some kinds of reason that are always decisive. On some views, for example, we always have decisive reasons not to act wrongly.

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

There is often nothing that we have decisive reasons to do, or *most* reason to do, because we have *sufficient* reasons, or *enough* reason, to act in any of two or more ways. Our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to act in any of the other possible ways. We might have sufficient reasons, for example, to eat either a peach or a plum or a pear, to choose either law or medicine as a career, or to give part of our income either to *Oxfam* or to some other similar aid agency, such as *Medecins Sans Frontieres*. When neither of two conflicting reasons is stronger, that is seldom because these reasons are precisely equally strong. Though there are truths about the relative strength of different reasons, these truths are often very imprecise.

Reasons can be related in more complicated ways. Some facts give us reasons, for example, to ignore some other reasons. If I am judging who deserves some prize, that would give me a reason to ignore the fact that one of the contestants is my best

friend. And some facts give us reasons, not in all cases, but only when combined with certain other facts. I shall mainly be discussing simpler reasons.

When we have decisive reasons, or most reason, to act in some way, this act is what we *should* or *ought* to do in what we can call the *decisive-reason-implying* senses. [50] Even if we never use the phrases 'decisive reason' or 'most reason', most of us often use 'should' and 'ought' in these reason-implying senses. There is a similar sense of 'must'. These words imply reasons of different strengths. I might say that you *should* see some film, that you *ought* to give up smoking, and that you *mustn't* touch some live electric cable. Though the word 'should' is used more often, and the word 'must' has more force, I shall mostly use the less ambiguous word 'ought'.

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

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

> *rational* if we have beliefs about the relevant facts whose truth would give us sufficient reasons to act in this way, [51]
>
> what we *ought rationally* to do, or *rationally required*, if these reasons would be decisive,
>
> *less than fully rational* if we have beliefs whose truth would give us clear and decisive reasons *not* to act in this way,

and

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

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

Some people would say that you do have a reason to run away, which is provided by your false belief that this act would save your life. But if we say that false beliefs can give people reasons, we would need to add that these reasons do not have *normative force*, in the sense that they do not count in favour of some act. And we would have to ignore such reasons when we are trying to decide what someone has most reason to do. It is better to say that *all* reasons have normative force, and that false beliefs can give people what *merely appear* to be reasons, or what I shall call, more briefly, *apparent* reasons. In the case of the angry snake, given your false belief that running away would save your life, you have an apparent reason to run away. When we have beliefs whose truth would give us a reason to act in some way, we have either a real or an apparent reason to act in this way. In either case, if this reason would be decisive, we ought rationally to act in this way.

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

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

and

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

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

It is worth explaining why, though it is facts that give us reasons, the rationality of our acts depends instead on our beliefs. When we are trying to decide what we or others ought to do, what matters are the reason-giving facts. In the case of the

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

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

To be fully rational, we may also need to meet certain other *rational requirements*, by avoiding certain kinds of inconsistency and other mismatch between our intentions, beliefs, and other mental states. We may be rationally required, for example, not to have contradictory intentions, and to intend to do what we believe that we ought to do. Though these requirements raise several interesting questions, I shall say little about them. Questions about reasons are, I believe, more fundamental. And while it often matters greatly whether we are wanting what we have reasons to want, and doing what we have reasons to do, it seldom matters, or matters much, whether we are being inconsistent and thereby failing to meet some rational requirement. [52] Some people claim that, to be rational, we don't need to respond to reasons or apparent reasons, since it is enough to meet these rational requirements. I shall later give some arguments against this view.

There are some other, similar questions that I shall mention briefly and then set aside. When we are deciding what to do, and we don't know all of the relevant facts, we must base our decision on what we believe, and on the available evidence. In such cases, we can ask what we *should* or *ought* to do in what we can call the *evidence-relative* senses. It may seem that, in such cases, we ought to try to do what we have most reason to do. But such attempts may be too risky, or too unlikely to succeed. We often ought to act in ways that are more likely to achieve less ambitious aims. If many people's lives are in danger, for example, we ought to do what would certainly save most of these people, rather than doing what has only a small chance of being the act that would save them all.

It is of great practical importance what we ought to do in cases that involve risk or uncertainty. These questions have been well discussed by many philosophers, decision theorists, and others. Certain questions about reasons, though more fundamental, have been less well discussed. These are also questions about which

different people more deeply disagree. Since I shall be mainly discussing these questions, I shall mostly consider cases in which we know all of the relevant, reason-giving facts.

These claims have been about about *normative* reasons. When we have such a reason or apparent reason, and we act *for this reason*, this becomes our *motivating* reason. If I avoid walnuts, for example, my motivating reason might be that, as my doctor has told me, eating them would kill me. This distinction is clearest when we have only a motivating reason for acting in some way. If you ran away from the angry snake, your motivating reason would be provided by your false belief that this act would save your life. [53] But, as I have said, you have no normative reason to run away. You merely think you do. In an example of a different kind, we might claim: 'His reason was to get revenge, but that was no reason to do what he did'. Since I shall not be discussing why people act as they do, I shall say little about motivating reasons.

As well as asking what we ought to do in the decisive-reason-implying sense, and what we ought rationally to do, we sometimes ask what we ought to do in one of several *moral* senses. Most of these senses differ in at least two ways from the decisive-reason-implying sense. First, we often have decisive reasons that are not moral reasons. If I need to catch some train, for example, I may have a decisive reason to leave some meeting now. If I hate commuting, I may have most reason to live close to where I work. These may not be things that I ought morally to do. Second, when we believe that we ought morally to act in some way, we may not believe that we have decisive reasons to act in this way. On some views, we might have *no* reason to do what we ought morally to do. In these chapters I shall first discuss reasons, turning only later to morality.

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

2 Reason-involving Goodness

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

> *good,* in what we can call the *reason-implying* sense, we mean that there are certain kinds of fact about this thing's nature, or properties, that would in some situations give us or others strong reasons to respond to this thing in

> some positive way, such as wanting, choosing, using, producing, or preserving this thing.

Some book may be good, for example, by being enjoyable, or inspiring, or containing useful information. Some medicine may be the best by being the safest and the most effective. These facts may give us or others reasons to read this book, or to take this medicine. There are similar senses of 'better', 'bad', 'worse', and 'worst'. [54]

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

> Ice formed on the butler's upper slopes, [55]

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

> A MAN A PLAN A CANAL: PANAMA,

I would not intend these uses of 'good', 'bad', and 'best' to be reason-implying. Moles could not have reasons to wear spectacles, nor do we have reasons to be amused by the ice on the butler's upper slopes. And many uses of 'good' mean only that something meets certain standards. But the most important uses of 'good' and 'bad' are, I believe, reason-implying.

When something is in this sense good, T. M. Scanlon claims, this thing's goodness could not give us reasons. Such goodness is the property of having *other* properties that might give us certain reasons, and the second-order fact that we had these reasons would not itself give us any reason not to act in this way. [56]

This view needs, I think, one small revision. If some medicine or book is the best, these facts could be truly claimed to give us reasons to take this medicine, or to read this book. But these would not be *further, independent* reasons. These reasons would be *derivative*, since their normative force would derive entirely from the facts that made this medicine or book the best. That is why it would be odd to claim that we had *three* reasons to take some medicine: reasons that are given by the facts that this medicine is the safest, the most effective, *and* the best. Since such derivative reasons have no independent normative force, it would be misleading to mention them in such a claim. [57]

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

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

When people say that something would be good for us, or in our interests, these people often mean that this thing would have good effects on our health, or our bank balance. In my intended wider sense, something is *intrinsically* or in itself good for us if it is one of the features of our lives in which our well-being consists, because these are the features that make our lives worth living. Something is *instrumentally* good for us if it has effects that are intrinsically good for us. On *hedonistic* theories, our well-being consists, roughly, in pleasure and happiness, and avoiding pain and suffering. On theories that appeal to *substantive goods*, our well-being may also partly consist in some other states or activities, such as loving and being loved, being morally good and acting well, and various other kinds of achievement. On *desire-based* theories, our well-being consists in the fulfilment of some of our desires, such as our informed desires about our own life. On any plausible theory, hedonism covers at least a large part of the truth, so my examples will often involve hedonic well-being.

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

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

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

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

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

I use the phrase 'point of view' in something close to its literal sense, not the looser sense in which we talk of the reasons that we might have from a financial, aesthetic, or other such point of view. We have an impartial point of view when we are considering possible events that would affect or involve people who are all

strangers to us. When our actual point of view is not impartial, we can think about possible events from an imagined impartial point of view. We can do that by imagining possible events that are relevantly similar, except that the people involved are all strangers to us.

We have impartial reasons, I believe, to care equally about everyone's well-being. That is a substantive belief, not something that is implied by my definition of an impartial point of view. It has been widely believed that we have reasons to care more about the well-being of certain kinds of people, such as those who are morally good, or those who have the greatest abilities. We can also note that, when our *point of view* is impartial, that does not ensure that *we* are impartial. We might care more about the well-being of certain strangers, such as those who are more similar to us, or those whose faces we like. But we would have no *reasons*, I believe, to care more about the well-being of these people. [58]

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

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

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

When we call one of two events

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

It would be in this sense better, I believe, if some plague or earthquake killed fewer people, or if any person or other animal ceased to be in pain. This kind of goodness is *impersonal* in the sense that, when we call some event in this sense good, we don't mean that this event would be *good for* some person or group of people. But many events are impersonally good because they are good for one or more people. The benefits to these people are what make these events impersonally good. And since everyone has reasons to want such events to occur, such impersonal goodness involves *omnipersonal* reasons.

If some possible event would be in these senses good for someone, or impersonally good, this fact could be truly claimed to give us a reason to want this event to occur. But as before, this reason would be derivative, since this reason's force would derive from the facts that would make this event good for this person, or impersonally

good. When we use 'good for' and 'good' in these senses, these are merely briefer ways of implying that there are such other, reason-giving facts. Unlike the concept of *a reason*, and the decisive-reason-implying concept *should* or *ought*, these versions of the concept *good* are not fundamental.

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

CHAPTER 2 OBJECTIVE THEORIES

3 Two Kinds of Theory

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

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

Our desires have *objects*, which are *what* we want. These objects are all *events* in the sense that includes acts and states of affairs. We can be correctly said to want things of other kinds. I might want an apartment in Venice, a glass of water, and a piano teacher. Some fugitive may be wanted by the police. But what we really want is to own, live in, drink, be taught by, find, use, or have some other relation to some thing or person. Rather than saying that we want some event *to occur*, I shall say, for short, that we want this event.

Our desires are *teleological* or *telic* when we want some event as an *end*, or for its own sake. Our desires are *instrumental* when we want some event as a *means*, because this event would or might cause some other event that we want. [59] We want some acts or other events both as an end and as a means to some other end. Two such events would be a thrilling seach for some important truth; and, when we want to have a child, making love. When we decide to try to fulfil some telic desire, we thereby make this desire's fulfilment one of our *aims*.

We often have long chains of instrumental desires, but such chains all begin with, or are grounded on, some telic desire. I might want medical treatment, for example, not for its own sake, but only to restore my health, and I might want health only so that I can finish writing some great novel, and I might want to finish this novel only

to achieve posthumous fame. This desire might also be purely instrumental, since I might want to achieve such fame only to refute my critics, or to increase the income of my heirs. But if I want posthumous fame for its own sake, this telic desire would begin this particular chain.

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

As well as wanting such other things, some people do not even want pleasure as an end. Suppose that we know some relentlessly ambitious politician, whom we find gambling in a casino, sipping champagne. When we ask this man what he is doing, he replies 'Enjoying myself'. Given our knowledge of this man's character, this reply is baffling. This man never does anything merely for enjoyment. He then explains that his doctor warned that, unless he allows himself such pleasures, his health will worsen, thereby hindering his pursuit of power. Our bafflement disappears. This man wants these pleasures, not for their own sake, but only as a means.

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

Object-given reasons are provided by the facts that make certain outcomes worth producing or preventing, or make certain things worth doing for their own sake. In most cases, these reason-giving facts also make these outcomes or acts good or bad for particular people, or impersonally good or bad. So we can also call these objective reasons and theories *value-based*. [60]

According to another group of theories, our reasons for acting are all provided by, or

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

These two kinds of theory are very different. According to Objectivists, though many reasons for acting can be claimed to be given by the fact that some act would achieve one of our aims, these reasons *derive their force* from the facts that give us reasons to *have* these aims. These are the facts that make these aims relevantly good, or worth achieving. According to Subjectivists, we have no such reasons to have our aims. Some Subjectivists even claim that it is *we* who, with our desires or choices, make things good. While defending such a view, Christine Korsgaard writes:

> most things are good because of the interest human beings have in them. . . Objectivism reverses this relation. . . Instead of saying that what we are interested in is therefore good, the objectivist says that the goodness is in the object, and we ought therefore to be interested in it. [61]

It is of great importance whether our desires can, in this way, make their objects good.

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

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

our money now, though we know that some of this money would later bring us greater benefits. By fulfilling such desires and preferences, many of us make our lives go worse. In these and many other ways, subjective and objective theories often disagree about what we have *most* reason to do.

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

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

4 Responding to Reasons

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

Similar claims apply to our *epistemic* reasons to have particular beliefs. These reasons are provided by facts that are related to the *truth* of some belief, by being evidence for its truth, or by logically implying this belief, or in some other way. If we see dark grey clouds, for example, that gives us some reason to believe that it will soon rain. If we know that gold weighs more than lead, which weighs more than iron, these facts give us a decisive reason to believe that gold weighs more than iron. When we are aware of facts that give us decisive reasons to have some belief, we can respond to these reasons by coming to have and continuing to have this belief. But our responses to such reasons are seldom voluntary. We could seldom choose *not* to believe what we have such decisive reasons to believe. If this despot threatens to kill me unless, one minute from now, I no longer believe that 2 + 2 = 4, I could not

choose to lose this belief.

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

Though we can seldom choose how we respond to our reasons to have particular beliefs and desires, our responses to these reasons are not things that merely happen to us, like an automatic knee-jerk, or our slipping on a banana skin. Our being rational consists in part in our responding to such reasons or apparent reasons in these non-voluntary ways. We can be asked *why* we believe something, or want something, and we can often give our reasons. [62]

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

The answer is likely to be No. Suppose first that, as well as failing to believe that your life is in danger, you also fail to believe that the smoke and flames give you any reasons to have this belief. You could not then correct your mistake, since you would not believe that you had made any mistake. You could not choose to believe, for these epistemic reasons, that your life is in danger, since you would not believe that you had these reasons.

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

There might be exceptions. Suppose next that, though you believe that the smoke and flames give you decisive reasons to believe that your life is in danger, you don't

yet have this second belief. We can perhaps imagine that you would then be able to take a further mental step, by choosing to make yourself have this belief for these reasons. But your response to these epistemic reasons would still be only partly voluntary. When you saw that smoke and flames were rising up the stairs of your hotel, you did not choose to believe that these facts gave you decisive reasons to believe that your life is in danger.

There are other reasons why our responses to most epistemic reasons could not be voluntary. For us to have knowledge of the world around us, our beliefs must be reliably caused by our visual and other perceptual experiences, or by our awareness of other facts that give us epistemic reasons to have these beliefs. Such causation could not be reliable if we could freely choose all of our beliefs. And to have knowledge of necessary truths, such as logical or mathematical truths, we must also respond to some epistemic reasons in rational but non-voluntary ways, by recognizing or realizing what follows from what, and what must be true. [63]

Similar claims apply to our desires and preferences. We can seldom choose what it is that we want or prefer, because we cannot choose either what we have reasons to want, or how strong these reasons are. What we can choose is only which of our desires we try to fulfil. Our responses to these reasons might become somewhat more voluntary than they are now. That would be, in some ways, better, since we could then more easily transform our desires, attitudes, and emotions, by making ourselves become the kind of person that we have reasons to want to be. We might be able to ensure, for example, that we shall never lose our youthful ideals. But such abilities would also be dangerous, like our recently discovered mechanical ways of moving our bodies at great speed. If we changed ourselves for the worse, our new, deliberately chosen desires might lead us not to undo such mistakes.

5 State-given Reasons

Our reasons to have some desire are provided, I have claimed, by facts about this desire's *object*, or the event that we want. Such reasons I am calling *object-given*. Many people assume that we can also have *state-given* reasons to have some desire. Such reasons would be provided by certain facts, not about some desire's object, but about our state of having this desire. We would have such reasons when our having some desire would be in some way good, either as an end or as a means. [64]

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

	telic and intrinsic	instrumental
object-	The event that we want	This event would

given	would be in itself good, or worth achieving	have good effects
state-given	Our wanting this event would be in itself good	Our wanting this event would have good effects

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

Similar claims apply to our reasons to have beliefs. Since our epistemic reasons are related to the truth of *what* we believe, these reasons can also be called *object-given*. Many people assume that we can also have *state-given* reasons to have certain beliefs. Such reasons would be provided by facts that would make our *having* some belief in some way good. It is often claimed, for example, that we have such reasons to believe that God exists and that we shall have a life after death. These reasons would not be epistemic, or truth-related, but *goodness-related*, or *value-based*. Such alleged reasons to have beliefs are sometimes called *practical* or *pragmatic*.

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

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

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

6 Hedonic Reasons

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

Telic reasons are *intrinsic* when they are provided by facts about some possible event's intrinsic properties or features, or what this event would *in itself* involve. We might have such reasons, for example, to want to make someone feel less lonely, or to see the sublime view from the summit of some mountain, or to understand how life or the Universe began. We might also have *extrinsic* telic reasons to want some possible event, which would be provided by facts about this event's relation to other events. But such reasons do not need to be separately considered, since such events would be *extrinsically* good by making some longer sequence of events, of which they were one part, *intrinsically* better. [65]

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

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

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

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

Similar remarks apply, I believe, to many aesthetic experiences. It is sometimes claimed that we have reasons to enjoy, or be thrilled or in other ways moved by, great artistic works. In many cases, I believe, this claim is false. We can have reasons to *want* to enjoy, or be thrilled or moved by, these artistic works. But these are not reasons to *enjoy*, or to *be* thrilled or moved by, these works. We do have reasons to admire some novels, plays or poems, given the importance of some of the ideas that they express. But poetry is what gets lost in the translation, even if this translation expresses the same ideas. And we never have reasons to enjoy, or be moved by, great music. If we ask what makes some musical passage so marvellous, the answer might be 'Three modulations to distant keys'. This answer describes a *cause* of our response to this music, not a *reason*. Modulations to distant keys are like the herbs, spices, or other ingredients that can make food delicious. When someone neither enjoys nor is moved by some great musical work, this person is not in any way less than fully rational, by failing to respond to certain reasons. In comparing music with food in this way, I am not belittling music, ranking it below novels, plays, or poems. Music is at least as great as the other arts. Without music, Nietszche plausibly (though falsely) said, life would be an error. But music is also the lost battlefield and graveyard of most general aesthetic theories.

Since these claims are controversial, we can return to those non-aesthetic sensations that people like or dislike. Though these sensations are not in themselves good or bad, they are parts of complex mental states that *are* good or bad. When we are in pain, what is bad is not our sensation but our conscious state of having a sensation that we dislike. If we didn't dislike this sensation, our conscious state would not be bad. What these sensations feel like may in part depend on whether we dislike

them. Such sensations might be claimed to be in themselves bad when their quality is affected in certain ways by our disliking them. On this view, it would still be true that, if we didn't dislike these sensations, neither they nor our conscious state would be bad, nor would we be failing to respond to some reason. [66]

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

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

If we call these states desires, we should remember that, given the differences between these states and our other desires, true claims about these states may not apply to our other desires. There are some other important and often ignored differences between these states and our meta-hedonic desires.

First, many people believe that our desires can *create* or *confer* value or disvalue. Korsgaard, for example, writes that something can be 'objectively good as an end *because* it is desired for its own sake.' [67] On this view, we create value by valuing things, and things matter by mattering to us. This view may seem to be supported by the examples of pleasure and pain. Our hedonic likings and dislikings *do*, as I have said, make our conscious states good or bad. If we fail to distinguish between these likings or dislikings and our meta-hedonic desires, we may believe that these desires make their objects good or bad.

Korsgaard's remarks provide one example. To illustrate her claim that something can be good 'because it is desired for its own sake', Korsgaard writes: 'chocolate gets its value from the way it affects us. We *confer* value on it by liking it'. [68] Such examples do not, I believe, show that our *desires* can create or confer value, or disvalue, by making what we want to have, or to avoid, good or bad. Our future pleasures or pains are not made to be good or bad by our present desires to have these pleasures, and to avoid these pains. And when we are in great pain, by having some sensation that we intensely dislike, what makes our conscious state bad

is our intense dislike, not our present desire not to be in this conscious state. Since our meta-hedonic desires do not make their objects good or bad, the examples of pleasure and pain do not support the view that our other desires have such value-creating power. Though it is good to have sensations that we *like*, nothing is good merely because we *want* this thing.

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

7 Irrational Preferences

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

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

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

to prefer it. So this man's preferences are strongly contrary to reason, and irrational.

Consider next some man who has a *bias towards the next year*. This imagined man cares equally about his future well-being throughout the next year, but he cares only half as much about his well-being in later years. Rather than having five hours of pain eleven months from now, he would prefer to have nine hours of pain twelve months from now. Such preferences are also irrational. If we would have some future pain just over rather than just under a year from now, that is no reason to care now about this pain only half as much.

No one has these attitudes to time. But many of us have what I have called the *bias towards the near*. Unlike these two imagined attitudes, this bias does not draw wholly arbitrary distinctions. But suppose that, because you have this bias, you want some ordeal to be briefly postponed, though you know that this postponement would make your ordeal much worse. Rather than having one minute of slight pain later today, you prefer to have one hour of agony tomorrow. This preference would also be, though more weakly, irrational. Many people often act on less extreme preferences of this kind, thereby making their lives go worse. [69]

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

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

CHAPTER 3 SUBJECTIVE THEORIES

8 Subjectivism about Reasons

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

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

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

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

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

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

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

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

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

Subjectivists can defend (B), however, in a different way. Suppose that I want to eat the two remaining apples that are on some tree. I also want to climb a ladder so

that I can reach the higher apple. Suppose next that this tree's owner allows me to eat only one of these apples, and lets me choose which apple I shall eat. If instrumental desires gave us reasons, I would have more reason to choose the higher apple. If I chose the lower apple, I would then fulfil only my desire to eat this apple. If I chose the higher apple, I would fulfil not only my desire to eat this other apple, but also my instrumental desire to climb this ladder so that I can reach this apple. This reasoning is obviously mistaken. Since I want to climb this ladder, not for its own sake, but only as a means of reaching this apple, I have no further, independent reason to fulfil this desire. My reason to climb this ladder derives entirely from, and adds nothing to, my reason to fulfil my desire to eat this higher apple. [70]

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

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

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

In some cases, however, our *telic* desires or aims depend on false beliefs. I might want to hurt you, for example, because I falsely believe that you deserve to suffer, or because I want to avenge some injury that I falsely believe you have done me. Subjectivists ought to deny that this desire gives me a reason. When they consider such cases, many Subjectivists claim that reasons are provided only by telic desires or aims that are *error-free*, in the sense that they do not depend on false beliefs. [71] According to what we can call

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

this fact makes it true that

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

According to these Objectivists, when it is true that

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

this fact makes it true that

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

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

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

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

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

Subjectivism about Reasons: Some possible act is

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

just when, and because,

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

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

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

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

(2) stopping smoking would lengthen my life.

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

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

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

In assessing subjective theories, it will be enough to consider *what* these theories imply that we have reasons to do, ignoring these disagreements about which are the facts that give us these reasons. When I say that, on these theories, reasons are provided by certain facts about our desires, aims, or choices, I shall mean that these are among the facts that make it true that we have these reasons.

Subjectivism about Reasons is now very widely accepted. Many writers take it for granted that we have subject-given reasons. Korsgaard for example writes that, if some act 'is a means to getting what you want. . . no one doubts that this is a reason'. [72] Williams writes: 'Desiring to do something is of course a reason for doing it.' [73] In many books and articles, Subjectivism is not even claimed to be the best of several views, but is presented as if it were the only possible view. So it is of great importance whether this view is true.

9 Why People Accept Subjective Theories

We ought, I believe, to reject all subjective theories, and accept some objective theory. Our practical reasons are all object-given and value-based. [74]

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

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

have value-based object-given reasons to try to fulfil such desires.

Second, we often have such desires because we believe that we have such reasons. We are often motivated by the belief that some act or outcome would be good or best, in the reason-implying sense. When our desires depend on our beliefs that we have such reasons, we may fail to distinguish between these desires and these beliefs. [75]

Third, some people accept desire-based theories about well-being. According to some of these theories, the fulfilment of some of our present desires would be in itself good for us. If that were true, we would have value-based reasons to fulfil these desires.

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

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

Sixth, when we could fulfil *other people's* desires, or help these people to achieve their aims, these facts may give us *non*-derivative reasons to act in these ways. When other people have some desire or aim that they have no reason to have, these people may have no reason to try to fulfil this desire or achieve this aim. But *we* may have such reasons. In helping other people to fulfil or achieve their desires or aims, we respect these people's autonomy, and avoid paternalism. Other people's desires, aims, or choices are often, in this respect, like votes, which should be given just as much weight even when the voters have no reason to vote as they do. [76] Many people accept desire-based or choice-based theories because they are democrats, liberals, or libertarians, who believe that we should not tell other people what they ought to want, or choose, or do. Nozick, for example, claims that a substantive value-based theory 'opens to the door to despotic requirements, externally imposed'. [77]

Seventh, when we have some aim, and we believe that some possible act would be

the only or the best way to achieve this aim, it may be true that we ought rationally to act in this way. Some people assume that, in such cases, we must have a reason to do what we ought rationally to do. But that is not so. When we believe falsely that some act would achieve our aim, we may have no reason to act in this way. Though you ought rationally to run away from the angry snake, you have no reason to run away.

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

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

Tenth, we have many reasons for acting that we wouldn't have if we didn't have certain desires. But these reasons are provided, not by the facts that our acts would fulfil these desires, but by certain other facts that causally depend on our having these desires. When we have some desire, for example, that may cause it to be true that this desire's fulfilment would be pleasant. In many cases, this fact would merely give us a further reason to fulfil this desire, since what we want would be in itself worth achieving. But such cases take their clearest form when we have no such reason to have some desire. When we play some kinds of game, for example, such as games without rewards whose outcomes depend on luck, we have no reason to want to win. But if we do want to win, that may make it true that we would enjoy winning, and this second fact would then give us a reason to try to fulfil this desire.

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

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

Suppose next that we must choose between two or more good possible aims, none of which would be more worth achieving than any of the others. Some examples are choices between different possible careers, or research projects, or between doing voluntary work for different aid agencies, or political campaigns. If there is one of these possible aims that we most strongly want to achieve, this fact may give us reasons to adopt this aim. But these reasons would again be given, not by the fact that our strongest desire is to achieve this aim, but by certain other facts that would depend on our having this desire. If one of these aims seems most appealing, for example, that may give us reasons to believe that we would find this aim's achievement most rewarding. The thought of this aim's achievement may give us pleasure in advance. And our strongly wanting to achieve this aim may make it easier for us to make the efforts and sacrifices that would be needed to achieve this aim. We may need such desires in our darkest hours, when we are losing energy or hope. As before, it would be these other facts, and not our desire itself, that would give us reasons to adopt and try to achieve this aim.

Similar claims apply to our decisions and aims. When we have decided to try to fulfil some desire, thereby making its fulfilment one of our aims, this decision may give us a further reason to try to fulfil this desire, thereby achieving this aim. But this reason would not be provided merely by the fact that we have made this

decision and adopted this aim. This reason would be provided by the fact that, if we do not act on this decision, we shall be less likely to achieve this aim, and more likely to waste our time. In some cases, however, neither is true, since we have nothing better to do than to reconsider some decision. If we have woken up in the middle of the night, for example, reconsidering our decision to adopt some aim may be less boring than simply waiting to drift back to sleep. In such cases, the fact that we have adopted some aim gives us no reason to keep and to try to achieve this aim, since this fact gives us no reason not to change our mind, and adopt some other aim instead. [78]

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

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

10 Analytical Subjectivism

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

substantive when this claim both

(a) states that something has some normative property,

and

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

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

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

> There is no science so filled with tautologies as ethics. [79]

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

(1) happiness is happiness,

and that

(2) acts that produce happiness produce happiness.

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

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

(3) felicity is happiness.

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

(4) acts that produce happiness are felicific.

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

(2) acts that produce happiness produce happiness,

and

(5) acts that are felicific are felicific.

As before, these are not substantive claims. Everyone who understands these

claims would accept them, because they are so obviously true. And everyone could consistently accept these claims whatever else they believe. (4) differs in these ways from the claim that

(6) acts that produce happiness are good.

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

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

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

they mean that

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

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

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

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

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

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

For Subjectivists about Reasons to make substantive claims, they must use the words 'reason', 'should', and 'ought' in the indefinable, normative senses that I discussed in Section 1. It is these substantive, non-analytical subjective theories that, in these chapters, I am discussing.

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

When making these claims, I shall use the word 'desire' in a wide sense, which covers any state of being motivated, or of wanting something to happen and being to some extent disposed to make it happen, if we can. My claims do not apply, however, to various complex states that involve desires. When we love someone, for example, we are motivated to act in certain ways. We care greatly about this person's well-being, and we want to do what would be best for him or her. Though our loving someone partly consists in our having such desires, we have strong reasons, I believe, to care about, and try to promote, the well-being of those we love. Such reasons are provided, not by the desires involved in loving someone, but by various other facts about our relations to those we love, such as facts about shared histories, or commitments, or reasons for gratitude, or by the facts that are involved in romantic or erotic love, or love for our parents, children, or other close relatives. [80] To illustrate this distinction, we can suppose that I meet several strangers, all of whom need my help. If I had a strong desire to help one of these strangers, perhaps because I like her face, that would at most give me only a weak reason to help this stranger rather than any of the others. Love, in its various forms, is very different from such a desire.

11 The Agony Argument

Subjective theories can have implausible implications. Suppose that, in

> *Case One*, I know that some future event would cause me to have some period of agony. Even after ideal deliberation, I have no desire to avoid this agony. Nor do I have any other desire or aim whose fulfilment would be prevented either by this agony, or by my having no desire to avoid this agony.

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

This case might be claimed to be impossible, because my state of mind would not be *agony* unless I had a strong desire *not* to be in this state. But this objection overlooks the difference between our attitudes to present and future agony. Though I know that, when I am later in agony, I shall have a strong desire not to be in this state, I

might have no desire now to avoid this future agony.

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

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

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

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

Some *Analytical* Subjectivists would accept this conclusion. If these people claimed that I would have no reason to avoid this agony, their claim would not be normative, but a concealed tautology, which merely repeats my description of this imagined

case. These people would mean only that, after ideal deliberation, I am not motivated to move my hand away from the approaching fire. We could all agree that, in this trivial and misleading sense, I would have no reason to act in this way.

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

We can now argue:

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

Therefore

> Subjectivism is false.

We can call this *the Agony Argument*.

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

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

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

Third, even if there were no such actual cases, normative theories ought to have acceptable implications in merely imagined cases, when it is clear enough what such cases would involve. Subjectivists make claims about which facts give us reasons.

These claims cannot be true in the actual world unless they would also have been true in possible worlds in which there were people who were like actual human beings, except that these people did not want to avoid all future agony, or their desires differed from ours in certain other ways. So we can fairly test subjective theories by considering such cases.

Subjectivists might reply that, even in such possible worlds, there would be some telic desires that everyone must have, because without these desires these people could not even be rational agents, who can act for reasons. To be such agents, Bernard Williams suggests, we must have 'a desire not to fail through error', and some 'modest amount of prudence'. [81] But such claims are irrelevant here. We could be agents who act for reasons without wanting to avoid all future agony.

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

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

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

> (A) We all have reasons to have those desires that would be had by anyone who was fully rational.

(B) Anyone who was fully rational would want to avoid all future agony.

Therefore

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

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

Subjectivists cannot, however, make such claims. On subjective theories, we have no such object-given reasons, not even reasons to want to avoid future agony. Deliberative Subjectivists appeal to what we would want after deliberation that was *merely procedurally* rational. On these theories, *if* we have certain telic desires or aims, we may be rationally required to want, and to do, what would best fulfil or achieve these desires or aims. But, except perhaps for the few desires without which we could not even be agents, there are no telic desires or aims that we are rationally required to have. We can be procedurally rational whatever else we care about, or want to achieve. [82] As one Subjectivist, John Rawls, writes:

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently. [83]

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

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

> (C) we are rationally required not to have desires or preferences that draw some arbitrary distinction. [84]

By appealing to this 'minimal principle', Smith writes, Subjectivists can explain the irrationality of many desires and preferences, such as the preferences of my imagined man who cares about what will happen to him except on any future Tuesday. [85] This man's preferences are irrational, Smith claims, because they draw an arbitrary distinction. It would be similarly arbitrary, Subjectivists might claim, not to want to

avoid all future agony.

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

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

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

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

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

Smith also claims that

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

To illustrate this requirement, Smith supposes that we want to help only some of the people whom we know to be in desperate need. Our desires would be more coherent, and would 'make more sense', Smith claims, if we wanted to help all of these people.[86] But this claim assumes that

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

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

Consider next Smith's claim that we can be rationally required to have a more unified set of desires. Mere unity is not a merit. Our desires would be more unified if we were monomaniacs, who cared about only one thing. But if you cared about truth, beauty, and the future of mankind, and I cared only about my stamp collection, your less unified set of desires would not be, as Smith's claim seems to imply, less rational

than mine. Smith might reply that my set of desires would be more impressively unified if I had several coherent desires. But if I also wanted to collect match-boxes, drawing pins, ticket stubs, and plastic cups, your less unified set of desires would still be more rational than mine. And this appeal to coherence would again assume that we have object-given reasons to have our desires. Subjectivists deny that we have such reasons.

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

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

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

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

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

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

We can argue

> The nature of agony does give us such a reason.
>
> Therefore
>
> Subjectivism is false.

These arguments are, I believe, decisive.

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

CHAPTER 4 FURTHER ARGUMENTS

12 The All or Nothing Argument

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

Suppose that, in

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

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

In response to this objection, Subjectivists might reply that *Case Two* cannot be coherently imagined. Some writers claim that, if we really believed that it would be *us* who would later be in agony, and we also understood what this agony would be like, it is inconceivable that we might want ourselves to be later in this state. [87] But this claim is false. We can want what we know will be bad for us. It makes sense to suppose that someone wants to have some future period of agony, for its own sake. Nor could Subjectivists claim that, if we had this desire, that would make it impossible for us to be rational agents, who act for reasons.

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

The opposite is true. This fact *strengthens* this objection. If we find it hard to imagine that anyone might have this desire, that is because we assume what objective theories claim. The nature of agony, we believe, gives everyone very

strong reasons to want *not* to be in this state. According to subjective theories, we have no such object-given reasons. If that were true, it would *not* be hard to imagine that someone might want, for its own sake, to have some future period of agony. We could at most claim that this desire would be unusual, like the bizarre sexual desires that some people have. This case is hard to imagine because the awfulness of agony gives everyone such clear and strong reasons *not* to have this desire. It is hard to believe that anyone could be so irrational. [88]

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

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

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

To assess this reply, we can suppose that, in

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

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

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

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

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

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

Suppose that, if my headache returns while I am playing chess this afternoon, my pain would distract me, and would deny me the victory that I want. Subjective theories then imply that I have a reason to want to avoid this headache as a means of helping me to win this game, thereby fulfilling my desire. But we can suppose that, in *Case Three*, I have no such instrumental reason to want to avoid my future

period of agony. Since this period would be fairly brief, my avoiding this agony would not have any other effects that I want, or help me to fulfil or achieve any of my other present desires or aims. On these assumptions, (H) does not imply that I have any reason to want to avoid this agony.

Subjectivists might also claim that

> (I) when it is true either that
>
> > (a) our *having* some desire or aim would have effects that we want,
>
> or that
>
> > (b) we *want* to have this desire or aim,
>
> these facts give us a reason to have this desire or aim, or at least give us a reason to cause ourselves to have or to keep this desire or aim, if we can. [89]

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

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

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

If (J) were true, all such desires or aims would be rationally self-justifying. My desire to avoid this agony would give me a reason to have this desire. But if I wanted to *be* in agony, this fact would give me a reason to want to be in agony. If I wanted to waste my life, this fact would give me a reason to want to waste my life. *Whatever* we want, our having such informed desires would give us reasons to have them. Since these claims are clearly false, [90] Subjectivists must reject (J). Since Subjectivists cannot appeal to (J), these people must agree that, in this version of *Case Three*, my desire to avoid my future agony gives me no reason to have this desire. Since I have no other present desire or aim that gives me any reason to have this desire, these people must now admit that, on their view, I have no reason to want to avoid this agony.

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

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

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

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

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

Return now to the claim that

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

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

claim contradicts all subjective theories, and is clearly false. So Subjectivists cannot appeal to (F).

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

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

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

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

We can now return to *Case Two*, in which I want to have some future period of agony, not as a means, but as an end, or for its own sake. I have no other present desire or aim that would be either fulfilled or prevented by this future agony, or by my desire to have this agony. After ideal deliberation, I have decided to cause myself to have this agony, if I can. Since Subjectivists must accept (L), they must admit that, on their view, I have most reason to cause myself to be in agony for its own sake. This act would best fulfil my present fully informed telic desires, and is what, after ideal deliberation, I have chosen to do. If there is a fire nearby, and I have no other way to fulfil my desire, I would have a decisive reason to thrust my hand into this fire. That is very hard to believe. Given my description of this case, there are, I believe, no facts that count even weakly in favour of my thrusting my hand into this fire. And I would have decisive reasons not to cause myself to be in agony in this way.

There could be many other, similar cases. According to subjective theories, if we had such informed desires to hit our howling baby, or to smash some malfunctioning machine, these facts would give us reasons to hit our baby and smash this machine. If what we most wanted and chose was to frustrate all of our future desires, this fact would give us a decisive reason to frustrate all of these desires. If what we most wanted and chose was to waste our lives, and to achieve

other bad or worthless aims, these facts would give us decisive reasons to waste our lives, and to try to achieve these bad or worthless aims. These claims are also very hard to believe. These implications of subjective theories give us decisive reasons, I believe, to reject all such theories.

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

Subjectivists *cannot*, however, make such claims. These claims appeal to differences between the reason-giving features of the *objects* of these desires or aims. If we make such claims, we have moved to an objective theory, which appeals to such object-given reasons. Subjectivists cannot distinguish in these ways between desires or aims that do or don't give us reasons. We are considering cases in which we know all the relevant facts. In such cases, we can argue:

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

Therefore

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

Therefore

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

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

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

Here is an overlapping argument for this conclusion. According to Objectivists, we have instrumental reasons to want something to happen, or to act in some way, when this event or act would have effects that we have some reason to want. As that claim implies, every instrumental reason gets its normative force from some other reason. This other reason may itself be instrumental, getting its force from some third reason. But at the beginning of any such chain, there must be some fact that gives us a reason to want some possible event as an end, or for its own sake. Such reasons are provided by the intrinsic features that would make this possible event in some way good. It is from such telic value-based object-given reasons that all instrumental reasons get their normative force.

Subjectivists must reject these claims. According to these people, instrumental reasons get their force, not from some telic reason, but from some telic desire or aim. We can have desire-based reasons to have some desire, and we can have long chains of instrumental desire-based reasons and desires. But at the beginning of any of *these* chains, as we have seen, there must always be some desire or aim that we have no such reason to have. And as my examples help us to see, we cannot defensibly claim that such desires or aims give us reasons. I would have no reason to thrust my hand into the fire. We would have no reason to hit our howling baby, or to waste our lives, or to try achieve countless other bad or worthless aims. So subjective theories are built on sand. Since all subject-given reasons would have to get their normative force from some desire or aim that we have no such reason to have, and such desires or aims cannot be defensibly claimed to give us any reasons, we cannot be defensibly claimed to have any subject-given reasons. We cannot have any such reasons to have any desire or aim, or to act in any way. [91]

13 The Incoherence Argument

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

As I have said, however, good theories about reasons must be able to be applied successfully to merely imaginary cases. Nor have I appealed only to such cases. I have argued that, in many actual cases, subjective theories imply that we have no reasons to want ourselves or others to avoid future periods of agony, or to have

future periods of happiness, and no reason to have many other good aims. And though subjective theories often have acceptable implications, this fact does not support these theories, since these theories have such implications only when they overlap with the best objective theories.

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

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

According to an implausible imaginary theory, which we can call

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

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

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

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

According to

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

When applied to actual people, these subjective theories often have acceptable implications. Since most of us often have the aims which the facts that we have considered give us decisive reasons to have, subjective theories often imply that we ought to try to achieve these aims. But that is not what these theories claim. In their claims about what we ought to do, these theories imply that we have no reasons to have our aims. When subjective aim-based theories have acceptable implications, that is because most actual people assume that they do have such reasons, and have aims that respond to them. These theories can seem plausible,

we might say, only because most people do not believe what these theories claim.

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

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

Many of these people also claim that

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

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

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

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

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

reason to accept (M). We would have no reason to believe that these better informed desires or aims have any higher reason-giving status, or are desires or aims that we have more reason to try to fulfil. [92]

Some Subjectivists make the weaker claim that

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

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

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

> frees us from assessing the rationality of a choice by means of the . . . task of assessing the thing chosen: we do not need to identify especially rational ends. [93]

To choose rationally, on Korsgaard's view, we needn't assess the merits of what we choose, since nothing *has* any such merits, by having any reason-giving features. But most Subjectivists do not see that, given their assumptions, we have no reason to try to have and to fulfil such better informed desires or aims. If Subjectivists cannot appeal to (M), (N), or (P), as I have just argued, that undermines the subtler and more plausible versions of Subjectivism, such as the Deliberative Theory and the Informed and Error-Free Desire Theories. These theories are incoherent, since they assume both that

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

and that

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

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

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

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

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

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

These people can make this claim because (T) does not assume that the things we want, or the possible outcomes of our acts, may have intrinsic reason-giving features. On the Telic Desire Theory, the relevant facts do not include facts about what these outcomes would be like, except when these are facts about what would best fulfil our actual present desires or aims. These Subjectivists can also coherently claim that

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

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

Though the Telic Desire Theory is not incoherent, it has several implausible implications which have led many Subjectivists to move to other, subtler theories. And my other objections apply. On this theory, we often have no reason to want avoid future agony, or to be happy, and we might have decisive reasons to cause ourselves to be in agony for its own sake, to waste our lives, and to try to achieve other bad or worthless aims. [94]

The Incoherence Argument, I have claimed, undermines the subtler and more plausible versions of Subjectivism. There is another, more positive way to state what this argument shows. When many Subjectivists appeal to what we would want or choose if we knew all the facts about the possible outcomes of our acts, these people rightly assume that these outcomes may have reason-giving features. Most of these people assume, for example, that we have object-given reasons to want to be

happy, and to avoid agony. These people are not really Subjectivists. When these people make Subjectivist claims, they are not correctly stating what they actually believe.

14 Reasons, Motives, and Well-Being

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

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

According to subjective theories about reasons, no events could be in this sense better than others, since there are no events that, from an impartial point of view, everyone would have more reason to want. It could not be better, for example, if some child's life were saved. There have been many people whose fully informed desires would not be better fulfilled when any child's life were saved. And even if everyone had such desires, subjective theories do not imply that everyone has *reasons* to have these desires, by having reasons to want any child's life to be saved. But that is what is meant by the claim that, in this impartial-reason-implying sense, it would be better if some child's life were saved. [95]

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

To reapply my earlier definition, if we call some possible life

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

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

If we accept some subjective theory about reasons, we cannot use 'best for someone' in this reason-implying sense. Subjective theories imply that there are no self-

interested reasons. Such reasons are provided by facts about the intrinsic features of future events that would make these events good or bad for us. Subjectivists deny that we have such reasons.

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

Of those who accept subjective theories about reasons, many use 'best for someone' in some sense that differs from the reason-implying sense. One example is the definition proposed by Rawls when he presents his *thin theory of the good*. On this definition,

> a person's good is determined by what is for him the most rational plan of life. [96]

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

> would adopt if he possessed full information. It is the objectively rational plan for him and determines his real good. [97]

If we call some life

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

Though it is a normative question which kinds of deliberation are procedurally rational, and in other ways ideal, it is a psychological question what, after such deliberation, someone would in fact choose. [98] On such views, there are no telic desires or aims that we are all rationally required to have, except perhaps those desires without which we could not even deliberate, choose what to do, and act.

The most rational plan of life for someone, Rawls writes, is the plan

> which would be chosen by him with full deliberative rationality, *that is*, with full awareness of the relevant facts and after a careful consideration of the consequences. [99]

We can be deliberatively rational in Rawls's sense whatever we have as our aims or ends. Rawls elsewhere claims that, from the fact that someone is *ideally rational*, we can infer nothing about what this person does or would want, or approve. [100] There is nothing, Rawls assumes, that we have any object-given reasons to want as an end.

To illustrate his theory of the good, Rawls imagines a man whose chosen plan is to spend his life counting the numbers of blades of grass in various lawns. Rawls writes that, on his theory, 'the good for this man is indeed counting blades of grass'. [101] This imagined man, Rawls assumes, would enjoy spending his life in this way. But on Rawls's theory, that assumption is not needed. It would be enough that, after rationally considering the relevant facts, this man would in fact choose this plan of life. For another example, consider

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

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

This example might be claimed to be unrealistic, because no one would choose a life of unrelieved suffering. As I have said, however, it is irrelevant whether such cases actually occur. Rawls does not assume that any actual person would choose to spend his life counting blades of grass, and Rawls rightly applies his theory to his merely imagined man. Any acceptable normative theory must be able to be applied successfully to such imaginary cases. And though it is hard to believe that anyone would choose a life of unrelieved suffering, that is because it is hard to believe that anyone could be so irrational as to choose a life that is so obviously bad in the reason-implying sense. On Rawls's view, however, no life could be bad for someone in this sense, since we have no object-given reasons. In Rawls's words, 'There is no way to get beyond deliberative rationality.' [102]

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

deliberation, Blue would choose.

Rawls intends, however, to be claiming more than this. Rawls's proposed sense of 'best for someone' is intended to replace the ordinary sense of this phrase, by giving us a clearer way of saying everything that we might want to say. [103] And Rawls, I believe, would want to say that it would be better for Blue if Blue's life did not consist of unrelieved suffering.

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

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

and others mean that

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

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

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

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

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

This means

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

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

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

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

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

Similar claims apply to actual cases. Subjective theories imply that we have no object-given reasons to want ourselves or others to live happy lives, and no such reasons to have any other good aim. And, as I have argued, Subjectivists cannot defensibly claim that we have *subject-given* reasons to have such aims, or to care about anything for its own sake. Such reasons would have to be provided by some desire or aim that we have no reason to have, and such desires or aims cannot be defensibly claimed to give us any reasons. So we can now conclude that, on these widely accepted views, *nothing matters*.

Some Subjectivists would admit that, on their view, nothing matters in an impersonal sense. It is enough, these writers claim, that some things matter to particular people. [104] But this reply shows how deep the difference is between the two kinds of theory that we have been considering. According to objective theories, some things matter in the normative sense that we have *reasons* to care about these things. When Subjectivists claim that some things matter to particular people, they mean only that these people *do* care about these things. That is not a normative but a merely psychological claim. We all know that people care about certain things. We hoped that philosophers, or other wise people, would tell us more than that.

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

15 Arguments for Subjectivism

These bleak views are seldom defended. Most Subjectivists take it for granted that reasons are provided by certain facts about our desires or aims. [105]

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

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

Therefore

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

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

Some Subjectivists argue:

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

Therefore

> All reasons for acting are desire-based. [106]

But (5) is false. We cannot defensibly claim that, whenever people are motivated to act for some reason, this reason must be subject-given and desire-based *rather* than object-given and value-based. That claim would have to assume that, for some

reason to be object-given and value-based, it must be impossible for anyone to be motivated to act for this reason. And that assumption would be absurd. If some act would achieve some aim that is good or worth achieving, some of us might be motivated to act for this reason.

These Subjectivists might reply

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

Therefore

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

Therefore

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

But (6) either confuses normative and motivating reasons, or claims that, whenever we act, we thereby give ourselves a normative reason for acting as we do. That claim would falsely assume that, any act however crazy would partly justify itself. [107] In taking (6) to imply (7), this argument falsely assumes that we cannot have reasons on which we fail to act. And (7) falsely assumes that value-based reasons might also be desire-based.

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

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

someone has a reason is, or consists in, some such causal or psychological fact.

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

> For the philosophical naturalist, concerned to place normativity within the natural order, there is nothing plausible for normative force to be other than motivational force. . . [108]

> there seems nothing for value to be, on deepest reflection, wholly apart from what moves, or could move, valuers, agents for whom something can matter. [109]

Object-given value-based reasons cannot be regarded in such ways, since we have such reasons even if we would *not* be moved or motivated to act upon them. [110]

Of the writers who give such reductive accounts, most claim to be describing normative reasons. But on such views, I believe, there aren't really any normative reasons. There are merely causes of behaviour. Things matter only in the sense that some people care about these things, and these concerns can move these people to act. [111]

Such Naturalist accounts of reasons are, I believe, mistaken. I defend this belief in Part Five, but I shall make one remark here. If Metaphysical Naturalism were true, we could not have reasons to have any particular beliefs. Such epistemic reasons are also irreducibly normative, and are therefore open to the same Naturalist objections. So it could not be true that we *ought* to accept Naturalism, nor could we have any reasons to accept this view. For us to be able to argue rationally about whether Naturalism is true, Naturalism must be false. [112]

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

CHAPTER 5 RATIONALITY

16 Practical and Epistemic Rationality

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

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

We can next look more closely at how the rationality of our desires and acts depends on our beliefs. My claims about our desires would also apply to our aims. Our desires and acts *causally* depend on our beliefs when we have these desires, and act in these ways, because we have these beliefs. Some desire might causally depend on some wholly irrelevant belief. We can imagine my wanting to go to sleep because I believe that 7 is a prime number. But if my desire directly depended on this belief, I would be mentally ill, or have some kind of local brain damage. 7's being a prime number gives me no reason to want to go to sleep. In most cases, when some desire depends on some belief, this relation is not merely causal. I may want to go to sleep because I believe that, unless I get some sleep, I shall perform badly in some interview tomorrow. Since this desire would be a rational response to what I believe, this desire would be not only caused by, but also *justified* by, my belief. [113] I shall now briefly sketch my view about how our desires and acts can be, or fail to be, justified by our beliefs.

The rationality of some of our desires depends only on their *intentional objects*, which are the possible events that we want, with the features that we believe these events

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

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

I would add:

> (B) In most cases, it is irrelevant whether these beliefs are true, or rational. Some of the exceptions involve certain normative beliefs.
>
> (C) When our beliefs are inconsistent, some of our desires or acts may be rational relative to some of our beliefs, but irrational relative to others. When we have no beliefs about the relevant, reason-giving facts, there may be nothing that we ought rationally to do.
>
> (D) Our having some desire is in one way rational when and because this desire itself is rational. But in some cases we could rationally cause ourselves to have some irrational desire. Our having this desire would then be, in a different way, rational. It could also be rational to cause ourselves to act irrationally. I discuss such cases further in Appendices B and C. [114]

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

Many people would reject some of these claims. Our desires are irrational, Hume suggests, just when these desires causally depend on false beliefs. [115] But false beliefs can be rational, and so can desires that depend on false beliefs.

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

irrational?

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

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

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

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

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

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

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

(1) smoking protects my health

and that

(2) I am now smoking,

I believe that

(3) I am now protecting my health.

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

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

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

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

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

To vary my example, suppose that I want to go to some crowded and noisy party because I believe that I shall enjoy it. This belief is irrational because I ought to have learnt by now that I never enjoy such parties. On the view expressed by (F),

given the irrationality of my belief, my desire to go to this party is in one way irrational. In another way, however, my desire is rational, since it is rationally supported by my beliefs. It is rational to want what I believe that I shall enjoy. And if I wanted to go to this party because I had the rational belief that I would *not* enjoy it, my desire would be in one way irrational.

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

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

We should not, I suggest, make either of these misleading claims. When some belief is epistemically irrational, this irrationality can be plausibly and usefully claimed to be *inherited* by any other belief that depends on this belief. But it is not worth claiming that some belief's irrationality is also inherited by any desire or act that depends on this belief. Given the differences between epistemic and practical reasons, we should turn to another, simpler view. We should claim that only beliefs can be epistemically irrational. Using a different metaphor, we might say that, when some belief is epistemically irrational, this irrationality can, like a virus, *infect* some of our other beliefs. But with a few exceptions to which I shall soon turn, this irrationality cannot be transmitted over the gap between our beliefs and our desires or acts. Our desires and acts are best called irrational only when, in having some desire or acting in some way, we are failing to respond to clear and strongly decisive *practical* reasons or apparent reasons not to have this desire, or not to act in this way.

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

reasons, while responding badly to the other set. We might be practically rational but epistemically irrational, or practically irrational but epistemically rational.

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

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

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

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

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

17 Beliefs about Reasons

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

preferences could be unreasonable.

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

I have claimed that, in *most* cases, our desires are rational if these desires depend upon beliefs whose truth would give us sufficient reasons to have these desires. I have also claimed that, in such cases, it is irrelevant whether our beliefs are true, or rational. These claims do not apply when our desires partly depend on certain *normative* beliefs. It may be relevant whether *these* beliefs are true, or rational. Suppose that we falsely and irrationally believe both that some fact gives us a reason to have some desire, and that this desire is rational. If these beliefs were true, we would have a reason to have this desire, and this desire would be rational. That does not make it true that we actually have such a reason, nor does it make this desire rational. Similar claims apply to our acts. If we falsely and irrationally believe that we have a reason to act in some way, or that some act would be rational, that does not give us such a reason, nor does it make this act rational. Practical rationality is not so easily achieved. [116]

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

As before, however, that claim applies only to most cases. When our beliefs are about practical reasons, these kinds of rationality and reason overlap. As Scanlon notes, many of our desires can be more fully described as states of being motivated by the belief that something would be good, or worth achieving, in the reason-implying sense. [117] Given this relation between these desires and beliefs, the rationality of these desires *does* in part depend on the rationality of these beliefs. And if we have irrational beliefs about practical reasons, and about what we ought rationally to want or to do, our having such beliefs makes us in one way practically irrational.

There is a similar overlap between practical reasons and certain epistemic reasons. We have a practical reason, for example, to want to avoid being in agony, and an

epistemic reason to believe that we have this practical reason. The nature of agony both gives us this practical reason, and gives us this epistemic reason by making it obviously true that we have this practical reason.

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

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

To illustrate this disagreement, suppose that

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

and

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

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

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

Crimson's view is closer to the views that many actual people accept. Crimson believes that, though we have reasons to care about all of our future, we have much stronger reasons to care about our nearer future. Crimson therefore believes that he

has decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain later today. Crimson has this preference, so he chooses to have the agony.

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

When Scanlon discusses someone with Scarlet's preference, he writes that 'such a person would not be irrational, but only substantively mistaken'. We should call someone irrational, Scanlon suggests, only when this person 'fails to respond to what he or she acknowledges to be relevant reasons'. [118]

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

Crimson's preference is less irrational, since this preference does not draw an arbitrary line, and it is not implausible to believe that we have reasons to care more about our nearer future. [119] But Crimson's version of this view is much too extreme. It is irrational to believe that we have decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain later today. Since Crimson's preference matches his belief about his reasons, he too avoids one kind of irrationality. But in preferring this agony to this slight pain, Crimson is failing to respond to a clear and strongly decisive reason, and his preference matches his belief only because both are irrational.

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

People are most *clearly* irrational, Scanlon claim, when they fail to respond to what they themselves acknowledge to be reasons. This claim is in one way true, since such people are less than fully rational even according to their own beliefs. If these people were accused of not being fully rational, they would plead guilty. But that

does not justify the claim that only such people should be called irrational. On Scanlon's view, even if we often fail to respond to very clear and decisive reasons, we could avoid irrationality merely by having no beliefs, or false beliefs, about which facts give us reasons, and about which desires or acts are rational. We ought, I believe, to reject this view. Scarlet's attitude to future Tuesdays is irrational even though he believes it to be rational. And if we have rational beliefs about practical reasons, and we admit our failures to respond to these reasons, we may be less irrational than those who have irrational beliefs and much greater unadmitted faults.

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

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

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

I have rejected Scanlon's claim that, when people like Scarlet and Crimson prefer an hour of agony to a minute of slight pain, these people's preferences are not irrational. There may, however, be no disagreement here. I am using 'irrational' in its ordinary sense, to mean, roughly, 'deserves strong criticism of the kind that we also

express with words like "foolish", "stupid", and "crazy"'. At one point Scanlon suggests that we should use 'irrational' in what he calls a narrower sense, which applies only to people who fail to respond to what they themselves believe to be reasons, or who are inconsistent in certain other ways. [120] If Scanlon is using 'irrational' in this narrower sense, his view may not conflict with mine. When Scarlet prefers an hour of agony to a minute of slight pain, his preference is not, I agree, in *this* sense irrational. And Scanlon might agree that Scarlet is making a very great substantive mistake, and that, compared with Pink's preference for an extra minute of slight pain, Scarlet's preference for an hour of agony deserves much stronger rational criticism. If this is Scanlon's view, however, it would be misleading for him to say that only Pink's preference is irrational, since that would suggest that *Pink's* preference deserves stronger criticism. We ought, I believe, to use 'irrational' in its ordinary, wider sense. If we believe that one of two preferences deserves much stronger rational criticism, we shouldn't say that only the other preference is irrational.

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

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

Unlike Scarlet and Crimson, moreover, many of these actual people have rational desires and preferences. Though these people believe that they have no reason to care about their future well-being, they do care. And they may care equally about the whole of their future, so that they would never postpone some ordeal if they believed that this would merely make this ordeal more painful. Such people

respond rationally to the facts that give them reasons to care about their future well-being, and they *do*, in this way, respond to these reasons. Their mistake is only in their failing to believe, at the conscious level, that they have these reasons. Some Subjectivists may even have such beliefs, and act upon them in their non-academic lives, ignoring or rejecting these beliefs only when they teach or write. (This is like the way in which many economists believe, but only when they teach or write, that interpersonal comparisons of well-being make no sense.)

18 Other Views about Rationality

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

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

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

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

According to another common view, our desires are rational when our having them has good effects. But if some whimsical despot credibly threatens to torture me unless, one hour from now, I want to be tortured, that would not make this desire rational. This despot's threat might make it rational for me to cause myself to have

this irrational desire, if I can. My *having* this desire would then be, in one way, rational. But this desire *itself* would still be irrational. This would be a case of rational irrationality. [121]

According to some writers, the rationality of our desires partly depends on certain other facts about their origin. Our desires are rational, these writers claim, if they were formed through autonomous deliberation, and irrational if they were formed in certain other ways, such as by indoctrination or hypnosis. We ought, I believe, to reject such views. Our desires may be rational even if we were hypnotized or indoctrinated into having them. If we care little about our future, for example, we might be hypnotized into having such rational concern. Or we might be indoctrinated into loving our enemies, and wanting to do at least one good deed in every day. Such love and such desires are, I believe, fully rational. Suppose next that, after autonomous deliberation, we want to starve ourselves to death, thereby losing what would have been a happy life, or we have some other desire for something that is wholly undesirable. The autonomous origin of these desires would not make either them, or us, rational. On the contrary, we would be *less* irrational if, rather than forming these desires through autonomous deliberation, we were made to have them by some form of outside interference, like hypnosis.

According to some other, similar views, the rationality of our desires depends, not on how we came to have them, but on what would cause us to lose them, or on whether they would survive certain tests. Our desires should be called rational, Richard Brandt suggests, if these desires would survive our being given some course of cognitive or belief-based psychotherapy. On this account, our desires might be rational because we are incurably insane. That is not a helpful claim.

According to another group of views, our desires or preferences are irrational when they are *inconsistent*. Two beliefs are inconsistent if they could not both be true. This definition cannot be applied directly to desires, since desires cannot be true. But two desires can be inconsistent, many writers claim, in the sense that these desires could not both be fulfilled.

Such inconsistency involves no irrationality. Suppose that, after some shipwreck, I could save either of my two children, but not both. Even when I realize this fact, I could rationally go on wanting to save both my children. If we know that two of our desires cannot both be fulfilled, that might make it irrational for us to *aim* or *intend* to fulfil both desires. But these desires may still be in themselves rational, and it may still be rational for us to have them. When our desires are, in this sense, inconsistent, that might make our having them unfortunate. As I have claimed, however, that does not make such desires irrational.

For inconsistency to be a fault, it must be defined in a different way. Though desires cannot be true or false, many desires depend on beliefs about what is good or bad, and these beliefs might be inconsistent, so that they could not all be true. Our desires might be claimed to be derivatively inconsistent when they depend on such inconsistent normative beliefs.

That would be true, it may seem, if we both wanted something to happen, and wanted it not to happen. In having these desires, we might seem to be inconsistently assuming that it would be both better and worse if this thing happened. But in most cases of this kind, we are assuming that some event would be in one way good and in another way bad. For example, I might want to finish my life's work, so as to avoid the risk of dying with my work unfinished, and also want *not* to finish my life's work, so that, while I am alive, I would still have important things to do. Such desires and normative beliefs involve no inconsistency. For two of our desires to be irrationally inconsistent in this belief-dependent way, these desires must depend on beliefs that the very same thing would be both good and bad in the very same way. It is not clear that it would be possible to have such beliefs and desires; but, if it were, the objection that appeals to inconsistency would here be justified.

When we turn to larger sets of preferences, there is more scope for inconsistency. We might prefer B to A, C to B, and A to C. Such preferences are called *cyclical*. If these were *mere* preferences which did not depend on normative beliefs, it is not clear that such a set of preferences could be claimed to be irrational. This claim is often defended with the remark that, if we had such cyclical preferences, we could be exploited. We might be induced to pay three sums of money first to have B rather than A, then to have C rather than B, and then to have A rather than C. Our money would be wasted, since we would be back with A, where we started. But this objection appeals, not to any inconsistency in such a set of preferences, but to their bad effects. And if we had such preferences, that might have some good effects. Suppose that, whenever our situation changed in some way that we preferred, that change would give us some pleasure. If we had three such cyclical preferences about three easily changeable situations X, Y, and Z, this would be, in a minor way, good for us. We could go round and round this circle, getting pleasure from every move. This merry-go-round would be, hedonically, a perpetual motion machine.

Things are different when such preferences depend on certain normative beliefs. Suppose that we have these preferences because we believe that X is better than Y, which is better than Z, which is better than X. Such beliefs would be inconsistent if, as we can plausibly and I believe truly claim, the relation *better than* is *transitive*.[122] On this view, just as I can't be taller than you if you are taller than someone who is taller than me, X can't be better than Y if Y is better than Z which is better than X. If such beliefs are inconsistent, that could be claimed to make such preferences

derivatively irrational. Though cases that involve such preferences are theoretically very interesting, they do not, I believe, have much practical importance. [123]

The rationality of our desires does not depend, I have claimed, either on their origin, or on their consistency with our other desires. Of those who propose these criteria, some may be misled by presumed analogies with beliefs. The rationality of most of our beliefs *does* depend either on their origin, or on their consistency with our other beliefs, or both. There are relatively few beliefs whose rationality depends only on their content: or *what* we believe. That is true of beliefs about some necessary truths or falsehoods, such as mathematical or logical beliefs. Some belief is intrinsically irrational, for example, if what we believe is some obvious contradiction. But most of our beliefs are *empirical* and *contingent*, in the sense that they are beliefs about how the observable spatio-temporal universe happens to be. There are some empirical beliefs whose rationality depends only on their content. Two examples may be Descartes' belief 'I exist,' and the more cautious Buddhist belief 'This is the thinking of a thought'. Perhaps these beliefs must be true, in a way that makes them intrinsically rational. But few empirical beliefs are of this kind. Some empirical beliefs---such as the belief of some psychotic person that he is Napoleon or Queen Victoria---might seem to be, simply in virtue of their content, irrational. But the irrationality of even these beliefs is still mostly a matter of their origin, and of whether they conflict with our other beliefs. The rationality of most empirical beliefs cannot depend only on their content, because such beliefs are true only if they match the world. What we can rationally believe about the world depends on our other beliefs, our perceptual experiences, and the other evidence available to us.

No such claims apply to our intrinsic telic desires. The rationality of these desires does not depend on how they arose, or on their consistency with our other desires. When we want something as an end, or for its own sake, the rationality of this desire depends only on our beliefs about this desire's object, or what we want. These desires are rational, as objective value-based theories claim, when they depend on beliefs whose truth would make their objects in some way good, or worth achieving. This is the central, fundamental truth that is either ignored or denied by most of the theories that we have been considering.

In rejecting these analogies between the rationality of our beliefs and our desires, I am not forgetting that many of our desires depend upon normative beliefs. These beliefs are about truths that are not empirical and contingent, but necessary. Undeserved suffering, for example, could not have failed to be in itself bad. For such normative beliefs to be rational, we do not need to have evidence that they match the actual world, since these beliefs would be true in any possible world.

CHAPTER 6 MORALITY

19 Sidgwick's Dualism

Objective theories about reasons can differ in several ways. One difference is in the range of events that these theories claim to be good or bad in the reason-implying senses. One of two outcomes would be worse, some theories claim, only if it would be worse *for* one or more people. That, I shall argue, is not true. Nor is it only outcomes that are worth achieving, since some acts are in themselves good. And some things may be worth doing only for their own sake.

Objective theories also differ in their claims about whose well-being we have reasons to promote. We can next consider three such theories. According to

> *Rational Egoism*: We always have most reason to do whatever would be best for ourselves.

According to

> *Rational Impartialism*: We always have most reason to do whatever would be impartially best.

Some act of ours would be impartially best, in the reason-implying sense, if we do what, from an impartial point of view, everyone would have most reason to want us to do. On one view, what would be impartially best is whatever would be, on balance, best for people, by benefiting people most.

In his great, drab book *The Methods of Ethics*, Sidgwick qualifies and combines these two views.[124] According to what Sidgwick calls

> *the Dualism of Practical Reason*: We always have most reason to do whatever would be impartially best, unless some other act would be best for ourselves. In such cases, we would have sufficient reasons to act in either way. If we knew the relevant facts, either act would be rational.[125]

Of these three views, Sidgwick's, I believe, is the closest to the truth. According to Rational Egoists, we could not have sufficient reasons to do what would be worse for

ourselves than some other possible act. That is not true. We might have such reasons, for example, when and because our act would make things go impartially much better. I would have sufficient reasons to injure myself if that were the only way in which some stranger's life could be saved. According to Rational Impartialists, we could not have sufficient reasons to do what would be impartially worse than some other possible act. That is not true. We might have such reasons, for example, when and because our act would be much better for ourselves. I would have sufficient reasons to save my own life rather than the lives of several strangers.

On Sidgwick's view, we have both impartial and self-interested reasons for acting, but these reasons are not *comparable*. That is why, whenever one act would be impartially best but another act would be best for ourselves, we would have sufficient reasons to act in either way. No reason of either kind could be outweighed by any reason of the other kind.

Some reasons are *precisely* comparable in the sense that there are precise truths about their relative weight or strength. According to some desire-based subjective theories, all reasons are precisely comparable, since there are precise truths about the relative strengths of all of our desires. According to value-based objective theories, when we must choose between two things that are very similar, such as two cherries or two copies of some book, we may have precisely equal reasons to choose---or, as we could better say, *pick*---either of these things. And when we are comparing reasons of the same kind---such as reasons that are provided by differences in the costs of what we might buy, or differences in the length of otherwise similar pleasures and pains---the strengths of these reasons may be precisely comparable. But when we compare most reasons of different kinds, these reasons are much less comparable.

Two such dissimilar reasons might be provided by the greater length of one of two possible pains and the greater intensity of the other. If we must choose between one brief but intense pain and another pain that would be much longer but much less intense, one of these possible experiences might be worse, in the sense that we would have more reason to prefer the other. But there could not, I suggest, be any precise truth about the relative strength of these reasons. One of these pains could not, for example, be 2.36 times worse than the other. Even in principle, there is no scale on which we could precisely compare the strengths of our reasons to avoid two such different pains. These claims might be challenged, because the length and intensity of pains both contribute to the same kind of badness. But there are other, clearer cases. There are only very imprecise truths about the relative strength of many other different kinds of reason, such as economic and aesthetic reasons, or our reasons to keep our promises and to help strangers. Such reasons *are* comparable, however, since some weak reasons of either kind could be weaker than, or be outweighed by, some strong reasons of the other kind.

According to Sidgwick's Dualism, in contrast, impartial and self-interested reasons are *wholly* incomparable. *No* impartial reason could be either stronger or weaker than *any* self-interested reason. Views of this kind are hard to defend. Suppose that we are choosing between some architectural plans for some new building. When neither of two conflicting reasons outweighs the other, we could rationally act in either way. If economic and aesthetic reasons were wholly incomparable, it would therefore be true both that

> (1) we could rationally choose one of two plans because it would make this building cost one dollar less, even though this building would be very much uglier,

and that

> (2) we could also rationally choose one of two other plans because it would make this building slightly less ugly, even though this building would cost a billion dollars more.

We can perhaps imagine how one of these choices might be rational, since we might have reasons to give absolute priority either to this building's beauty, or to its cost. But it would be most implausible to claim that we could rationally make *both* these choices. As this example suggests, to defend Sidgwick's view that impartial and self-interested reasons are wholly incomparable, it is not enough to claim that these reasons are of different kinds.

Sidgwick's defence of his view appeals in part to the rational significance of personal identity. Given the unity of each person's life, we each have strong reasons, Sidgwick claims, to care about our own well-being, in our life as a whole. [126] And given the depth of the distinction between different people, it is rationally significant that one person's loss of happiness cannot be compensated by gains to the happiness of others. Sidgwick here appeals to the *separateness of persons*, which has been claimed to be 'the fundamental fact for ethics.' [127]

Sidgwick's Dualism also rests on what Thomas Nagel calls our *duality of standpoints*. [128] We live our lives from our own personal point of view. But we can also think about the world, and all the people in it, as if we had the impartial point of view of some detached observer. When we ask what we have most reason to do, we reach different answers, Sidgwick claims, from these two points of view. [129] From our own point of view, self-interested reasons are *supreme*, in the sense that we always have most reason to do whatever would be best for ourselves. From an impartial point of view, impartial reasons are supreme, since we always have most reason to do whatever would be impartially best. [130]

Suppose next that one possible act would be impartially best, but that some other act

would be best for ourselves. Impartial and self-interested reasons would here conflict. In such cases, we could ask what we had most reason to do all things considered. But this question, Sidgwick claims, would never have a helpful answer. We could never have more reason to act in either of these ways. 'Practical Reason' would be 'divided against itself', and would have nothing to say, giving us no guidance. [131] This conclusion seemed to Sidgwick deeply unsatisfactory.

Sidgwick's reasoning seems to be this:

> (A) When we try to decide what we have most reason to do, we can rationally ask this question either from our own personal point of view or from an imagined impartial point of view.
>
> (B) When we ask this question from our personal point of view, the answer is that self-interested reasons are supreme.
>
> (C) When we ask this question from an impartial point of view, the answer is that impartial reasons are supreme.
>
> (D) To compare the strength of these two kinds of reason, we would need to have some third, neutral point of view.
>
> (E) There is no such point of view.

Therefore

> Impartial and self-interested reasons are wholly incomparable. When such reasons conflict, no reason of either kind could be stronger than any reason of the other kind.

Therefore

> In all such cases, we would have sufficient reasons to do either what would be impartially best, or what would be best for ourselves. If we knew the facts, either act would be rational.

We can call this the *Two Viewpoints Argument*.

Sidgwick's view is, I believe, partly true. But we ought to reject this argument, and revise this view.

We should reject premise (A). It can be worth asking what we would have most reason to want, or prefer, if we were in the impartial position of some outside observer. By appealing to what everyone would have such impartial reasons to want or prefer, we can more easily explain one important sense in which outcomes

can be better or worse. But when we are trying to decide what we have most reason to do, we ought to ask this question from our actual point of view. We should not ignore some of our actual reasons merely because we would not have these reasons if we had some other, merely imagined point of view.

We should also reject (D). To be able to compare partial and impartial reasons, we don't need to have some third, neutral point of view. We can compare these two kinds of reason from our actual, personal point of view.

When we compare these reasons, we can next reject premise (B). On Sidgwick's view, we could rationally do what we knew would be only very slightly better for ourselves, and would be impartially very much worse. For example, we could rationally save ourselves from one minute of discomfort rather than saving a million people from death or agony. If we acted in such a way, the main reactions of others would be horror and indignation. But our question here is: Would this act be rational?

Some people would answer Yes. According to these people, if we knew that this act would best fulfil our present desires, or would be best for us, this act, however horrendous, *would* be rational. Of those who hold such views, however, many use 'rational' in either the present-desire-based sense or the egoistic sense. If these people claimed that this act would be rational, some of them would mean that, in doing what would best fulfil our present desires, we would be doing what would best fulfil these desires. Others would mean that, in doing what we would be best for ourselves, we would be doing what would be best for ourselves. We can ignore such trivial claims. When I ask whether this act would be rational, I am not using 'rational' in either of these senses. I am asking whether this act would deserve one kind of criticism. We act rationally, I believe, only when we have beliefs about the relevant facts whose truth would give us sufficient reasons to act as we do.

In my imagined case, we know the relevant facts. Would we have sufficient reasons to save ourselves from mild discomfort, rather than saving a million people from death or agony? The answer, I believe, is No. This horrendous act would not be rational.

Such acts would not be rational, we might add, because they would be morally wrong. Sidgwick assumes that our self-interested reasons cannot be weaker than, or be outweighed by, our reasons to avoid acting wrongly. We should reject this assumption.

We might also reject Sidgwick's claim that we could always rationally do whatever we knew would make things go best. As an *Act Consequentialist*, Sidgwick believes that such acts would always be morally right. Most of us reject this view, since we

believe that certain acts would be wrong even if they would make things go best. The wrongness of such acts, we might claim, would often give us decisive reasons not to act in these ways.

I shall soon turn to questions about morality, and about our reasons to avoid acting wrongly. But we can first revise Sidgwick's view in other ways. This view overstates the rational importance of personal identity. Sidgwick rightly claims that we have reasons to be specially concerned about our own future well-being. But we have other, similar reasons. Our reasons to care about our future are at least in part provided, not by the fact that this future will be *ours*, but by various psychological relations between ourselves as we are now and our future selves. Most of us have partly similar relations to some other people, such as our close relatives, and those we love. These are the people, I shall say, to whom we have *close ties*. Our relations to these people can give us reasons to be specially concerned about their well-being. [132] We can have reasons to benefit these people that are much stronger than some of our reasons to benefit ourselves. So we should reject Sidgwick's claim that, when assessed from our personal point of view, self-interested reasons are supreme.

As well as having these *personal* and *partial* reasons to care about the well-being of ourselves and those to whom we have close ties, we also have *impartial* reasons to care about everyone's well-being. Some of Sidgwick's claims imply that we have such reasons only when we consider things from an impartial point of view. But that is not so. Imagining himself as an egoist, Nagel writes:

> Suppose I have been rescued from a fire and find myself in a hospital burn ward. I want something for the pain, and so does the person in the next bed. He professes to hope that we will both be given morphine, but I fail to understand this. I understand why he has reason to want morphine for himself, but what reason does he have to want *me* to get some? Does my groaning bother him? [133]

This egoistic attitude would be, as Nagel remarks, 'very peculiar.' Unless we have been taught to accept some desire-based subjective theory, or we lack the concept of a reason, most of us rightly believe that we have some reason to want any stranger's pain to be relieved. [134] And we have such impartial reasons even when our actual point of view is not impartial. As I have said, we can have reasons to benefit strangers that conflict with, and are much stronger than, some of our self-interested reasons. Rather than saving ourselves from some minor harm, we would have much stronger reasons to save many strangers from death or agony.

Sidgwick's view, however, is partly right. Our partial and impartial reasons are, I

believe, only *very imprecisely* comparable. According to what we can call

> *wide value-based objective views*: When one of our two possible acts would make things go in some way that would be impartially better, but the other act would make things go better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either of these ways.

The word 'often' allows for various exceptions. Different wide value-based objective views make conflicting further claims about when it would *not* be true that we had sufficient reasons to act in either of these ways. We ought, I believe, to accept some view of this kind. [135]

To illustrate such a view, we can suppose that, in

> *Case One*, I could either save myself from some injury, or act in a way that would save some stranger's life in a distant land,

and that, in

> *Case Two*, I could save either my own life or the lives of several distant strangers.

In both cases, on most people's views, I would be *morally* permitted to act in either way. I would also be *rationally* permitted, I believe, to act in either way. In *Case One* I would have sufficient reasons either to save myself from some injury or to save this stranger's life. And I might have such reasons whether my injury would be as little as losing one finger, or as great as losing both legs. In *Case Two*, I would have sufficient reasons to save either my own life or the lives of the several strangers. And I might have such reasons whether the number of these strangers would be two or two thousand. Though my reason to save *two* strangers would be *much* weaker than my reason to save *two thousand* strangers, both these reasons might be neither weaker nor stronger than my reason to save my own life. If these claims are true, the relative strength of these two kinds of reason is very imprecise.

There is such great imprecision, we can claim, because these reasons are provided by very different kinds of fact. Our impartial reasons are *person-neutral*, in the sense that these reasons are provided by facts whose description need not refer to us. One example is the fact that some event would cause great suffering. We all have reasons to regret anyone's suffering, and to prevent or relieve this person's suffering if we can, whoever this person may be, and whatever this person's relation to us. We have such reasons to prevent or to regret the suffering of any *sentient* or conscious being. When we are in pain, as Nagel writes,

> the pain can be detached in thought from the fact that it is mine without losing any of its dreadfulness. . . suffering is a bad thing, period, and not just for the sufferer. . . *This experience* ought not to go on, *whoever* is having it. [136]

Our personal and partial reasons are, in contrast, *person-relative.* These reasons are provided by facts whose description must refer to us. We each have such reasons to be specially concerned about the well-being both of *ourselves* and of those other people who are in certain ways *related to us*. Though I would have reasons to prevent both my own pain and the pain of any distant stranger, my relation to *myself*, and to *my* pain, is very different from my relation to that stranger, and to that stranger's pain. That is why these reasons are so imprecisely comparable.

According to some wide value-based views, when we are choosing between morally permissible acts, our reasons to give ourselves some benefit are always stronger than, or outweigh, our reasons to give the same benefit to strangers; but this difference is very imprecise. On one such view, we are rationally required to give to our own well-being more weight than we give to any stranger's well-being, but this greater weight could be as little as twice as much or as great as a hundred or a thousand times as much.

These views are, I believe, too egoistic. We could often rationally give equal or even greater weight to some stranger's well-being. Suppose that, like Nagel, I am in pain in some hospital ward, and the only dose of morphine belongs to me. I would have sufficient reasons, I believe, to give this morphine to the stranger in the next bed. And I would have such reasons even if this stranger's pain was less bad than mine.

Such acts are rational, it might be claimed, only when we are denying ourselves some fairly small benefit. Suppose instead that, in

> *First Shipwreck*, I could use some life-raft to save either my own life or the life of a single stranger. This stranger is relevantly like me, so our deaths would be, for each of us, as great a loss.

When the stakes are as high as this, we may seem to be rationally required to give significant priority, or much greater weight, to our own well-being. If that is true, I would not have sufficient reasons to save this stranger rather than myself. This act, even if morally admirable, would not be fully rational.

I am inclined to believe that this act *might* be fully rational. This stranger's well-being matters just as much as mine. And if I gave up my life to save this stranger, this act would be generous and fine. These facts might, I believe, give me sufficient reasons to act in this way. [137]

There is, I must admit, a strong objection to this view. I believe that, as Sidgwick claims, we have reasons to be specially concerned about our own well-being. And in this imagined case, my death would be impartially as bad as the stranger's death. Since I would have *equal* impartial reasons to save either myself or this stranger, my

self-interested reasons might be claimed to break this tie, or tip the scale, giving me decisive reasons, all things considered, to save myself. [138]

These reasons may not, however, be decisive. Even when the stakes are very high, we may not be rationally required to give any priority to our own well-being. We might be able to defend a revised version of Sidgwick's view. According to what we can call

> *Permissive Dualism:* When we are choosing between two morally permissible acts, of which one would be better for ourselves and the other would be better for one or more strangers, we could rationally either give greater weight to our own well-being, or give roughly equal weight to everyone's well-being.

Different versions of this view make different further claims. Though such views do not rationally *require* us to give greater weight to our own well-being, they may *permit* us to give *much* greater weight to our own well-being. And they *do* require us *not* to give much greater weight to any stranger's well-being. On some versions of this view, for example, I could rationally save one of my fingers rather than saving some stranger's life, but I could *not* rationally save some *stranger's* finger rather than saving *my* life. In permitting us to give such great priority to our own well-being, but requiring us *not* to give such great priority to the well-being of strangers, Permissive Dualism recognizes and endorses our reasons to be specially concerned about our own well-being.

Suppose next that, in

> *Second Shipwreck*, I could save either some stranger's life or the life of someone to whom I have close ties, such as one of my children, or some friend.

As Permissive Dualists could claim, I could not rationally choose to save this stranger. I ought morally to give priority to my child. I would have other strong non-moral reasons to act in this way, such as the reasons that are involved in my love for my child or friend. And if I saved this stranger rather than my child or friend, this act would *not* be generous and fine.

Similar claims might apply to *First Shipwreck*. I might have young children who depend on me, or have other obligations to certain other people. That might make it wrong for me to save some stranger rather than myself, since I could not then care for my children, or fulfil these other obligations. This stranger might have similar obligations that his death would cause to be unfulfilled, but those obligations would not be mine. And if my death would be bad for those who love me and are loved by me, that would give me other decisive reasons to save my life. So in this version of *First Shipwreck*, I would be rationally required to save myself.

Suppose next that I have no such reason-giving and obligation-involving ties to

certain other people. In this other version of this case, I am inclined to believe that I could rationally choose to give up my life to save this stranger. In such cases, we may be rationally permitted to ignore our reasons to be specially concerned about our own well-being. But we need not here decide whether that is true, or whether my act, though morally admirable, would be less than fully rational.

20 The Profoundest Problem

We can now turn to the relations between reasons and morality. According to

> *Moral Rationalism*: We always have most reason to do our duty. It could not be rational to act in any way that we believe to be wrong.

According to

> *Rational Egoism*: We always have most reason to do what would be best for ourselves. It could not be rational to act in any way that we believe to be against our own interests.

Many people accept both these views. Most of these people believe that duty and self-interest never conflict, since each of us will have some future life in which, if we have done or failed to do our duty, we shall get the happiness or suffering that we deserve. That is claimed by most of the world's great religions.

Sidgwick doubted that we shall have some future life, and he thought it to be likely that, in some cases, duty and self-interest conflict. If there are such cases, Sidgwick claims, that would raise 'the profoundest problem in ethics'. [139]

Sidgwick's problem was in part that Moral Rationalism and Rational Egoism both seemed to him intuitively very plausible, but that, if duty and self-interest sometimes conflict, these views cannot both be true. If we had to choose between two acts, of which one was our duty but the other would be better for ourselves, these views imply that we would have most reason to act in each of these ways. That is inconceivable, or logically impossible. Just as we could not keep most of our money in each of two different wallets, we could not have most reason to act in each of two different ways. So if duty and self-interest sometimes conflict, we would have to reject or revise at least one of these views.

When they consider these alternatives, some writers reject Moral Rationalism. Thomas Reid, for example, claims that, if it would be against our interests to do our duty, we would be 'reduced to this miserable dilemma, whether it be best to be a knave or a fool'. [140] We would be knaves if we didn't do our duty, but fools if we did. Other writers reject Rational Egoism. According to these people, we could

never have sufficient reasons to act wrongly, not even if that was our only way to save ourselves from great pain or death.

Sidgwick found such claims incredible. Rather than rejecting one of these views, he revised them both. According to another version of Sidgwick's Dualism, which we can call

> *the Dualism of Duty and Self-Interest:* If duty and self-interest never conflict, we would always have most reason both to do our duty and to do what would be best for ourselves. But if we had to choose between two acts, of which one was our duty but the other would be better for ourselves, reason would give us no guidance. In such cases, we would not have stronger reasons to act in either of these ways. If we knew the relevant facts, either act would be rational. [141]

Partly because he accepted this view, Sidgwick passionately hoped that duty and self-interest never conflict. If there are such conflicts, he writes,

> the whole system of our beliefs as to the intrinsic reasonableness of conduct must fall. . . the Cosmos of Duty is thus really reduced to a Chaos, and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure. [142]

These magnificently sombre claims are, however, overstatements. Sidgwick believed that in most cases duty and self-interest do not conflict. Sidgwick's view implies that, in these many cases, we would have most reason to do our duty, at no cost to ourselves. In such a world, the cosmos of duty would not be a chaos. Nor would our whole system of beliefs about what is reasonable conduct fall if we concluded that, when duty and self-interest conflict, we could reasonably, or rationally, act in either way. But it would be bad if, in such cases, we and others would have sufficient reasons to act wrongly. The *moralist's problem*, we might say, is whether we can avoid that conclusion. And it would be disappointing if, in such cases, reason gave us no guidance. We may hope that, in at least in some of these cases, there would be something that we had most reason to do. The *rationalist's problem*, we might say, is whether that is true.

These problems might take other forms. Sidgwick assumes that, if we had sufficient reasons to act wrongly, these reasons would be self-interested. We should not make that assumption, since we can have other strong reasons to act wrongly. Some of these reasons are personal and partial, but not self-interested. We might have sufficient reasons to act wrongly, for example, if some wrong act was our only way to save from great pain or death, not ourselves, but our close relatives, or other people whom we love.

We might also have strong impartial reasons to act wrongly. As an Act Consequentialist, Sidgwick claims that we ought always to do whatever would make things go best. Most of us reject this view, since we believe that some acts would be wrong even if they would make things go best. It might be wrong to kill someone, for example, even when that is the only way in which many other people's lives could be saved. Even if this act would be wrong, however, the fact that we would be saving many people's lives, thereby making things go best, might be claimed to give us sufficient reasons to act in this way. If that were true, this would be another kind of case in which we could rationally act wrongly.

There is a third possibility. On Sidgwick's view, we always have sufficient reasons to do our duty, and to avoid acting wrongly. We can call this view *Weak Moral Rationalism*. If we are Subjectivists about Reasons, we must reject this view. Rawls for example claims that, if our present informed desires would be best fulfilled by acting unjustly, we would not have sufficient reasons to do what justice requires.[143] According to such subjective theories, we might have no reason to do our duty, and decisive reasons to act wrongly. It might then be *irrational* for us to do our duty.

To cover these various possibilities, we can revise Sidgwick's description of what he calls 'the profoundest problem'. When we are choosing between different possible acts, we can ask:

> Q1: What do I have most reason to do? Do I have sufficient or decisive reasons to act in any of these ways?
>
> Q2: What ought I morally to do? Would any of these acts be wrong?

These questions might, it seems, have conflicting answers, since we might sometimes have sufficient or decisive reasons to act wrongly. Our problem is to decide whether we do or could have such reasons, and, if that is true, what further conclusions we should draw.

In considering these questions, it will help to distinguish between two conceptions of normativity. On the *reason-involving* conception, normativity involves reasons or apparent reasons. On the *rule-involving* conception, normativity involves requirements, or rules, that distinguish between what is *correct* and *incorrect,* or what is *allowed* and *disallowed*. Certain acts are required, for example, by the law, or by the code of honour, or by etiquette, or by certain linguistic rules. It is illegal not to pay our taxes, dishonourable not to pay our gambling debts, and incorrect to eat peas with a spoon, to spell 'committee' with only one 't', and to use 'refute' to mean 'deny'. Such requirements or rules are sometimes called 'norms'.

These conceptions of normativity are very different. On the rule-involving conception, we can create new normative truths merely by introducing, or getting some people to accept, some rule. Legislators can create laws, and anyone can create the rules that define some new game. When Shakespeare wrote, there were regularities but no rules about the spellings of English words. Later writers of English have created such rules. In contrast, on the reason-involving conception, there is normativity only when there are true or apparent normative reasons. We cannot create such reasons merely by getting people to accept some rule.

These conceptions may conflict. When there are such rules or requirements, we may have reasons to follow them. But these reasons are mostly provided, not by the mere existence or acceptance of these rules, but by certain other facts, most of which depend on some people's acceptance of these rules. If we drive on the correct side of the road, we shall be less likely to crash. If we use words with their correct spelling and meaning, that may make us seem better educated, and help us to be understood. When there are no such reason-giving facts, we may have no reason to follow some rule or requirement. We may have no reason, for example, to follow some fashion, or to refrain from violating some taboo. When I was told, as a child, that I shouldn't act in certain ways, and I asked why, it was infuriating to be told that such things are *not done*. That gave me no reason not to do these things.

Many of these claims do not apply to *moral* requirements. On some views, it is we who create these requirements. That is true, I believe, only in limited and often superficial ways. What we can create are only the particular forms that, in different communities, more fundamental, universal, and uncreated requirements take. For example, it is true everywhere that some people ought to care for those other people who cannot care for themselves, such as young children and those who are disabled by disease or old age. In most communities it is mostly close relatives who have such responsibilities. But that is not true everywhere.

There are also various uncreated rational requirements. For example, if we believe that we have decisive reasons to act in some way, we might be rationally required either to act in this way, or to give up this belief. And if we believe that some act is our only way to achieve some aim, we might be rationally required either to act in this way or to give up this aim.

Moral requirements often conflict with requirements of other kinds. We can be legally required, for example, to act wrongly. And many men have believed that, though it would be morally wrong to fight some duel, it would be dishonourable not to fight. Most of us would believe that, in these two kinds of case, moral requirements are more important. These requirements are often called *overriding*. But it would be trivial to claim that moral requirements are *morally* more important, or *morally* overriding. Legal requirements are *legally* overriding, and the code of honour is overriding in this code's terms. To be able to make significant claims

about the relative importance of these conflicting requirements, we need some impartial, neutral criterion.

Reasons provide such a criterion. We can compare the strengths of our reasons to follow these requirements. The men who fought duels had at most weak reasons to follow the code of honour, and they had strong moral reasons not to fight. And when we are legally required to act wrongly, we may have decisive moral reasons to break the law. Moral requirements may thus be more important in the reason-implying sense than the requirements of the code of honour, or the law.

It would be similarly trivial to claim that rational requirements are rationally overriding. So we should ask whether we have reasons to follow these requirements. It is a difficult question how much these requirements matter in the reason-implying sense. Following these requirements might be good, not in itself, but only as a means. And in appealing to claims about what matters in the reason-implying sense, we are not assuming that rationality matters.

We can next note one difference between moral and rational requirements. When we are deciding what to do, we often ought to ask whether any of our possible acts would be morally required, or wrong. But we need not ask which acts would be rational. That question arises only when we consider our own past acts, or the acts of others, and we ask whether these acts make us or others open to certain kinds of criticism. Compared with questions about what we ought to do or have reasons to do, questions about rationality are much less important. [144]

When we are deciding what to do, as I have said, we have two main questions:

Q1: What do I have most reason to do?

Q2: What ought I morally to do?

Of these questions, it is the question about reasons that is wider, and more fundamental. And if these questions often had conflicting answers, because we often had decisive reasons to act wrongly, that would undermine morality. For morality to matter, we must have reasons to care about morality, and to avoid acting wrongly. No such claim applies the other way round. If we had decisive reasons to act wrongly, the wrongness of these acts would not undermine these reasons.

These claims might be denied. When I claim that the wrongness of these acts would not undermine these reasons, I mean that we would still have these reasons. It might be similarly claimed that, even if we had decisive reasons to act wrongly, *morality* would not be undermined, since these acts would still be wrong.

This defence of morality would be weak. It could be similarly claimed that, even if we had no reasons to follow the code of honour, or the rules of etiquette, this code

and these rules would not be undermined. It would still be dishonourable not to fight some duels, and still be incorrect to eat peas with a spoon. But these claims, though true, would be trivial. If we had no reasons to do what is required by the code of honour, or by etiquette, these requirements would have no importance. If we had no reasons to care about morality, or to avoid acting wrongly, morality would similarly have no importance. That is how morality might be undermined.

It might next be objected that, in making these claims, I am appealing to the reason-involving criterion of importance. I am assuming that something is important only when and because we or others have reasons to care about this thing. But I have not defended this criterion. And like morality or the code of honour, the reason-involving criterion cannot support itself. Just as it would be trivial to claim that morality is *morally* important or that rationality is *rationally* important, it would be trivial to claim that reasons are important in the *reason-implying* sense.

As this objection rightly claims, we cannot show that reasons matter by appealing to claims about reasons. But justifications must end somewhere. And if reasons are fundamental, we should not expect that we could justify the reason-involving criterion of importance, by appealing to some other, deeper criterion.

Reasons *are*, I believe, fundamental. Something matters only if we or others have some reason to care about this thing. It would have great importance if morality did not in this sense matter, because we had no reason to care whether our acts were right or wrong. To explain and defend morality's importance, we can claim and try to show that we do have such reasons. Morality might have supreme importance in the reason-implying sense, since we might always have decisive reasons to do our duty, and to avoid acting wrongly. But if we defend morality's importance in this way, we must admit that the deepest question is not what we ought morally to do, but what we have sufficient or decisive reasons to do.

In the rest of these chapters, I shall mostly discuss morality. If reasons are more fundamental, as I have just claimed, it may seem that I should continue to discuss reasons. But we have sufficient reasons for turning to morality.

First, we can plausibly assume that we do have strong reasons to care about morality, and to avoid acting wrongly. In discussing morality, we shall in part be discussing these reasons. And these are among the reasons that most need discussing, because they raise some of the hardest questions.

Second, before we can judge the strength of our reasons to avoid acting wrongly, we must answer certain questions about which acts are wrong. One example is the question whether, as Act Consequentialists believe, we ought to sacrifice our life if we could thereby save the lives of several strangers. If that were true, we could

more plausibly claim that we might have sufficient or even decisive reasons to act wrongly. According to the overlapping sets of beliefs that most people accept, which Sidgwick calls *common sense morality*, we are morally permitted to give some kinds of strong priority to our own well-being. We might have no duty to sacrifice our life, however many strangers we could thereby save. If morality's requirements are in such ways much less demanding, it is less plausible to claim that we can have sufficient or decisive reasons to act wrongly.

There is another way in which, in discussing morality, we shall be discussing reasons. On several plausible moral principles or theories, whether some act is wrong depends on what, in certain actual or imagined situations, we or others would have most reason or sufficient reason to consent to, or agree to, or to want, or choose, or do. To know what these principles and theories imply, we must answer questions about reasons. That is like the way in which, to know about the nature and properties of atoms, we must answer questions about sub-atomic particles.

CHAPTER 7 MORAL CONCEPTS

21 Acting in Ignorance or with False Beliefs

Before we start to ask which acts are wrong, it will help to discuss what we mean by 'wrong', and what we are believing when we believe that some act is wrong. These questions are about *moral* senses of 'wrong', and the concepts that these senses express. We can ignore non-moral senses, such as the sense in which we might give the wrong answer to some question, or open some cereal packet at the wrong end.

It is often assumed that the word 'wrong' has only one moral sense. This assumption is most plausible when we are considering the acts of people who know all of the morally relevant facts. We can start by supposing that, when we think about such acts, we all use 'wrong' in the same sense, which we can call the *ordinary* sense. In many cases, however, we don't know all of the relevant facts, and we must act in ignorance, or with false beliefs. When we think about such cases, we can use 'wrong' in several partly different senses. Some of these senses we can define by using the ordinary sense. Some act of ours would be

> *wrong* in the *fact-relative* sense just when this act would be wrong in the ordinary sense if we knew all of the morally relevant facts,
>
> *wrong* in the *belief-relative* sense just when this act would be wrong in the ordinary sense if our beliefs about these facts were true,

and

> *wrong* in the *evidence-relative* sense just when this act would be wrong in the ordinary sense if we believed what the available evidence gives us decisive reasons to believe, and these beliefs were true. [145]

Acts are in these senses *right*, or at least *morally permitted*, when they are not wrong, and they are what we *ought morally* to do when all of their alternatives would be in these senses wrong.

Some writers claim or assume that, even when we are considering the acts of people who don't know all of the morally relevant facts, it is enough to ask which of these people's acts would be wrong, or were wrong, in the ordinary sense. Other writers claim that one of the senses I have just defined *is* the ordinary sense. [146] These claims

are, I believe, mistaken. We ought to use 'wrong' in all these senses. If we don't draw these distinctions, or we use only some of these senses, we shall fail to recognize some important truths, and we and others may needlessly disagree.

To illustrate these points, we can suppose that, as your doctor, I must choose between different ways of treating you. I am a bad doctor, since I have various unjustified beliefs about what, given the evidence, are the likely effects of different treatments. I also have some reasons to wish that you were dead. This story could continue in several ways. Suppose that, in

> *Case One*, I give you some treatment that I believe and hope will save your life, but which kills you, as it was almost certain to do,

and that, in

> *Case Two*, I give you some treatment that I believe and hope will kill you, but which saves your life, as it was almost certain to do.

According to some people, it is enough to use 'right' and 'wrong' in their belief-relative senses. On this view, it is enough to claim that I acted rightly in *Case One*, because I did what I believed would save your life, and that I acted wrongly in *Case Two*, because I did what I believed would kill you.

It is *not* enough to make these claims. We should also claim that, in *Case One*, I acted wrongly in the fact-relative and evidence-relative senses, since I killed you, as on the available evidence my act was almost certain to do. If I had asked some fully informed adviser what I ought to do, this person should not have told me that I ought to do what he or she knew would almost certainly kill you. We should similarly claim that, in *Case Two*, I acted rightly in the fact-relative and evidence-relative senses, since my act saved your life, as it was almost certain to do. I did what any fully informed adviser ought to have told me that I ought to do.

Suppose next that, though certain treatments nearly always cure people who have your particular disease, and certain other treatments would nearly always kill such people, your case is one of the rare exceptions. And suppose that, in

> *Case Three,* I give you some treatment that is almost certain to kill you, but which saves your life, as I hoped and unjustifiably believed that it would,

and that, in

> *Case Four,* I give you some treatment that is almost certain to save your life, but which kills you, as I hoped and unjustifiably believed that it would.

According to some people, it is enough to use 'right' and 'wrong' in their evidence-

relative senses. On this view, if some believer in sorcery tried to kill some enemy by sticking pins into a wax dummy, this person would not be acting wrongly. It is not wrong to stick pins into a wax dummy, since there is no evidence that such acts do any harm. And I acted rightly, in *Case Four*, when I gave you a treatment that, on the available evidence, was almost certain to save your life. But I acted wrongly in *Case Three* when I gave you a treatment that was almost certain to kill you.

As before, it is not enough to make these claims. We should not say only that I acted rightly, in *Case Four*, since my act was almost certain to save your life. We should also claim that I acted wrongly in the belief-relative and fact-relative senses, thereby murdering you. Murders should not be ignored.

Nor is it enough to say that, in *Case Three*, I acted wrongly by doing what was almost certain to kill you. We should also claim that I acted rightly in the fact-relative and belief-relative senses, since I intentionally saved your life. In failing to believe that my act would almost certainly kill you, I may be guilty of negligence, since I may have failed to read the recent medical journals, as I ought to have done. But it might instead be true that I conscientiously read these journals, and my mistake was only that I failed to believe what the evidence reported in these journals gave me decisive reasons to believe. Though I would then be at fault for medical incompetence, my failure to respond to these epistemic reasons would not be morally wrong.

According to some other people, it is enough to use 'right' and 'wrong' in their fact-relative senses. But suppose that, in

> *Case Five*, I give you some treatment that, as I justifiably believe, is almost certain to save your life, but which in fact kills you.

It is not enough to claim that, since I killed you, I acted wrongly. We should also claim that I acted rightly in the belief-relative and evidence-relative senses. It is morally important that I justifiably believed that my act was almost certain to save your life. Suppose instead that, in

> *Case Six*, I give you some treatment that, as I justifiably believe, will almost certainly kill you, but which in fact saves your life.

It is not enough to claim that, since I saved your life, I acted rightly. We should also claim that I acted wrongly in the belief-relative sense, because I believed that my act would kill you, as I intended it to do.

It would be possible to draw these distinctions without using these different senses of 'right' and 'wrong'. We might use only the evidence-relative senses. We might then claim that, though I did not act wrongly in *Case Four* when I murdered you, I had morally decisive reasons not to act in this way, and my act was blameworthy, giving me reasons for remorse and giving others reasons for indignation. Or we

might use only the belief-relative senses. We might then claim that, though I did not act wrongly in *Case One* when I tried to save your life, I had morally decisive reasons not to act in this way, because my act killed you, as I should have known that it was almost certain to do. Or we might use only the fact-relative senses. We might then claim that, though I did not act wrongly in *Case Six* when I saved your life, my act was blameworthy, because I was trying to kill you. But if we use 'wrong' in only one of these three senses, we may be misunderstood by those who use 'wrong' in only one or both of the other two senses. We and others may mistakenly believe that we are disagreeing. When we consider cases in which people do not know all of the morally relevant facts, there is no one sense of 'wrong' that everyone uses. So it is best to distinguish and use all these three senses.

We can next ask which of these senses are most important. As some of my claims have implied, that depends on which questions we are asking. We can start with questions about blameworthiness, which we can take to include questions about reasons for remorse and indignation. What is most important here is what, when acting, people believe. We should claim that

> (A) when some act is wrong in the *belief-relative* sense, because this act would be wrong if the agent's non-moral beliefs were true, this fact makes this act blameworthy.

In *Cases Two, Four*, and *Six*, for example, I act in ways that I believe will kill you. These acts would all be wrong if my beliefs were true, since killing you would be wrong. So (A) rightly implies that these acts were all blameworthy.

It might be similarly claimed that

> (B) when some act is wrong in the *fact-relative* sense, because this act would be wrong if the agent knew the relevant facts, this fact makes this act blameworthy.

But we ought to reject this claim. Remember that, in

> *Case Five*, I kill you by doing what I justifiably believe will save your life.

Since this act would be wrong if I knew that it would kill you, (B) implies that this act was blameworthy. But that is clearly false. When I learn that I have killed you, I shall be appalled. But since I justifiably believed that my act would save your life, this act was not blameworthy. And I have no reason for remorse, nor do others have any reason for indignation.

Here is a wider objection to (B). Suppose that, in

Case Seven, I save your life by doing what I justifiably believe will save your life.

It is clear that, in this case, my act was *not* blameworthy, since this act wasn't in any sense wrong. Though my act kills you in *Case Five* but saves your life in *Case Seven*, this difference is, from my point of view, entirely a matter of luck. In calling this difference a matter of *luck* from my point of view, I mean that I could not have known that one of these acts would kill you, and that this fact was in no way under my control. Though the difference between these cases is entirely a matter of luck, (B) implies that my act was blameworthy in *Case Five* but not in *Case Seven*. (B) therefore implies that

(C) an act's blameworthiness might entirely depend on luck.

When children are learning what it is for acts to be blameworthy, some of them have beliefs that assume or imply (C). Some of these children believe, for example, that well-intentioned acts are blameworthy when these acts have bad effects, even if these effects were wholly unpredictable. And some adults have had similar beliefs, such as the belief that we can inherit blameworthiness and guilt for the sins of our ancestors. These sins were not under our control. But when we understand blameworthiness better, we realize that (C) is false. Since (B) implies (C), we ought also to reject (B). When some act is wrong in the fact-relative sense, this fact does not make this act blameworthy.

There are two alternatives to (C). According to what we can call

the Kantian view, an act's blameworthiness cannot depend on luck.

According to

the semi-Kantian view, an act's blameworthiness cannot depend *entirely* on luck. But when two acts are blameworthy in some way that does not depend on luck, one of these acts may be *more* blameworthy in some way that *does* depend on luck.

This view is in itself less plausible than the Kantian view, since it is hard to see how blameworthiness might *partly* depend on luck. But this semi-Kantian view is sometimes claimed to have more plausible implications. [147] Return for example to

Case Two, in which I save your life by doing what I believe will kill you,

and

Case Four, in which I kill you by doing what I believe will kill you.

These acts are both wrong in the belief-relative sense, since if my beliefs were true

these acts would both kill you, as I intend them to do. In the *fact*-relative sense, however, my act is wrong only in *Case Four*. Though my act kills you in *Case Four* but saves your life in *Case Two*, this difference is, from my point of view, entirely a matter of luck. So, on the Kantian view, these acts are equally blameworthy. According to some semi-Kantians, that is not so. These people believe that

> (D) when acts are blameworthy because they are wrong in the belief-relative sense, these acts are more blameworthy if they are also wrong in the fact-relative sense.

On this view, though my attempts to kill you are both blameworthy, my act is more blameworthy in *Case Four*, because this attempt succeeds. Though attempted murder is blameworthy, murder deserves more blame, and gives me and others reasons for greater remorse and greater indignation.

Some semi-Kantians might also claim that

> (E) when acts are blameworthy because they are wrong in the belief-relative sense, these acts are more blameworthy if they are also wrong in the *evidence*-relative sense.

But remember that, in

> *Case Four*, I kill you by giving you a treatment that, on the evidence, was almost certain to save your life, but which I unjustifiably believed would kill you.

Suppose next that, in

> *Case Eight*, I kill you by giving you a treatment that I justifiably believed would kill you.

These acts are both wrong in the belief-relative and fact-relative senses, since they both kill you, as I believed they would. (E) implies that, in *Case Eight*, my act is more blameworthy, because this act is also wrong in the evidence-relative sense. We ought, I believe, to reject this claim. Murder can be plausibly regarded as more blameworthy than attempted murder. But we cannot plausibly regard murder as more blameworthy if and because the murderer's beliefs about the likely effects of his act were epistemically justified, because these beliefs were better supported by the available evidence. The most that we could claim is that, if potential murderers have such justified beliefs, these people are more dangerous, because their attempts to kill other people are more likely to succeed. That is not a difference in blameworthiness. [148]

On the Kantian view, all such attempts to kill are equally blameworthy, whether or not these acts succeed, or were likely to succeed. It is equally blameworthy to shoot

someone and hit, to shoot someone and miss, and to stick pins into a wax dummy believing irrationally that this way of killing someone will succeed. We cannot deserve less blame merely because we are either less successful in hitting our intended target, or are epistemically irrational.

This Kantian view is, I believe, true. Though murder can be plausibly regarded as more blameworthy than attempted murder, this claim's plausibility can be sufficiently explained, I believe, in other ways, some of which I mention in a note.

We can next define a fourth relevant sense of 'wrong'. Some act is

> *wrong* in the *moral-belief-relative* sense just when the agent believes this act to be wrong in the ordinary sense.

On one fairly plausible view, which we can call

> *the Thomist View*, when people believe that they are acting wrongly, that is enough to make their act wrong, even if this act would not otherwise be wrong.

Suppose, for example, that it would not be wrong to use artificial contraceptives, or to perform an early abortion, or to help someone to die in a swifter, better way. On this Thomist view, such acts *would* be wrong if they were done by people who mistakenly believed them to be wrong. As Thomists add, however, when people believe that some act would be *right*, that is *not* enough to make this act right. Conscientious SS officers often acted wrongly, even when they believed their acts to be right, or to be their duty.

Even if we reject this view, it seems clear that

> (F) in most cases, when someone acts in some way that this person believes to be wrong, that makes this act blameworthy.

Of the facts that can make acts blameworthy, this may be the most important. In some cases, however, people do what they believe to be wrong because they are half-aware that their act is not wrong, but morally required. One example may be Huckleberry Finn when he helped a runaway slave to escape. [149] Some such acts may not be blameworthy. But in most cases, an act's blameworthiness depends on whether this act is wrong in the belief-relative and moral-belief-relative senses.

We can next ask which are the most important senses of 'ought', 'right', and 'wrong' when we are trying to decide what to do. In the cases that we have been discussing, and many others, the rightness of our acts depends on the goodness of their effects or possible effects. It is often assumed that

> (G) in such cases, we ought to try to act in the way that would be right in the fact-relative sense, because this act would make things go best.

In my medical examples, (G) has acceptable implications. In trying to do what would save your life, I would be trying to do what would make things go best. But in many other cases (G) is false. Consider

> *Mine Shafts:* A hundred miners are trapped underground, with flood waters rising. We are rescuers on the surface who are trying to save these men. We know that all of these men are in one of two mine shafts, but we don't know which. There are three flood-gates that we could close by remote control. The results would be these:

		The miners are in	
		Shaft A	Shaft B
	Gate 1	We save 100 lives	We save no lives
We close	Gate 2	We save no lives	We save 100 lives
	Gate 3	We save 90 lives	We save 90 lives

Suppose next that on the evidence available and as we believe, it is equally likely that the miners are all in Shaft A or all in Shaft B. If we closed either Gate 1 or Gate 2, we would have a one in two chance of doing what would be right in the fact-relative sense, because our act would save all of these hundred people. If we closed Gate 3, we would have *no* chance of doing what would be in this sense right. But this is clearly what we ought to do, since by closing Gate 3 we shall be certain to save ninety of these people.

When I claim that we ought to close Gate 3, I am using 'ought' in the ordinary sense. This act is also what we ought to do in the more precise belief-relative and evidence-relative senses, since the hundred miners *are*, as we justifiably believe, equally likely to be in either shaft. Since it would be wrong for us to try to act rightly in the fact-relative sense by closing either of the other gates, we ought to reject claim (G). On a rough statement of the true view, which we can call

> *Expectabilism*: When the rightness of some act depends on the goodness of this act's effects or possible effects, we ought to act, or try to act, in the way whose outcome would be *expectably-best*. [150]

In calling some act's outcome 'expectably-best', we do *not* mean that we expect this

act to produce the best outcome. In this example, the outcome would be expectably-best if we closed Gate 3, though this act would be certain *not* to produce the best outcome, as our act might do if instead we closed one of the other gates. To decide which of our possible acts would make things go *expectably-best*, we take into account both how good the effects of the different possible acts might be, and the probabilities, given our beliefs or the available evidence, that these acts would have these effects. When what matters is only the number of lives that are saved, some act's outcome would be expectably-best if this is the act that would save the greatest *expectable number* of lives. The expectable number that some act would save is the number of lives that this act might save, multiplied by the chance that this act would save these lives. In *Mine Shafts,* for example, if we closed either Gate 1 or Gate 2, the expectable number of lives saved would be 100 multiplied by a chance of one in two, or by 0.5. This number would be 50. If we closed Gate 3, this expectable number would be 90, since this act would be certain to save 90 lives.

We can similarly claim that, whenever we don't know what effects our acts would have, the expectable goodness of some act's effects is, roughly, the goodness of these possible effects multiplied by the chance that this act would have these effects. [151] Expectabilism applies to all cases, including those in which we know which act would in fact make things go best. This act's outcome would be expectably-best.

I have just rejected the view that, when we don't know what effects our acts would have, we ought to try to do what would in fact make things go best. It is sometimes claimed that, if we reject this view, we cannot explain why we ought, in many cases, to try to discover more of the facts, so that we can make better informed decisions. But this claim is mistaken. We ought to try to get more information whenever acting in this way would itself make things go expectably-best. In important cases, that is often true. In *Mineshafts*, if we could easily find out where the miners are, trying to find that out would make things go expectably-best, since we would then be very likely to save all these people.

There is another reason why, when we are trying to decide what to do, we can ignore the fact-relative senses of 'ought', 'right', and 'wrong'. We cannot try to do what is right in the fact-relative *rather than* the belief-relative sense. Suppose I believe that, to save your life, I must act in a certain way. Though I know that my belief might be false, I cannot try to do what *would in fact* save your life rather than doing what *I now believe* would save your life, since what *I now believe* is that acting in this way *would in fact* save your life. We cannot base our decisions on the facts except by basing our decisions on what we now believe to be the facts. In the same way, as Sidgwick points out, though we know that our moral beliefs may be mistaken, we cannot try to do what is really right rather than what, at the time of acting, we believe to be right. [152]

I claimed earlier that, when we ask whether some act was blameworthy, or whether the agent has reasons for remorse and others have reasons for indignation, what is most important is whether this act was wrong in the belief-relative and moral-belief-relative senses. I have just claimed that, when we are choosing between different possible acts, we need not ask what we ought to do in the fact-relative sense. And when the rightness of our acts depends on the goodness of their effects, we ought to try to do, not what would in fact make things go best, but what on the evidence, or given our beliefs, would make things go expectably-best. These claims may seem to imply that it has little importance which acts are right or wrong in the fact-relative senses.

There is, however, one way in which these fact-relative senses can be claimed to be fundamental. As well as asking, in some actual case, whether some act would be wrong, we can ask wider questions about which moral beliefs are true, and which moral principles or theories we ought to accept and try to follow. We ought to try to answer some of these questions, or at least to think about some other people's answers. Though we cannot try to do what is really right rather than what we now believe to be right, we ought to try to have true moral beliefs, since we shall then be less likely to act wrongly.

In trying to answer such questions, it is best to proceed in two stages. We can first ask which acts would be wrong if we knew all of the morally relevant facts. These are questions about which acts would be wrong, in such cases, in what I have called the ordinary sense. But these are also questions about which acts would be wrong in the fact-relative sense. Acts are in this sense wrong when these acts would be wrong in the ordinary sense if we knew all of the relevant facts.

After answering these questions, we can turn to questions about what we ought morally to do when we don't know all of the relevant facts. These questions are quite different, since they are about how we ought to respond to risks, and to uncertainty. As in the case of non-moral decisions, though these questions have great practical importance, they are less fundamental. These are not the questions about which different people, and different moral theories, most deeply disagree. Given the difference between these two sets of questions, they are best discussed separately. So I shall often suppose that, in my imagined cases, everyone would know all of the relevant facts. We can then ask what we ought to do in the simplest, fact-relative sense. In many other cases these distinctions do not matter, so I shall often use 'best' to mean 'best or expectably-best'.

There is much more to be said about the relations between these and some other similar senses of 'ought' and 'wrong'. There are difficult questions, for example, about when and how people who have different beliefs, or are aware of different evidence, can disagree about what someone ought to do. My aim has been only to argue that we need to distinguish these senses, and to decide which senses are most

relevant to the kind of moral question we are asking.

We can next return briefly to questions about what, in the non-moral senses, we should do, or ought to do. These are questions, we can now say, about what we *ought practically* to do. We can call some possible act

> what we *ought practically* to do in the *fact-relative* sense just when and because this act is what we have decisive reasons, or most reason, to do.

This fact-relative sense of 'ought' is what I am calling the *decisive-reason-implying* sense. When we are considering cases in which people know all of the relevant, reason-giving facts, it may be enough to use this sense of 'ought'. In many cases, however, people do not know, or have false beliefs about, these relevant facts. In such cases, we can call some act

> what we *ought practically* to do in the *evidence-relative* sense just when this act would be what we had decisive reasons to do, if we believed what the available evidence gives us decisive reasons to believe, and these beliefs were true.

We can similarly call some act

> what we *ought practically* to do in the *belief-relative* sense just when this act would be what we had decisive reasons to do, if our beliefs about these facts were true.

We can also call some act

> what we *ought practically* to do in the *normative-belief-relative* sense just when this act is what we believe that we ought practically to do, or what we believe that we have decisive reasons to do.

As well as asking what we ought to do in these four senses, we can ask which acts are *rational*. We ought, I have claimed, to use the words 'rational' and 'irrational' to express certain kinds of praise or criticism. Questions about rationality are, in several ways, like questions about blameworthiness. The answers depend, for similar reasons, on the agent's beliefs. On the view that I defended earlier,

> (H) we ought rationally to act in some way when this act is what we ought practically to do in the belief-relative or normative-belief-relative senses.

In the case of the angry snake, for example, you ought rationally to run away, given your false belief that this act would save your life. In some cases, some act might be rational relative to our beliefs about the reason-giving facts, but irrational relative

to our normative beliefs, or *vice versa*.

According to some writers, whether we ought rationally to act in some way depends *only* on our normative beliefs, and our acts are irrational only if we are either failing to respond to what we believe to be decisive reasons, or failing to do what we believe that we ought to do. This is like the view that acts are blameworthy only if the agent believes them to be wrong. Such views are, I have claimed, too narrow. Acts can be blameworthy even if the agent believes them to be right, as in the case of the conscientious SS officer. We should similarly claim that, if we are aware of facts that give us what are clearly and strongly decisive reasons to act in some way, we ought rationally to act in this way even if we fail to believe that these facts give us such reasons. Similar claims apply to our desires and aims. When Scarlet prefers agony on next Tuesday to mild pain on any other day, his preference is irrational even though he is not failing to respond to what he believes to be a reason.

22 Other Kinds of Wrongness

We should distinguish, I have just claimed, between several moral senses of 'ought', 'right', and 'wrong'. I defined these senses by using a single sense, which I have called the *ordinary* sense. We can now ask whether we can explain this ordinary sense, and whether there is more than one such sense.

It can be unclear, or indeterminate, what we should claim to be part of the meaning of some word. It is unclear, for example, whether it is part of the meaning of the word 'cheetah' that cheetahs are hunters and have claws, or part of the meaning of 'war' that wars have to be declared. If we decide to include more in our accounts of the meaning of our words, we shall more often claim that some word has several senses. We might, for example, claim that the word 'war' has two senses, one of which applies only to wars that have been declared. I have already distinguished several senses of 'wrong', and I shall now distinguish several others. On a different account, to which I shall return, there is only one moral sense of 'wrong'. It is worth considering both accounts, but we need not choose between them.

Though I shall discuss the English word 'wrong', our questions are about the *concept* wrong, which is what is meant by this English word, and by words in other languages with sufficiently similar meanings. This concept refers to the *property* of wrongness. (When we claim that some word, phrase, or concept *refers to some property*, we are not thereby claiming that anything *has* this property. There are many properties that nothing has, such as the properties of being a Greek god, or a witch.) If there are different senses of 'wrong', these senses express different versions of the concept *wrong*, which refer to different kinds of wrongness.

Like the concept of *a reason*, and the decisive-reason-implying concepts *should* and *ought*, one version of the concept *wrong* is indefinable, in the sense that it cannot be helpfully explained in other terms. We can use this concept to define some other moral concepts. We can say that some act is

> *right,* or *morally permitted,* when this act would not be wrong,

and that some act is

> *our duty*, *morally required*, or what we *ought morally* to do, when it would be wrong for us *not* to act in this way.

We might instead define this version of the concept *wrong* by appealing to an undefined version of one of these other concepts. Some act would be wrong, we might say, when we ought not to act in this way. But though we can explain how these concepts are related to each other, this group of concepts all have a common element which we cannot helpfully explain merely by using words. Like the concept of *a reason*, and the decisive-reason-implying concepts *should* and *ought*, we must explain this concept in other ways, by getting people to think certain thoughts. To express this indefinable version of the concept *wrong*, I shall use the phrase '*mustn't-be-done*'. [153]

These moral concepts, I shall assume, also have other, definable versions. For example:

> In the *blameworthiness* sense, 'wrong' means 'blameworthy'.
>
> In the *reactive-attitude* sense, 'wrong' means 'an act of a kind that gives its agent reasons to feel remorse or guilt, and gives others reasons for indignation and resentment'.
>
> In the *justifiabilist* sense, 'wrong' means 'could not be justified to others'.
>
> In the *divine command* sense, 'wrong' means 'forbidden by God'.

These senses can be combined to form more complex senses. For example, when we claim that some act is wrong, we might mean that this act is blameworthy because such acts are unjustifiable to others. Or we might mean that this act mustn't-be-done because such acts are forbidden by God.

Some people use 'ought morally' and 'wrong' in reason-implying senses. In what we can call the *decisive-reason* senses,

> 'what we ought morally to do' means 'what we have decisive reasons to do',

and

'wrong' means 'what we have decisive reasons *not* to do'.

These senses are misleading, and should not be used. We often believe that we have decisive reasons to act in some way, though we do not believe that we ought morally to act in this way. And if Rational Egoists used these decisive-reason senses, they would claim that

(I) we ought morally to do whatever would be best for ourselves.

But Rational Egoism is best regarded, not as a moral view, but as an external rival to morality. On this view, we always have decisive reasons to do whatever would be best for ourselves, whether or not these acts would be morally wrong. [154]

In what we can call the *decisive-moral-reason* senses,

'what we ought morally to do' means 'what we have decisive *moral* reasons to do',

and

'wrong' means 'what we have such reasons not to do'.

These senses do not, I believe, have much importance. We already have the concept of what we have decisive reasons to do, and it adds little to claim that some of these reasons are moral reasons. It is also unclear which reasons should be called 'moral'. It is unclear, for example, whether our reasons to promote the well-being of others should all be called moral reasons. Whether we ought morally to act in some way cannot be helpfully claimed to depend on how we ought to answer such partly verbal questions.

In what we can call the *morally-decisive-reason* senses,

'what we ought morally to do' means 'what we have morally decisive reasons to do',

and

'wrong' means 'what we have such reasons not to do'.

Though these senses may seem very similar to the *decisive-moral-reason* senses, there are two important differences. First, when we ask whether we have morally decisive reasons to act in some way, we are not asking whether we have decisive reasons of the kind that should be called 'moral'. We are asking whether we have reasons to act in this way that *morally outweigh* any reasons that we may have not to act in this way. Second, to be able to state our moral beliefs by using 'wrong' in the decisive-moral-reason sense, we must believe that we always have decisive reasons

not to act wrongly. But if we claim instead that we have *morally* decisive reasons not to act in some way, that leaves it open whether these reasons are also *non-morally* decisive, or decisive *all things considered*. We could use 'wrong' in this sense even if we believed that, in some cases, we might have sufficient or decisive reasons to act wrongly.

Some people seem to use

> 'what we ought morally to do' to mean 'what we have the strongest impartial reasons to do'.

Some act is in this sense wrong when we have stronger impartial reasons to do something else. We can call these the *impartial-reason-implying* senses of 'ought' and 'wrong'. There are, as I have said, similar senses of 'good', 'bad', and 'best'. According to some Act Consequentialists:

> We ought always to do whatever would make things go best.

If this claim uses both 'ought' and 'best' in these impartial-reason-implying senses, it would mean

> (J) What we have the strongest impartial reasons to do is whatever would make things go in the way in which we all have the strongest impartial reasons to want things to go.

We can call this view *Impartial-Reason Act Consequentialism*. To express this sense of 'ought', we can use the phrase *ought-impartially*.

This sense of 'ought' differs significantly from more familiar moral senses. Sidgwick, for example, writes:

> the good of any one individual is of no more importance, from the point of view. . . of the Universe, than the good of any other. . . And. . . as a rational being I am bound to aim at good generally. . . not merely at a particular part of it. . . I ought not to prefer my own lesser good to the greater good of another.[155]

When Sidgwick claims that he *ought* not to prefer his own lesser good, he does not seem to mean that such a preference would be blameworthy, or unjustifiable to others, or that such an act would give him reasons for remorse and give others reasons for indignation. Sidgwick seems to mean only that, when assessed from an impartial point of view, his reason to give himself some lesser good is weaker than, or outweighed by, his reason to give some greater good to someone else.

This kind of Consequentialism may be better regarded, not as a moral view, but as

being, like Rational Egoism, an external rival to morality. Given this view's claim that we ought to sacrifice our lesser good for the greater good of others, it is much closer to morality. That makes this view, in some ways, a more serious rival. Impartial-Reason Act Consequentialism may be accepted by many people who would reject Rational Egoism, because they regard their own well-being as what Sidgwick calls a 'narrow' and 'ignoble end'. [156]

(J) may seem to be a trivial claim, which is close to a tautology. It is not, however, trivial to claim that acts can be right or wrong, and outcomes can be good or bad, in these impartial-reason-implying senses. On some widely accepted views about reasons, as I have claimed, there are no such acts or outcomes. And even if (J) were a tautology, Impartial-Reason Act Consequentialists could make other, substantive claims. If they are Hedonistic Utilitarians, for example, these people might claim

> (K) What we ought-impartially to do is whatever would produce the greatest sum of happiness minus suffering. [157]

These people may believe that we all have strong reasons to act in this way. And they might not act upon, or even have, moral beliefs that involve any of the more familiar senses of 'ought morally' and 'wrong'. These people may be convinced that it matters greatly how well things go, and they may be strongly motivated and often moved to act in ways that prevent or relieve suffering. But they may be doubtful whether any acts are duties, or mustn't-be-done, and doubtful about blameworthiness, and about reasons for remorse and indignation. That is one way in which this form of Consequentialism might be an external rival to morality. [158]

According to some writers, as I have said, there is only a single moral sense of 'wrong', 'right', and 'ought'. It would be implausible to make this claim about one of the definable senses. If we can use 'wrong' in one definable sense, we can surely use it in others. Nor is there any one definable sense that can be plausibly claimed to be the only sense that everyone uses. We cannot even claim that everyone uses 'wrong' to mean 'what we have morally decisive reasons not to do', since some people never or seldom use the concept of a reason.

It would be more plausible to claim that everyone uses 'wrong' in the indefinable sense that I am expressing with the phrase 'mustn't-be-done'. The blameworthiness and reactive-attitude senses might be claimed to appeal implicitly to this indefinable sense, because the attitudes of blame, guilt, remorse, and indignation all involve the belief that some act is wrong. In defining the morally-decisive-reason sense of 'wrong', we might have to use the word 'morally' indefinably. And some other definable senses might be claimed to express, not the belief that certain acts are wrong, but certain other beliefs about wrong acts. The divine command and justifiabilist senses might, for example, express the beliefs that

acts are wrong, in the sense that they mustn't-be-done, when and because these acts are forbidden by God, or unjustifiable to others. [159]

When some writers claim that words like 'wrong' and 'ought' have only one moral sense, they appeal to the fact that, even when we and other people have very different moral views, we regard ourselves as *disagreeing* with these other people. If we and others used these words in different senses, these writers claim, we could not be disagreeing with these other people, since we wouldn't be discussing the same questions.

This argument is weak. Different people may use 'wrong' or 'ought' in different definable senses that partly overlap. That may be enough to make disagreement possible. Suppose for example that, when I claim that some act is wrong, I mean that such acts are blameworthy because they are forbidden by God. When you claim that some act is wrong, you mean that such acts are blameworthy because they are unjustifiable to others. If I claimed that some act was wrong and you claimed that it wasn't, we would be disagreeing about whether this act was blameworthy. And when people use 'wrong' in such different senses, that may *increase* their disagreements. In the case just imagined, if we understood each other's use of 'wrong', you might believe that no acts are in my sense wrong, since you believe that no acts are blameworthy because they are forbidden by God. I might believe that no acts are in your sense wrong, since I believe that no acts are blameworthy because they are unjustifiable to others. We would then completely disagree, since each of us would reject all of the other's moral beliefs.

When different people in the same community use words like 'wrong' or 'ought' in such different, partly overlapping senses, these people have reasons to move to other, thinner senses, which they can all use. It would then be clearer when these people disagree, and what they are disagreeing about. In the case just imagined, if you and I both used 'wrong' to mean 'blameworthy', we would be able to agree that many acts are in this sense wrong, even though we disagreed about what makes these acts wrong.

In some cases, we can add, those who use 'wrong' or 'ought' in different senses may *not* be disagreeing. On Sidgwick's view, for example, I ought to give up my life if I could thereby save the lives of two strangers who are relevantly like me. If Sidgwick were using 'ought' in the blameworthiness or reactive-attitude or senses, most of us would reject this claim. We would believe that, if I saved myself rather than these two strangers, my act would not be blameworthy, and I would have no reason to feel remorse, nor would these strangers or others have any reasons to be indignant. But Sidgwick might mean only that I would have stronger impartial reasons to save the two strangers. That claim would not conflict with other people's moral beliefs.

Consider next those cases in which the rightness of our acts depends on the goodness of their effects. In such cases, some people claim that

(L) we ought to do what would make things go best,

and others claim that

(M) we ought to do what would make things go expectably-best.

If (L) uses 'ought' in the fact-relative sense, and (M) uses 'ought' in the evidence-relative sense, these claims do not conflict, and we could accept them both. Nor would either claim conflict with a version of (M) that used 'ought' and 'expectably-best' in belief-relative senses.

There is another avoidable disagreement. According to some writers, we ought to do certain things, such as keeping our promises, saving people's lives, and doing what would make things go expectably-best. According to some other writers, we ought to *try* to do these things. We ought, I believe, to make both these claims. We should not claim only that we ought to *do* these things, since it is morally important whether we tried to do them. We may deserve no blame, for example, if we tried but failed to keep some promise, or to save someone's life. Nor should we claim only that we ought to *try* to do certain things, since it is often morally important whether our acts succeed. If our attempt to keep some promise fails, for example, it may be true that we ought to act in some other way instead. When we claim that we ought to do something, we should often be taken to mean that we ought to do this thing or at least try to do it.

It is unimportant whether the various senses that I have described should be called different senses of 'wrong', which refer to different kinds of wrongness. It is enough to distinguish these senses, and the concepts that they express. We can then decide which of these concepts are most worth using.

In making that decision, we can return to the question of how much morality matters in the reason-implying sense. If some possible act would be wrong, does this fact give us a reason not to do it? If so, how strong are such reasons?

The answers depend in part on what we mean by 'wrong', and on the kind of wrongness to which our use of 'wrong' refers. Suppose first that, in claiming that some act is wrong, we mean that we have decisive moral reasons not to act in this way. These reasons would be provided by the facts that made some act wrong. Two examples might be the facts that some act would be a lie or would cause pointless suffering. On this view, the fact that

(N) some act is wrong

would be the higher-order fact that

(O) there are certain other facts that give us decisive moral reasons not to act in this way.

This higher-order fact would not give us a *further*, *independent* reason not to act in this way. Though we might claim that an act's wrongness always gives us a reason not to do it, this reason would be *derivative*, since its normative force would derive entirely from these other reason-giving facts. So if we used 'wrong' only in this decisive-moral-reason sense, we could claim that

(P) when some act would be wrong, this fact would not give us any further reason not to act in this way.

On this view, it would have no practical importance whether some act would be wrong. When we were trying to decide what to do, it would always be enough to ask whether we had decisive reasons for or against acting in any of the possible ways. If we decided that we had such reasons, we could then ask whether these were *moral* reasons, so that our act was wrong in the decisive-moral-reason sense. But this would not be a question about what we ought to do, or had reasons to do. This question would be merely conceptual, like the question of which are the kinds of reason that can best be called legal, or aesthetic. [160] So we have little reason, I believe, to use this sense of 'wrong'.

Many people assume that an act's wrongness does give us strong or even decisive further reasons not to do it. If these people use 'wrong' in the decisive-moral-reason sense, their assumption would be false, in the way that I have just described. That does not show that these people cannot be using 'wrong' in this sense, since these people may not have seen the point that I have just made. But most of us, I believe, use 'wrong' in one or more other senses. And when certain acts would be wrong in these other senses, we *can* claim that the wrongness of these acts gives us further, independent reasons not to act in these ways.

Suppose first that we use 'wrong' in the indefinable sense. When we claim that some act is in this sense wrong, we are not claiming that this act has what Scanlon calls the 'purely formal, higher-order property' of having other, reason-giving properties. [161] We are claiming that this act has the highly distinctive substantive property of being something that *mustn't-be-done*. Though I believe strongly that some acts are in several other senses wrong, it seems to me a more open question whether any acts have this indefinable property. But if they do, we could plausibly claim that, when some act mustn't-be-done, that gives us a very strong reason not to do it. This is one of the senses of 'wrong' with which it seems most plausible to

claim that

> (Q) when some act would be wrong, this fact always gives us a decisive reason not to do it.

(Q) would be just as plausible, though for significantly different reasons, if we used 'wrong' to mean 'forbidden by God'.

If we use 'wrong' in the other definable senses, we could similarly claim that an act's wrongness gives us independent reasons not to do it. When some act would be blameworthy, unjustifiable to others, and is an act that would give us reasons for remorse and give others reasons for indignation, these facts would all give us further reasons not to act in this way. We should not, however, claim that these facts would always give us our *strongest* reasons not to act wrongly. If some act would cause great suffering, for example, that might give us a much stronger reason than the reasons given by the facts that this act would be blameworthy and unjustifiable to others.

As I have said, we need not choose between these senses of 'wrong', and the concepts that they express. It is worth using several of these concepts, asking, for example, which acts are wrong in the indefinable, justifiabilist, reactive-attitude, or blameworthiness senses. In the rest of this book I shall use 'ought morally' and 'wrong' vaguely, in some combination of these senses.

There are some deep and difficult questions about how we should understand these normative concepts, and about whether acts can have the properties to which these concepts refer. Except in Part Five, I shall say little about these *meta-ethical* questions. Such questions will be easier to answer when we have made more progress in our thinking about practical and epistemic reasons, and about morality. As Rawls and Nagel claim, our moral theories 'are primitive, and have grave defects', and 'ethical theory. . . is in its infancy.' [162]

Rather than proposing a new moral theory, I shall try to learn from some existing theories, hoping to get somewhat closer to the truth. I shall start with Kant, because he is the greatest moral philosopher since the ancient Greeks. When Kant presents his famous formulas, his aim, he writes, is to find 'the supreme principle of morality'. [163] I shall ask whether he succeeds.

PART TWO PRINCIPLES

CHAPTER 8 POSSIBLE CONSENT

23 Coercion and Deception

According to Kant's best-loved principle, often called

> *the Formula of Humanity*: We must treat all rational beings, or persons, never merely as a means, but always as ends. [164]

To treat people as ends, Kant claims, we must never treat them in ways to which they could not consent. In explaining the wrongness of a lying promise, for example, Kant writes

> he whom I want to use for my own purposes with such a promise cannot possibly agree to my way of treating him.[165]

Korsgaard comments:

> People cannot assent to a way of acting when they are given no chance to do so. The most obvious instance of this is when coercion is used. But it is also true of deception. . . knowledge of what is going on and some power over the proceedings are the conditions of possible assent. [166]

Onora O'Neill similarly writes:

> if we coerce or deceive others, their dissent, and so their genuine consent, is in principle ruled out. [167]

Korsgaard concludes:

> According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrong-doing to others. [168]

These remarks suggest this argument:

> It is wrong to treat people in any way to which they cannot consent.

People cannot consent to being coerced or deceived.

Therefore

Coercion and deception are always wrong.

It is sometimes right, however, to treat people in ways to which they cannot consent. When people are unconscious, for example, they cannot consent to life-saving surgery, but that does not make such surgery wrong.

Kant's claim, Korsgaard might say, applies only to acts whose nature makes consent impossible. Deception, unlike surgery, is such an act. For people to be able to consent to our way of treating them, they must know what we are doing. If people knew that we were trying to deceive them, we would be unable to deceive them. So we cannot possibly deceive people with their consent. This might be why, unlike surgery, deception is always wrong. [169]

But consider

> *Fatal Belief*: I know that, unless I tell you some lie, you will believe truly that *Brown* committed some murder. Since you could not conceal that belief from Brown, he would then murder you as well.

If I say nothing, you could reasonably complain with your dying breath that I ought to have saved your life by deceiving you. I could not defensibly reply that, since I could not have deceived you with your consent, this way of saving your life would have been wrong. My life-saving lie *would* be like life-saving surgery on some unconscious person. Just as this person would consent to this surgery if she could, you would consent to my deceiving you. It is a merely technical problem that, if I asked you for your consent, that would make my deceiving you impossible. We could solve this problem if you had the ability to make yourself lose particular memories. After you had given your consent, you could deliberately forget our conversation, so that my lie could save your life. Since you would consent to my deceiving you if you could, my lie would be morally as innocent as some lie that was needed to give someone a surprise party.

Similar remarks apply to coercion. People could not consent to being coerced, it might be claimed, because if people gave consent they would not be being coerced, and if they were being coerced they could not freely give consent. But we can freely consent to being later coerced in some way. Before the discovery of anaesthetics, many people freely consented to being later coerced during painful surgery. And we can freely consent to some kinds of coercion even while we are being coerced. Most of us would vote in favour of everyone's continuing to be legally coerced, by threats of punishment, to pay fair taxes and obey good laws. I would consent to being coerced to be less untidy. Though deception and coercion

are often wrong, what makes them wrong is not, I believe, the fact that these are acts whose nature makes consent impossible.

24 The Consent Principle

Return now to Kant's claim that

> (A) it is wrong to treat people in any way to which they cannot possibly consent. [170]

People cannot consent, Korsgaard writes, 'when they are given no chance to do so.' O'Neill similarly writes, 'To treat others as persons we must allow them the *possibility* either of consenting to or of dissenting from what is proposed'. [171] These remarks assume that Kant means

> (B) It is wrong to treat people in any way to which they cannot possibly consent because we have not given them the possibility of giving or refusing consent.

When we treat people in some way, they can often give or refuse consent in a *declarative* sense, by telling us or others that they do or don't consent. Korsgaard and O'Neill use 'consent' in a different and more important sense. People can give or refuse consent in this *act-affecting* sense if they have what Korsgaard calls 'power over the proceedings', because they will be treated in some way only if they consent. So we can restate (B) as

> *the Choice-Giving Principle*: It is wrong not to give other people the power to choose how we treat them. [172]

If this were what Kant meant, we would have to reject Kant's claim, since the Choice-Giving Principle has implications that are clearly false. This principle mistakenly implies, for example, that we ought to let other people choose whether or not we give their student essays low grades, buy what they are trying to sell us, take back what they stole from us, report their crimes, or vote against them in some election. In most morally important cases, moreover, our choice between different possible acts would have significant effects on two or more people. We could not give to more than one of these people the power to choose how we shall act, as would be shown if two of these people made conflicting choices. So the Choice-Giving Principle also mistakenly implies that, in all these cases, whatever we did would be wrong.

There is, I believe, a better way to interpret Kant's remarks. Korsgaard and O'Neill assume that, when Kant claims

> (A) It is wrong to treat people in any way to which they cannot possibly consent,

he means

> (C) It is wrong to treat people in any way to which they cannot consent in the act-affecting sense because we have not given them the power to choose how we treat them.

I suggest that Kant means

> (D) It is wrong to treat people in any way to which they *could not* consent in the act-affecting sense, *if* we gave them the power to choose how we treat them.

It might be objected that, if we gave people this power, they *could* choose that we act in any of the possible ways, so there would never be any act to which these people could not consent. If this were the kind of impossibility that Kant had in mind, (D) would be trivial, since (D) would never imply that some act is wrong. But there is another kind of impossibility. When people say 'I cannot possibly consent to your proposal', they hardly ever mean that giving consent is not one of the choices that is open to them. These people often mean that they could not *rationally consent*, because they have decisive reasons to refuse consent. Kant, I suggest, means

> (E) It is wrong to treat anyone in any way to which this person could not rationally consent.

I shall call this the *Consent Principle*. [173]

We have several reasons to believe that Kant is appealing to this principle. While the Choice-Giving Principle is obviously false, the Consent Principle might be true, which makes it more likely to be what Kant means. When Kant claims that we could not do something, he often means that we could not rationally do this thing. [174] Kant also writes that, if he treated someone wrongly, this person

> could not possibly agree to my way of treating him, *and so himself contain the end of this act*. [175]

If Kant were claiming that we ought to let other people choose how we treat them, he would have no reason to add that, for our treatment of someone to be justified, this person must be able to 'contain the end of this act', by sharing this act's aim. When we let other people choose how we shall treat them, we are not acting with some aim that these people might be unable to share. Kant must mean that, when *we* are choosing how we shall treat other people, we ought always to act with some aim that these people would be able to share. Nor would it be enough if these

people could *conceivably* share our aim, since many unjustifiable aims could conceivably be shared. We ought to act only with some aim that other people could *rationally* share, so that they could rationally consent to our way of treating them.

Kant's remark about shared ends or aims, though helping to explain his claims about consent, also adds a less plausible idea. Even if other people could rationally share our aim, we may be acting wrongly if and because these people could not rationally consent to our way of achieving this aim. Though you could rationally share my aim that my tame tiger be fed, you could not rationally consent to being what my tiger eats. And even if other people could *not* rationally share our aim, we may not be acting wrongly if these people could rationally consent to our act. Though you could not rationally share my aim of reciting someone's name a thousand times, you could rationally consent to my reciting your name. So, compared with the question whether other people could rationally share our aims, it is more important whether these people could rationally consent to our acts.

Kant's claims about consent give us an inspiring ideal of how, as rational beings, we ought all to be related to each other. It is worth asking whether we could achieve this ideal. We cannot always let everyone choose how we treat them. But we might be able to treat everyone only in ways to which they could rationally consent. And if that is possible, Kant may be right to claim that this is how everyone ought always to act.

25 Reasons to Give Consent

Whether we could achieve Kant's ideal depends on which are the acts to which people could rationally consent. Rawls suggests that, in proposing the Consent Principle, Kant assumes that

> (F) people could rationally consent to some act if and only if, or *just when*, they could will it to be true that the agent's maxim is a universal law. [176]

Rawls is referring here to another of Kant's proposed statements of the supreme principle of morality. According to Kant's

> *Formula of Universal Law*: It is wrong to act on any maxim that we could not will to be a universal law.

By our *maxims* Kant means, roughly, our policies and underlying aims. We need not yet consider in what sense maxims might be universal laws.

Kant does not, however, commit himself to (F). And this assumption would be a mistake. Suppose that I am your doctor, and I ask you whether you consent to my

giving you some medical treatment. For it to be rational for you to consent, you would need to have beliefs about whether I am a well-qualified and conscientious doctor, and about what effects this and the other possible treatments would be likely to have. But you wouldn't need to have beliefs about whether I am acting on some maxim, or policy, that you could will to be a universal law.

To support his suggestion that Kant assumes (F), Rawls appeals to Kant's remark that all of his various principles are merely different statements of 'precisely the same law'. [177] Rawls takes this remark to imply that Kant's other principles 'cannot add to the content' of Kant's Formula of Universal Law. Rawls therefore proposes that we should try to interpret Kant's other principles in ways that make them add nothing, because they contain no other ideas. [178]

Kant is a greater philosopher than this proposal assumes. Kant himself goes even further in underrating his achievements, since he denies that he is presenting even one new principle. [179] The truth is that, in the cascading fireworks of a mere forty pages, Kant gives us more new and fruitful ideas than all the philosophers of several centuries. Of the qualities that enable Kant to achieve so much, one is inconsistency. If we ignore some of Kant's claims because they conflict with others, we may miss some of what Barbara Herman calls the 'untapped theoretical power and fertility' of Kant's ideas. [180]

Kant's Consent Principle is one example. It is surprising that this principle has been so little discussed. This principle has great appeal, and is worth considering as a separate moral idea, not merely as another way of stating Kant's Formula of Universal Law. So in asking what this principle implies, I shall not assume (F).

When we ask whether someone could rationally consent to some act, our question should be about consent in the *act-affecting* sense. It is not worth asking whether people could rationally consent to being treated in some way, if their refusal of consent would be a mere declaration, or protest, which would either make no difference to how others would treat them, or might make others treat them even worse. If that were true, it might be rational for these people not to protest, even if they were being treated in ways that were very bad for them, and very wrong.

Our question should also be about *informed* consent. When people do not know what effects some act might have, it is irrelevant whether they could rationally consent to this act. People could rationally consent to being grossly maltreated, if they did not know what was being done to them. For these reasons, we can restate the Consent Principle as

> CP: It is wrong to treat people in any way to which they could not rationally consent in the act-affecting sense, if these people knew the relevant facts, and we gave them the power to choose how we treat them.

We should be counted as *treating* people in some way when we know that our act, or one of its possible alternatives, would or might affect these people in some way, or be an act with which they would have some personal reason to be concerned. That could be true even when our way of acting would not causally affect these people. Two examples would be failing to save someone's life, or breaking a promise to someone who is dead.

When people know the relevant facts, they could rationally consent to some act just when these facts would give them sufficient reasons to consent. People have *sufficient* reasons to consent to some act when these reasons are not weaker than any reasons they might have to refuse consent. So the Consent Principle could be more briefly stated as

> CP2: It is wrong to treat people in any way to which they would not have sufficient reasons to consent in the act-affecting sense.

In stating this principle in these ways, I assume that we are rational insofar as we respond to reasons or apparent reasons. On some other views about rationality, CP and CP2 state different principles, which might have different implications. If you accept such a view, you should take the Consent Principle to be stated by CP2. When I ask whether someone could rationally consent to some act, I shall be asking whether this person would have sufficient reasons to consent.

For the Consent Principle to succeed, it must both be in itself plausible, and have plausible implications. This principle must not require too many acts that seem to us to be clearly wrong, or *condemn*---in the sense of implying to be wrong---too many acts that seem to us to be clearly morally required. If this principle both implies and plausibly supports many of our best considered intuitive moral beliefs, we could justifiably use this principle to guide some of these beliefs, by revising or extending them.

What the Consent Principle implies depends on our assumptions about which facts give us reasons. If we assume either some desire-based subjective theory, or Rational Egoism, the Consent Principle would not be plausible, and would mistakenly condemn many permissible or morally required acts. Suppose, for example, that in

> *Earthquake*, two people, *White* and *Grey,* are trapped in slowly collapsing wreckage. I am a rescuer, who could prevent this wreckage from either killing White or destroying Grey's leg.

White, Grey, and I, we should assume, are all strangers to each other; nor do we differ in any other morally relevant way. We should make similar assumptions about my later imagined cases. If these are the only morally relevant facts, it is

clear that I ought to save White's life. We can next suppose that, if I saved Grey's leg, that would be much better for Grey, and would much better fulfil Grey's present fully informed desires. According to both desire-based subjective theories, and Rational Egoism, Grey could not then rationally consent to my failing to save her leg, so the Consent Principle would mistakenly imply that it would be wrong for me to save White's life. [181] Similar claims apply to countless other cases. There are countless right acts to which, according to both subjective theories and Rational Egoism, some people could not rationally consent. If we accept any of these theories, as many people do, we must reject the Consent Principle. That may be one reason why this principle has been so little discussed.

We ought, I have claimed, to accept some *wide value-based objective* theory. On such views, when one of two possible choices would make things go in a way that would be impartially better, but some other choice would make things go better either for ourselves or for those to whom we have close ties, we often have sufficient reasons to make either choice. *Earthquake*, I believe, is one such case. If Grey could choose how I would act, she would have sufficient reasons, I believe, to make either choice. Grey could rationally choose that I save her leg, since this choice would be much better for her. But she would not be rationally required to make this choice. Grey could rationally choose instead that I save White's life. Grey could rationally regard White's well-being as mattering about as much as hers, and White's loss in dying would be much greater than Grey's loss in losing her leg.

White, in contrast, could not rationally choose that I save Grey's leg. We could rationally choose to benefit some stranger, I believe, even if our choice would make us lose a somewhat greater benefit. But there is too great a difference between the possible benefits to White and Grey. White would not have sufficient reasons to give up her life so that I could save Grey's leg. [182] So the Consent Principle rightly requires me to save White's life, since this is the only act to which both Grey and White could rationally consent.

Suppose next that, in

> *Lifeboat*, I am stranded on one rock, and five people are stranded on another. Before the rising tide drowns all of us, you could use a lifeboat to save either me or the five. We are all young, and would lose as much in dying.`

Though some people would believe that you ought to give me some chance of being saved---which might be a chance of one in six or even one in two---most people would believe, more plausibly, that you ought to save the other five people.

If I could choose how you will act, could I rationally choose that you save the five rather than me? Some people would answer No. These people might agree that, if I chose to give up my life to save five strangers, this choice would be morally

admirable. But this choice, they believe, would also be irrational. On this view, since I could not rationally consent to your saving the five rather than me, the Consent Principle implies that it would be wrong for you to save the five. That is an unacceptable conclusion. So if we accept this view, we would have to reject the Consent Principle.

We ought, I believe, to reject this view. Though I could rationally choose that you save me, I could also rationally choose, I believe, that you save the five. I would have sufficient reason to give up my life if I could thereby save five strangers.

Could the five rationally consent to your saving me rather than them? The word 'consent' may be misleading here, since we may assume that each of the five could give consent only on her own behalf. But we should not make that assumption. When we apply the Consent Principle, we should ask whether, if each of the five could give or refuse consent to your act in the act-affecting sense, thereby choosing how you will act, this person could rationally choose that you save me rather than the five. The answer is clearly No. Suppose that *Green* is one of the five. Green would not have sufficient reasons to choose that you save me rather than saving *both* Green *and* four other people. Green would have both strong personal and strong impartial reasons not to make this choice. On these assumptions, the Consent Principle rightly implies that you ought to save the five, since this is the only act to which both I and each of the five would have sufficient reasons to consent.

As these examples suggest, whether we could rationally consent to some act depends in part on the benefits or burdens that would come to us or other people in the different outcomes that would be produced by this and the other possible acts. It makes a difference both how great these benefits or burdens would be, and to how many people they would come. It also makes a difference, I believe, how badly off we and the other people are. And it may make a difference whether we or the others are responsible for various features of our situation. That might be true, for example, if some of us have worked to produce the possible benefits, or are responsible, through negligence or recklessness, for the possible burdens. There may be other acts to which we would not have sufficient reasons to consent even though these acts would not impose any significant burden on us. We can have strong reasons, for example, to refuse consent to other people's deciding how our lives will go, even when these people's decisions would not be bad for us.

Whenever people could not rationally give informed consent to being treated in some way, there must be facts about these acts which give these people decisive reasons to refuse consent. White, I have claimed, could not rationally consent to my saving Grey's leg rather than White's life, given the fact that White's loss would be so much greater than Grey's. This fact can also be claimed to make this act wrong. Similar claims apply to other cases. Whenever certain facts would give some people decisive reasons to refuse consent to being treated in some way, these facts

would also provide moral objections to these acts.

For the Consent Principle to be true, these moral objections must be decisive, since this principle condemns all acts to which anyone could not rationally consent. For this much stronger claim to be defensible, it must be always or nearly always true that

> (G) there is at least one possible act to which everyone would have sufficient reasons to consent.

In cases in which there was no such act, the Consent Principle would mistakenly imply that whatever we did would be wrong. (G) is least likely to be true when

> (H) each of our possible acts would impose some very great burden on at least one person, or would deny at least one person some very great benefit.

Such people would have very strong reasons to refuse consent to being made to bear such burdens, or being denied such benefits. One such case is *Lifeboat*, in which either I or the five will be denied the benefit of being saved from an early death. In this case, I have claimed, (G) is true. Though I would have very strong reasons to choose that you save my life, these reasons would not be decisive. I would have sufficient reasons, I believe, to consent to your saving the five rather than me. If I would have such reasons, that strongly supports the view that, at least in cases in which the stakes are lower, there would be at least one possible act to which everyone could rationally consent.

I shall return to the question whether there would always be such an act. If that is true, we could argue:

> Whenever someone could not rationally consent to some act, there must be certain facts that give this person decisive reasons to refuse consent to it. These facts provide moral objections to this act.
>
> These objections must be significantly stronger than the objections to any other possible act to which everyone *could* rationally consent.
>
> Whenever there are significantly stronger moral objections to one of two acts, this act is wrong.

Therefore

> It is wrong to act in any way to which anyone could not rationally consent.

Though this argument is rough, it is enough to show that the Consent Principle is in itself plausible.

This principle also has many plausible implications, since it condemns many of the acts that are most clearly wrong, such as many acts of killing, injuring, coercing, deceiving, stealing, and promise-breaking. Many of these acts treat people in ways to which they would not have sufficient reasons to consent.

26 A Superfluous Principle?

According to some writers, nothing is achieved by appealing to the possibility of rational consent. These writers concede that it may always be wrong to treat people in ways to which they could not rationally consent. But what is morally important, these writers claim, is not the fact that these people could not rationally consent to these acts, but the various facts that give these people decisive reasons to refuse consent.

In considering this objection, we can first distinguish two aims that any moral principle might achieve. This principle might provide a reliable *criterion* of wrongness, by truly telling us that all acts of a certain kind are wrong. This principle might also be *explanatory*, by describing one of the reasons why these acts are wrong, or one of the facts that make them wrong. According to the writers I have just mentioned, even if the Consent Principle is true, we do not need this principle as a criterion, nor is this principle explanatory.

This objection has most plausibility when we consider acts whose main effects would be on one person, with whom we cannot communicate and whose preferences we don't know. In such a case, we would have to make some decision on this person's behalf. Surgeons, for example, sometimes have to make decisions on behalf of their unconscious patients. In such cases, it may be enough to claim that we ought to try to do what would be best for this other person, or what would benefit this person most. It may not be worth adding that it would be wrong for us to act in any way to which this person could not rationally consent.

In most important cases, however, our choice between possible acts would have significant effects on two or more people. The view that I have just described might be widened to cover such cases. According to *Act Utilitarianism*, or

> *AU*: We ought always to do whatever would, on the whole, benefit people most, by giving people the greatest total sum of benefits minus burdens.

Act Utilitarians might claim that

> (I) everyone could rationally consent to all and only the acts that would, on the whole, benefit people most.

If (I) were true, AU and the Consent Principle would always *coincide*, by requiring all the same acts. These Utilitarians might then claim that AU is more fundamental, and that, since AU tells us how we ought always to act, the Consent Principle adds nothing to our moral thinking. But this claim would be false. If it were only these Utilitarian acts to which everyone could rationally consent, the Consent Principle would support AU. (I)'s truth would give us a further reason to believe that these acts were morally required, and a further reason to act in these ways.

(I) is not, I believe, true. There are many Utilitarian acts to which some people could not rationally consent, and many non-Utilitarian acts to which everyone could rationally consent. I shall give some examples later.

If the Consent Principle is true, this principle would be more than a reliable criterion of wrongness. Whenever someone could not rationally consent to being treated in some way, this fact would provide an objection to this act, and could be claimed to be one of the facts that would make this act wrong. The Consent Principle would have most importance when we must choose between many possible acts that would have significant effects on many people, whose interests or aims conflict. In such cases, if there is only one possible act to which everyone could rationally consent, this fact would give us a strong reason to act in this way, and might be enough by itself to explain why all the other possible acts would be wrong.

We have another reason to ask whether the Consent Principle is true. Even if we do not need to use this principle as a criterion of wrongness, it is worth asking whether we could achieve what I call *Kant's ideal*, by treating everyone only in ways to which they could rationally consent.

27 Actual Consent

It is often morally important whether people *actually* consent to being treated in some way, or whether, if they had the opportunity, these people *would in fact* consent. In such cases, it is not enough to ask whether people *could rationally* consent to some act. Some rapist might claim that his victim could have rationally consented to having sexual intercourse with him. Even if this claim were true, that would not justify this man's act. It may be objected that, since the Consent Principle does not require actual consent, this principle mistakenly ignores the moral importance of such consent.

That is not, however, true. Even if this man's victim could have rationally consented to having sexual intercourse with him, she could not have rationally consented to being raped, by having such intercourse forced on her despite her actual refusal of consent. In this and many other kinds of case, we could not

rationally consent to being treated in some way without our actual consent. Since the Consent Principle condemns all such acts, this principle does not ignore the moral importance of actual consent.

This principle might instead be claimed to give, implicitly, *too much* importance to actual consent. Consider

> *the Veto Principle*: It is wrong to treat people in any way to which they either do in fact, or would in fact, refuse consent.

Like the similar Choice-Giving Principle, this principle is clearly false. There are countless permissible or morally required acts to which some people either do or would refuse consent. In *Earthquake*, for example, even if Grey refuses her consent, I ought to save White's life rather than Grey's leg. And there is often no possible act to which everyone would in fact consent. Someone might now argue:

> It is wrong to treat people in any way to which they could not rationally consent.
>
> (J) No one could rationally consent to being treated in any way to which they either do in fact, or would in fact, refuse consent.

Therefore

> It is wrong to treat people in any way to which these people either do in fact, or would in fact, refuse consent.

If (J) were true, the Consent Principle would imply the Veto Principle. That would make the Consent Principle clearly false. [183]

Should we accept (J)? It may be confusing to ask whether people could rationally consent to some act to which they actually refuse consent, since these people could not at the same time both give and refuse consent. To make our question clearer, we can appeal to another version of the Consent Principle. According to

> CP3: It is wrong to treat people in any way to which, if they had known the relevant facts, these people could not have rationally given, in advance, their irreversible consent.

Our consent to some act is *irreversible* when we know that, if we later withdrew our consent, that would make no difference to how we would later be treated.

There are many acts to which we could not rationally give such irreversible consent in advance. For example, we could seldom rationally give such consent in advance to sexual acts to which, at the time of these acts, we refuse consent. That would

seldom be rational because the nature of most sexual acts is greatly affected by whether, at the time, both or all of the people involved actually consent.

There are also many acts, however, to which we *could* rationally give such irreversible consent. For us to have sufficient reasons to give such consent, it might have to be true both that

> (K) we have some reason to give irreversible consent, thereby restricting our future freedom,

and that

> (L) we shall not later learn some fact that might give us decisive reasons to regret that we earlier gave such consent.

But these conditions are often met. In many cases, for example, someone needs to know that someone else's consent is binding, and cannot be withdrawn. Suppose that, in *Earthquake*, once I had started to save White's life rather than Grey's leg, it would be dangerous for me to stop. Suppose next that Grey knows all of the relevant facts, and that Grey is just as able to make a good decision now as she will later be. On these assumptions, Grey could rationally make her decision now. We are not rationally required to postpone our decisions whenever we can. And Grey would have sufficient reasons, I have claimed, to choose that I save White's life rather than Grey's leg. If that is so, Grey would also have sufficient reasons to give irreversible consent to my later doing that. Grey could rationally say, 'Go ahead and save White's life, even if I later change my mind'.

When we apply the Consent Principle in the form stated by CP3, our aim is only to ask whether people could rationally consent to being treated in some way to which they in fact refuse consent. This question is easier to answer when we apply it to irreversible consent given in advance. In many actual cases, people would not in fact have sufficient reasons to give such consent in advance, thereby committing themselves in a way that would restrict their future freedom. But given the aims of our imagined thought-experiment, we can *suppose* that these people would have had sufficient reasons to make their decision in advance. Our question can be whether, on that supposition, these people would have had sufficient reasons to give their irreversible consent.

In many cases, I believe, people could rationally give such irreversible consent to being later treated in some way without their later actual consent. If that is true, we can reject premise (J) of the argument above. The Consent Principle does not imply the Veto Principle, and avoids at least the strongest objections to that principle.

Though we ought to reject the Veto Principle, we could plausibly accept a much weaker version of this principle. According to what we can call

> *the Rights Principle*: Everyone has rights not to be treated in certain ways without their actual consent.

When we claim that people have *rights* not to be treated in certain ways, we mean in part that, without these people's consent, such acts would be wrong. We can call these the *veto-covered* acts.

In stating this principle, it would often be hard to decide which are the acts that people have a right to veto. For this principle to be acceptable, these rights must be narrowly described. We should not, for example, claim that everyone always has a right not to be killed, since some killings are unavoidable, and some others are justified, as is true in some cases of self-defence. But we might claim that we all have certain more restricted rights, such as a right not to be killed for our own good without our consent. We might similarly claim that everyone has a right to veto what is done to their bodies, not only sexually but in other ways. On one view, for example, everyone has a right not to be kept alive by medical treatments to which they refuse consent.

As well as condemning veto-covered acts to which people refuse consent, the Rights Principle should require us to give people the opportunity to refuse consent. When we cannot give people this opportunity, because we cannot communicate with them, we ought to try to treat these people only in those veto-covered ways to which, *if* they had the opportunity, they *would* consent. When people cannot consent to some act, but we know that they would have given or refused consent, this fact would have similar moral significance. When we ask whether people *would in fact* consent to some act, that is quite different from asking whether these people *could rationally* give such consent. We might know that certain people would not in fact consent to some veto-covered act, even though it would be irrational for them to refuse consent. In cases that involve veto-covered acts, we might say, people have a right to be irrational, and to suffer the effects.

For consent to be morally significant, however, it must be given by people who have sufficient understanding of the relevant facts, and are able to consider these facts in a sufficiently clear-headed way. These conditions can be met by people who make some irrational decision. But the Rights Principle should not appeal to consent that is given by people who don't understand the most important relevant facts, or who are too young, or seriously mentally ill, or are affected by some other seriously distorting influence, such as being drunk, drugged, or threatened. Under such conditions, we can say, people cannot *validly* give or refuse consent. [184]

When people cannot validly consent to some act, we might ask whether, *if* these

people had been free from such distorting influences, they *would* have given such consent. But this question may be hard to answer. And there are other ways in which we could plausibly revise or extend the Rights Principle. Rather than appealing to the *hypothetical* consent that we believe that someone would have given at the time at which we act, we may be able appeal to this person's *actual* consent at some earlier time. In some cases, when people know that that they will later be affected by some distorting influence, they may validly give or refuse consent in advance to being later treated in some way. We may believe that we should later follow these earlier valid decisions. In some other cases, people cannot give valid consent at the time, and they have neither given nor refused consent in advance. In such cases, we may believe that we ought to try to treat these people only in ways that they would later *retroactively endorse*, since they would later be glad that we acted as we did. Unlike the claim that people *would* have given valid consent, which could not be confirmed, most predictions of later endorsement could be either confirmed or shown to be false. That would provide a useful check on our use of such predictions to justify our acts.

We might next qualify the Rights Principle, so that it reflects the fact that the conditions for valid consent are matters of degree. When people are under some influence that to some extent distorts their judgment, though not so greatly as to make their decisions invalid, we may give these decisions less moral weight.

To illustrate some of these points, we can return to the view that everyone has a right not to have surgery performed on them without their consent at the time. This right is often claimed to be absolute, in the sense that it has no exceptions. But there are, I believe, some exceptions. Suppose that, in

> *Surgery*, to save *Green's* life, we must operate on her without anaesthetics. This operation would be very painful, but it would give Green many more years of worthwhile life. Green gives irreversible consent to this operation in advance, permitting us to use force, if necessary, if the pain later leads her to change her mind.

Before the discovery of anaesthetics, many people rationally gave such irreversible consent to life-saving surgery. If Green gave such consent, and the pain did later lead her to change her mind, we would be justified, I believe, in using force to complete this surgery. The Rights Principle should permit this act. We might however believe that, since great pain is a seriously distorting factor, Green's withdrawal of consent would not be valid.

Suppose next that, in a different version of this case, Green refuses to give such consent in advance. We may believe that this refusal is decisive, concluding that we ought to let Green die. But we might instead believe that Green's refusal should be regarded as invalid, or should be given less weight, since the immediate prospect of

great pain is another distorting factor, making it too difficult for people to make rational decisions. On one version of the Rights Principle, we could justifiably impose this surgery on Green if the pain of the surgery would be brief, and we also have strong reasons to believe that Green would later endorse our decision, being glad that we had saved her life despite her refusal of consent both at the time and in advance. We might know that, in such cases, most people endorse such surgery as soon as their worst pain is over.

In such cases, however, there is another, less obvious distorting factor. When we consider experiences that are painful, most of us have a strong *bias towards the future*. Once our pain is over, we care about it much less, or not at all. That makes it harder to justify imposing painful life-saving surgery by appealing to the fact that, after such surgery is over, almost everyone retroactively endorses such acts. Given our bias towards the future, we may underestimate the strength of the reasons that we earlier had to want to avoid what is now past pain.

Suppose next that, in

> *Depression*, Blue decides to kill herself. We have strong reasons to believe that, if we forcibly prevented Blue's act, Blue's depression would soon lift, and the rest of her life would go well.

Many of us would believe that we could justifiably override Blue's decision, and use force to prevent her from killing herself. If we accept the Rights Principle, we might claim that severe depression is a sufficiently distorting factor, so that Blue's refusal of consent is not valid. But if we made this claim, our standards of validity would be high, and would often fail to be met. People who are severely depressed may know the relevant facts, nor are they clearly incapable of making rational decisions. It would be more plausible to claim that, though Blue's depression does not make her refusal of consent invalid, it makes her less able to make rational decisions, so that Blue's refusal might be morally outweighed by her decisions at other times. For example, if Blue has frequent temporary depressions, she may have consented in advance to our later using force to prevent her from killing herself while she is depressed. That may be enough to justify our act, though we would here be overruling Blue's *valid* refusal of consent at the time. And given the irreversibility of suicide, such acts might be justified even without such earlier consent. There is here an important asymmetry. If we frustrate Blue's attempt to kill herself, she could later try again, but if we allow her to kill herself, she could not later try to stay alive.

For an example of a different kind, suppose that, in

> *False Belief*, we could save Brown's life with a blood-transfusion. Brown refuses her consent, since she is a Jehovah's Witness who believes blood-tranfusions to

be wrong.

For people to give valid consent, I have said, they must know the relevant facts. If Brown knew these facts, she would know that blood-transfusions are *not* wrong, and she could then have rationally consented to our saving her life in this way. But we might believe that, since Brown actually refuses her consent, it would be wrong for us to save her life in this way. When people refuse consent to some act because they have certain kinds of false belief, such as certain moral or religious beliefs, we may believe that this refusal should be regarded as valid.

In these remarks, I have assumed that present consent matters more than past consent, which matters more than retroactive endorsement. It is worth asking why these differences in timing have such significance.

If I cannot communicate with you, I might ask which of my possible acts would be most likely to fulfil your desires or preferences. As I have said, though our own preferences give us only derivative reasons, we can have non-derivative reasons to try to fulfil other people's preferences. In trying to do what would fulfil your preferences, I would have no reason to give priority to what you *now* prefer. Suppose that I have reasons to believe both that you would now want me to act in one of two ways, and that you would later change your mind, and would be glad if I had acted in the other way. I also have reasons to believe that, when you later changed your mind, you would know more of the relevant facts, so that your later preference would be better grounded. On these assumptions, I believe, I could rationally and justifiably give priority to fulfilling this later preference.

As one example of this kind, we can suppose that, as your doctor, I must decide whether to treat you in some way. Since you are unconscious, I cannot ask for your consent, and can only try to predict what you would prefer, and choose. This treatment would cause you some pain in the near future, but it would later save you from much greater pain. I have good reasons to believe that you would now prefer me not to treat you in this way, but that when you later learnt how bad that greater pain would be, you would change your mind. Given these facts, I could plausibly believe that I should fulfil your predictable later, better informed preference.

Suppose next that, in a different version of this case, you *are* conscious, so that I can ask for your consent to my proposed treatment. If you refuse consent, this fact might clearly morally outweigh my plausible prediction that you would later regret having made this decision. Though I have no reason to give your present *preferences* priority over your future preferences, I do have reason, when you are able to decide how I shall treat you, to give priority to what you now *decide.*

To explain this difference, we can first note a similar fact about our attitudes to our own and other people's beliefs. When I am trying to reach the truth about some question, and I take into account other people's beliefs, I would have no reason to give greater weight to other people's *present* beliefs. If I had some way of knowing what other people would later believe, I might have good reasons to give greater weight to these people's future beliefs, since these beliefs would be better grounded. I might also have good reasons to give greater weight to some of these people's past beliefs, which were freer from some distorting influence. I must, in contrast, give priority to *my present* beliefs. I can believe, for example, that some claim is false, though I did earlier believe, or shall later predictably believe, that this claim is true. But I cannot believe that some claim is false though I *now* believe that this claim is true. We can never base our decisions on the truth *rather* than on what *we now* believe to be true.

Similar claims apply to our decisions. We must give some priority to what we now decide, since these decisions are based on what we now believe to be true. And even when our beliefs have not changed, or we believe that they will not change, we must give priority to what we now decide, since we cannot make our decisions from some past or future point of view. We have to live our lives from our own present point of view. These facts may explain why, when other people ought to act only with our consent, these people should also give priority to whether we *now* consent to their way of treating us.

28 Deontic Beliefs

The Consent Principle claims to describe only one of the ways in which our acts may be wrong. Acts may be wrong even though everyone could rationally consent to them.

Many such acts are wrong because some people do not, or would not, actually consent to them. That may be true, as I have said, of most kinds of direct interference with our bodies. Another much larger group of cases involve ownership. People do not always have a right to veto how we treat their property, since we could justifiably use or even destroy many kinds of property, despite the owner's refusal of consent, if that is our only way to save someone else from death or injury. But there are also many cases in which it would be wrong to use or destroy someone's property without this person's actual consent. If I do not have your consent, it may be wrong for me to live in your apartment, wear some of your clothes, and eat what is in your kitchen. In most cases, the Consent Principle would condemn such acts, since we could not have rationally consented in advance to other people's acting in such ways without our consent at the time. But if I had earlier been homeless, cold, and hungry, these facts might have given you sufficient reasons to consent in advance to my acting in these ways. The Consent Principle

would not then condemn my acts. Despite this fact, it might be wrong for me to live in your apartment, wear your clothes, and eat what is in your kitchen, without your *actual* consent to these acts.

There might also be acts that are wrong even if everyone involved actually and rationally gives their valid consent. Many people have that view, for example, about *voluntary euthanasia*: killing someone, as this person asks us to do, for his or her own good. And some acts are wrong for reasons other than the ways in which they treat other people, so that the question of consent does not arise. That is true of cruelty to animals, for example, and some believe it to be true of suicide.

Since acts can be wrong in other ways, or for other reasons, what the Consent Principle implies may in part depend on which acts would be wrong for such other reasons. So when we apply this principle, we must sometimes appeal to our beliefs about which acts are wrong. These beliefs I shall call *deontic*, and the reasons that might be provided by some act's wrongness I shall call *deontic* reasons.

It might be objected that, if we apply the Consent Principle in a way that appeals to these beliefs, our moral reasoning would be circular, or question-begging. Such reasoning could not support our beliefs about which acts are wrong.

This objection is, in part, correct. It could not be true both that

> (M) some act would be wrong because someone could not rationally consent to it,

and that

> (N) this person could not rationally consent to this act because it would be wrong.

For some act to be wrong *because* someone could not rationally consent to it, this person must have decisive *non-deontic* reasons to refuse consent. But people often have such reasons. In *Earthquake*, for example, White has such a reason to refuse consent to my saving Grey's leg rather than White's life. White could not rationally consent to this act, not because it would be wrong, but because White's loss in dying would be so much greater than Grey's loss in losing a leg. When applied to such cases, and many other kinds of case, the Consent Principle supports and helps to justify some of our deontic beliefs.

As I have said, however, we must sometimes apply the Consent Principle, in a way that appeals to our other deontic beliefs. Suppose that in a variant of *Earthquake*, which we can call

> *Means,* White and Grey are trapped in slowly collapsing wreckage. Though

> White's life is threatened, Grey is in no danger. I could save White's life, but only by using Grey's body as a shield, without Grey's consent, in some way that would destroy her leg.

Many of us would believe that, given Grey's refusal of consent, it would be wrong for me to save White's life in this way, by destroying Grey's leg. On this view, which we can here suppose to be true, it is wrong to act in any way that gravely injures someone, without this person's consent, as a means of benefitting someone else.

In applying the Consent Principle to this case, we can first set aside our assumption that this act would be wrong. If this act would not be wrong, this case would not be relevantly different from *Earthquake*. In both *Earthquake* and *Means*, either White will die or Grey will lose her leg. These cases would differ only in how the saving of White's life would be causally related to the loss of Grey's leg. Grey would have no strong reason to prefer to lose her leg in one of these ways. Neither, we can suppose, would be worse for her. In both cases, I believe, Grey could have rationally given her irreversible consent to my later saving White's life, even though Grey would then lose her leg. And in both cases, since White's loss would be so much greater than Grey's, White could not have rationally consented to my failing to save her life. On these assumptions, the Consent Principle would require me in *Means* to save White's life by destroying Grey's leg, since that is the only act to which both White and Grey could rationally consent.

Return now to our assumption that this act would be wrong. If the Consent Principle required this wrong act, that would be a strong objection to this principle. But this principle would not, I believe, require this act. If it would be wrong for me to save White's life by destroying Grey's leg, this act's wrongness would give White a sufficient reason to consent to my failing to act in this way. We all have sufficient reasons, I believe, to consent to someone's failing to benefit us, even when this benefit would be as great as the saving of our life, if this way of benefiting us would wrongly injure someone else.

Here is another way to defend this belief. We are discussing possible consent in the act-affecting sense. For White to be able to give or refuse such consent, we must suppose that I have given White the power to choose how I shall act. If White chose that I save her life by wrongly injuring Grey, she would be partly responsible for my wrong act. That would make it wrong for White to make this choice. And we always have sufficient reasons, I believe, not to make choices that would be morally wrong. I am not claiming here that it would be irrational for White to make this choice. Perhaps White could rationally choose that I act wrongly, since this choice would save White's life. But White would also have sufficient reasons to choose instead not to be partly responsible for this wrong act. Since White could rationally consent to my failing to save her life by destroying Grey's leg, the Consent Principle

would not mistakenly require this wrong act. [185]

It might next be objected that, since Grey *could* rationally consent to my saving White's life in this way, the Consent Principle mistakenly permits this act even when, because Grey actually refuses consent, this act would be wrong. But this objection misunderstands the Consent Principle. This principle claims to describe only one of the facts that make acts wrong. So, when this principle does not condemn this way of saving White's life, it does not thereby imply that this act is morally permitted.

Similar remarks apply to other cases. We are discussing cases in which some act of ours would be wrong, not even in part because someone could not rationally consent to this act, but for other reasons. We can argue:

> The Consent Principle requires some act only when someone would not have sufficient reasons to consent to our failing to act in this way.
>
> (O) Whenever some act would be wrong for other reasons, this act's wrongness would give everyone a sufficient reason to consent to our failing to act in this way.

Therefore

> The Consent Principle could never require acts that are wrong for other reasons.

We can similarly argue that this principle could never condemn acts that are morally required for other reasons. If some act is required, all of its alternatives would be wrong, and that would give everyone sufficient reasons to consent to this act.

On some views, premise (O) might be denied. Suppose that, in

> *Fire*, Black is trapped in burning wreckage, and will soon, if I do nothing, die a slow and painful death. I cannot save Black from this pain except by killing her now, before the increasing heat forces me to withdraw.

Suppose next that, knowing these facts, Black asks me to kill her. This act, I believe, would be morally justified. If that is true, Black could not rationally consent to my failing to benefit her, by giving her a swifter, painless death. On these assumptions, the Consent Principle requires me to kill Black, as she requests.

On one view, even in cases like *Fire*, such voluntary euthanasia is wrong. If it would be wrong for me to benefit Black by giving her this better death, would this act's wrongness give Black a sufficient reason to consent to my failing to act in this way? Some people might answer No. These people might agree that, in *Means*,

White could rationally consent to my failing to save her life by destroying Grey's leg. But White's reason to give such consent is provided by the fact that I could save White's life only by wrongly injuring someone else. No such claim applies to *Fire*. If I killed Black at her request, I would not be wrongly injuring anyone else. These people might believe that, given this difference, the wrongness of my killing Black would *not* give Black a sufficient reason to consent to my failing to benefit her in this way. On these assumptions, premise (O) would here be false, and the Consent Principle would require an act that would be wrong.

This example does not, I believe, provide a strong objection to the Consent Principle. Few people would believe both that this act would be wrong and that its wrongness would not give Black a sufficient reason to consent to my failing to act in this way. And we could plausibly reject this view.

Consider next a different version of *Fire*. Suppose that, though Black knows that my killing her would be better for her, she refuses her consent. Some people might believe both that this act would be wrong without Black's consent, and that Black could not have rationally consented in advance to my failing to give her, without her later consent, this swifter, better death. If these beliefs were both true, premise (O) would be false, since the Consent Principle would here require me to act wrongly. But I believe that, if it would be wrong for me to kill Black without her actual consent at the time, this act's wrongness would have similarly given Black sufficient reasons to consent in advance to my failing to act in this way.

For an example of a different kind, suppose that, in

> *Parents*, after some shipwreck, you and I each have a child whose life is in danger. I have a life-belt, which I could use to save either my child or yours.

Suppose next that, as most of us would believe, I ought to save my child. Could you rationally consent to my acting in this way?

On one view, the answer is No. If I gave you the power to choose how I would act, you ought to choose that I act wrongly, by saving your child. Though you would be partly responsible for my wrong act, your duty to protect your child would morally outweigh your reason not to choose that I act wrongly. Given this fact, and your other strong reasons to want me to save your child, you could not rationally consent to my failing to act in this way. On these assumptions, the Consent Principle would here require me to act wrongly, by saving your child rather than mine.

If we accept this view, and we have similar beliefs about other relevantly similar cases, we would have to revise the Consent Principle, so that it did not apply to this kind of case. According to

> CP4: It is wrong to treat people in any way to which they would not have sufficient reasons to consent, except when these people would not have such reasons because the case involves conflicting person-relative moral obligations.

Though this revision would restrict the scope of the Consent Principle, it would not make this principle less plausible. When we apply this principle, we appeal to a thought-experiment, by asking whether other people could rationally choose that we act in some way. We cannot usefully ask this question when it makes a moral difference whether it is we or someone else who chooses how we shall act. In such cases, it might be wrong for us to do what it would be right for someone else to choose that we do. Our thought-experiment would here lead us to ignore this fact. We should not expect that, in such cases, the Consent Principle could help us to decide which acts are wrong. Since we can give this explanation of why this principle should not be applied to cases of this kind, such cases would not cast doubt on the moral idea that this principle expresses.

This revision may not, however, be needed. We can ask

> Q1: Could we have a duty to choose, or bring it about, that someone else acts wrongly?

On some moral views, the answer is sometimes Yes. One such view is the kind of moral nationalism that was widely accepted in Europe before and during the First World War. On this view, if your nation is at war with mine, it might be my patriotic duty to try to get you to act wrongly, by unpatriotically giving me the information with which my nation's army can defeat yours.

Kant's answer to Q1 would be No. And if we are right to accept this answer, *Parents* does not undermine the Kantian ideal. On such a view, we can have what are in one sense conflicting personal-relative obligations. It might be my duty to save my child, and your duty to save yours, though my doing my duty would make it impossible for you to do yours. But in *Parents* I could act in a way to which you could rationally consent. Since it would be wrong for me to save your child rather than mine, you could not have a duty to choose that I act in this way, and this act's wrongness would give you a sufficient reason to consent to my doing my duty, by saving my child.

For a different objection, suppose next that, in

> *Equal Claims*, I could save either your life or Grey's.

It may seem that, in this case, you could not rationally consent to my saving Grey's life rather than yours. You would have strong personal reasons not to give such consent. And since your death would be impartially as bad as Grey's, these

personal reasons may seem to be decisive. Grey would have similar reasons not to consent to my saving your life rather than Grey's. The Consent Principle may seem here to fail, by mistakenly implying that, whatever I do, I shall be acting wrongly, since I shall be treating someone in some way to which this person could not rationally consent. We can plausibly claim, however, that I ought to give both you and Grey an equal chance of being saved. And if it would be wrong for me not to give you both an equal chance, this fact would give you both sufficient reasons to consent to this act.

My remarks about these cases do not prove that we could always justifiably follow the Consent Principle, thereby achieving Kant's ideal. Some people would reject these claims. And there may be other kinds of case in which, on plausible assumptions, there would be no possible act to which everyone could rationally consent.

These cases also show, however, that Kant's ideal makes a significant, substantive claim. For another example, suppose that, in

> *High Price*, I sell you some product that, as only I know, you could have bought much more cheaply elsewhere. [186]

Suppose next that, since you are not rich, you could not have rationally chosen to pay this higher price. The Consent Principle then implies that, in taking your money, I act wrongly. Some of us would believe that, since you freely consent to my taking your money, I do not act wrongly. But the Consent Principle is not obviously mistaken here. We can plausibly believe that, just as I ought to warn you if the product that I am selling is in some way defective, I ought to tell you that you could buy this product much more cheaply elsewhere.

In several other cases, I believe, the Consent Principle has implications that are plausible, though not undeniable. That makes it worth asking, of the most plausible views about both morality and rationality, which views are compatible with Kant's ideal.

29 Extreme Demands

Suppose next that, in

> *Self*, I am trapped with White in slowly collapsing wreckage. I could save either White's life or my leg.

On some views, this case is morally just like *Earthquake*. I ought to save White's life rather than my leg, since White's loss would be much greater than mine. Most of

us would have a different view. On this view, though it would be wrong for me to save some other stranger's leg rather than White's life, I would be morally permitted to save *my* leg. We ought to save any stranger's life when that would cost us very little. But the cost to me here would be too great.

What does the Consent Principle here imply? If White could choose how I would act, could White rationally choose that I save my leg rather than her life?

The answer may seem to be No. It may seem that White could not rationally consent to anyone's saving anyone's leg rather than White's life. But this view is too simple. We can have reasons to care, not only about *what* will be done, but also about *who* will be doing these things, and *why* they will be doing them.

To illustrate this point, it will help to lower the stakes. Suppose first that I could either save White from a week of pain, or save some other stranger from only one day of similar pain. There is no other relevant difference between White and this other stranger. On that assumption, I would have no reason to give less weight to White's well-being. And White could not rationally consent to my choosing, *for no reason,* to help the other stranger rather than saving White from her much greater burden. That choice would treat White as if she were inferior, or didn't even exist.

Suppose next that, rather than saving White from her week of pain, I could save myself from one day of pain. Though I would have no reason to care more about the well-being of one of two strangers, I do have reasons to care more about my own well-being. We all have reasons to be specially concerned about what happens to ourselves. Since everyone has such reasons, we could often rationally consent to other people's giving priority, for these reasons, to their own well-being. Though White could not rationally consent to my choosing, for no reason, to save some other stranger from a day of pain rather than saving White from her week of pain, White may have sufficient reasons to consent to my saving *myself* from this much smaller burden. *This* act would not treat White as if she were inferior, or didn't even exist.

In *Self*, however, the stakes are much higher. White may not have sufficient reasons to consent to my saving my leg rather than White's life.

Would it make a difference if, as most of us would believe, I would be morally permitted to save my leg rather than White's life? Perhaps not. There may be a difference here between permissibility and wrongness. If I could save White's life only by acting wrongly, as we have supposed to be true in *Means*, this act's wrongness, I have claimed, would give White a sufficient reason to consent to my failing to save her life. In *Self*, however, I could save White's life without acting wrongly. And even if I would be morally permitted to save my leg rather than White's life, this act's permissibility may not give White a sufficient reason to consent to my failing to save her life.

If this act's permissibility would *not* give White such a reason, White could not rationally consent to my failing to save her life, so the Consent Principle would require me to save White's life rather than my leg. This principle would here conflict with what most of us believe.

Though few people could save someone else's life only at the cost of a serious injury to themselves, there are many cases to which similar reasoning applies. We could very often either benefit ourselves or give some greater benefit to others. When the benefits to other people would be *much* greater, these people may not have sufficient reasons to consent to our failing to benefit them. Suppose that, in

> *Aid Agency,* I could either spend $200 on some evening's entertainment, or give this money to some efficient aid agency, such as *Oxfam*, which would use this money to save some poor person in a distant land from death, blindness, or some other great harm.

When applied to these two alternatives, the Consent Principle seems to imply that I ought to give this money to this aid agency. This poor person seems not to have sufficient reasons to consent to my failing to act in this way. [187] Similar claims will apply to me tomorrow, and on every other day. And similar claims apply, on every day, to most readers of this book. Compared with the more than a billion people who now live on around $2 a day, most readers of this book are *very* rich.

It would be no objection to the Consent Principle if, for these reasons, this principle requires the rich to transfer much of their wealth or income to the poor. Now that the rich could so easily save so many of the poor from death or suffering, any plausible principle or moral theory makes similarly strong demands. And though the rich are legally entitled to all their property, they may be morally entitled to much less than that. Kant writes:

> Having the resources to practice such beneficence as depends on the goods of fortune is, for the most part, a result of certain human beings being favoured through. . . injustice. [188]

And he is reported to have said:

> one can participate in the general injustice, even if one does no injustice. . . even acts of generosity are acts of duty and indebtedness, which arise from the rights of others. [189]

The Consent Principle may, however, be *too* demanding. After thinking seriously about what justice requires, and considering the relevant arguments, we may have to admit that we rich people ought all to transfer to the poor as much as a tenth of our wealth or income, or even a fifth. But the Consent Principle might require much more than that.

If this principle is too demanding, it could be revised. We might claim

> CP5: It is wrong for us to treat people in any way to which they would not have sufficient reasons to consent, except when, to avoid such an act, we would have to bear too great a burden.

In applying this version of the Consent Principle, we would have to decide when such burdens would be too great. When we consider the moral problems raised by extreme global inequality, that is a very difficult question. One problem is whether and how we should assess the cumulative costs of many small gifts. [190] But we could start by claiming that, in *Self*, I would be permitted to save my leg rather than White's life.

If the Consent Principle is too demanding, and must be weakened in this way, Kant's ideal of interpersonal relations may seem to be in principle impossible, since there would be some right acts to which some people could not rationally consent. But these acts would be right only in the sense that they would be morally permitted. There might be no morally *required* acts to which some people could not rationally consent. So we might still be able to achieve Kant's ideal. It might still be possible for everyone to act only in ways to which everyone could rationally consent. And there might always be at least one such act that would be right. In *Self*, for example, I could save White's life rather than my leg, and this admirable act would be right. If the Consent Principle is too demanding, this would at most imply that, to achieve Kant's ideal, we would have to do more for each other than we are morally required to do. That would not be surprising.

We have, I conclude, strong reasons to accept some version of the Consent Principle. This principle may be too demanding, and there may be some other ways in which it should be revised. But at least in most cases, it is wrong to act in ways to which some people could not rationally consent. When our acts would affect many people, and there is only one possible act to which everyone could rationally consent, this fact gives us a strong reason to act in this way, and may be enough to explain why such acts are morally required. And on some plausible assumptions, the Consent Principle could never go astray, by requiring acts that are wrong for other reasons, or condemning acts that are required.

The Consent Principle cannot, however, be what Kant was trying to find: the supreme principle of morality. [191] Some acts are wrong even though everyone could rationally consent to them. The Consent Principle states one of the ideas that are expressed in Kant's Formula of Humanity. Since we need at least one other principle, we can now turn to another part this formula.

CHAPTER 9 MERELY AS A MEANS

30 The Mere Means Principle

Using people, it is often claimed, is wrong. But this claim needs to be qualified. If we are climbing together, I might use you as a ladder, by standing on your shoulders. And I might use you as a dictionary, by asking you what some word means, or use you as a witness to my signing of my will. Such ways of using people are not wrong. What is wrong, Kant claims, is *merely* using people. As others say, 'You were just using me'.

According to what we can call

> *the Mere Means Principle*: It is wrong to treat anyone merely as a means. [192]

How can we use people without *merely* using them? In explaining this distinction, we can first compare how two scientists might treat the animals in their laboratories. One scientist, we can suppose, does her experiments in the ways that are most effective, regardless of the pain she causes her animals. This scientist treats her animals merely as a means. Another scientist does her experiments only in ways that cause her animals no pain, though she knows these methods to be less effective. This scientist, like the first, treats her animals as a means. But she does not treat them *merely* as a means, since her use of them is restricted by her concern for their well-being.

Similar claims apply to our treatment of each other. According to one rough definition,

> we treat someone *as a means* when we make any use of this person's abilities, activities, or body to help us to achieve some aim.

This definition needs to be qualified in certain ways. We should sometimes distinguish, for example, between doing something to someone as a means of achieving some aim, and treating this *person* as a means. Suppose that, to find out whether I have a broken rib, my doctor presses all over my chest, saying 'Tell me where it hurts'. My doctor is using my body, and hurting me, as a means of getting this information, but she isn't treating *me* as a means. To cover such cases, we might suggest that we do not treat someone as a means when our aim is to benefit this person, and we act with this person's consent.

According to another rough definition,

> we treat someone *merely* as a means if we both treat this person as a means, and regard this person as a mere instrument or tool: someone whose well-being and moral claims we ignore, and whom we would treat in whatever ways would best achieve our aims.

Frances Kamm rejects this second definition. Kamm objects that, if this were the sense in which, on Kant's principle, we must never treat people merely as a means, this principle would be too weak, and too easy to follow. On this definition, for example, if some slave-owner gave even slight weight to the well-being of his slaves, by letting them rest in the hottest part of the day, he would not be treating his slaves merely as a means. But this man surely treated his slaves in a way that Kant's principle condemns. [193]

This objection shows, I believe, not that we ought to revise this definition, but that we ought to revise Kant's principle. For a similar example, consider Kant's claim that

> (A) it is wrong for the rich to give nothing to the poor. [194]

Suppose that some rich man gives to the poor, in his whole life, a total of one dollar and 3 cents. Since this man gives something to the poor, (A) does not imply that he acts wrongly. As this example shows, (A) is too weak, since this man's failure to give more is wrong. The rich act wrongly, we should claim, if they give *too little* to the poor. This kind of wrongness is a matter of degree.

So is the wrongness, we might claim, of treating people merely as a means. On a stronger form of Kant's principle, which we can call

> *the Second Mere Means Principle*: It is wrong to treat anyone merely as a means, or to come close to doing that.

We *come close* to treating someone merely as a means when we both treat this person as a means and give too little weight to this person's well-being or moral claims. That is how my imagined slave-owner treated his slaves, even though he let them rest in the hottest part of the day. So this revised principle condemns this man's acts.

We can next claim that

> (B) we do *not* treat someone merely as a means, nor are we even close to doing that, if either
>
> > (1) our treatment of this person is governed or guided in sufficiently

important ways by some relevant moral belief or concern,

or

(2) we do or would relevantly choose to bear some great burden for this person's sake.

For some moral belief to be *relevant* in the sense intended in (1), this belief must require direct concern for the well-being or moral claims of the person whom we are treating in some way. Suppose that some other slave-owner never whips his slaves because he believes that such acts would be wrong. But what would make such acts wrong, he believes, is not the fact that he would be inflicting pain on his slaves, but the fact that he would be giving himself sadistic pleasure. If that is why this man never whips his slaves, this fact would not count against the charge that he treats his slaves merely as a means. Another example is Kant's view that cruelty to animals is wrong because it dulls our sympathy, making us more likely to be cruel to other people. [195] If it is only this moral belief that leads some scientist to avoid causing her laboratory animals any pain, she would be treating these animals merely as a means.

Since relevance and importance are both matters of degree, it is often unclear whether (1) is true. Some other slave-owner might refrain from whipping his slaves because he cares about their well-being. But this concern, though relevant, would not govern this man's acts in a sufficiently important way. In a case that is less clear, when my mother traveled on a Chinese river in the 1930s, her boat was held up by bandits, whose moral principles permitted them to take, from ordinary people, only half their property. These bandits let my mother choose whether they would take her engagement ring or her wedding ring. If these people treated my mother as a means, they did not treat her *merely* as a means. Were they *close* to doing that? I am inclined to answer No. But this is a borderline case, in which this question has no definite answer. [196]

For condition (2) to be met, it is not enough that we would be prepared to bear some great burden for someone's sake. This fact may not be sufficiently relevant to the acts that we are considering. Consider some man who loves his wife, and who, in some disaster, would give up his life to save hers. It may still be true that, in much of this man's ordinary domestic life, he treats his wife merely as a means.

Whether we are treating someone *as a means* depends only on what we are intentionally doing. Whether we are treating someone *merely* as a means depends also, I believe, on our underlying attitudes or policies. And that is in part a matter of what we would have done, if the facts had been different. Return to our scientists who both use laboratory animals in their research. Suppose that, in one experiment, both these scientists use the most effective method, which causes their

animals no pain. Though these scientists are acting in the same way, the first scientist would still be treating her animals merely as a means, since it would still be true that she *would* have used the most effective method even if that would have caused her animals great pain. And the second scientist would *not* be treating her animals merely as a means, because she would not have acted in that other way. Consider next these claims:

> He treats her merely as a means.
>
> On this occasion, in acting as he did, he treated her merely as a means.

The first claim is more natural, and it is often clearer whether such claims are true.

It is wrong, Kant claims, to treat any rational being merely as a means. On a similar but wider view, it is wrong to treat any sentient or conscious being merely as a means. These views rightly imply that it is wrong to *regard* any rational or sentient being as a mere tool, whom or which we could treat as we please. But Kant's claim seems also to imply that, in treating anyone merely as a means, we would be *acting* wrongly.

That may not be true. Consider some gangster who, unlike my mother's principled bandits, regards most other people as a mere means, and who would injure them whenever that would benefit him. When this man buys a cup of coffee, he treats the coffee seller just as he would treat a vending machine. He would steal from the coffee seller if that was worth the trouble, just as he would smash the machine. But though this gangster treats the coffee seller merely as a means, what is wrong is only his attitude to this person. In buying his cup of coffee, he does not act wrongly.

Consider next some Egoist, who treats others in whatever way he believes would be best for him. Kant claims

> he who intends to make a lying promise. . . wants to make use of another human being merely as a means. [197]

We could similarly claim that, when this Egoist *keeps* some promise to someone whose help he will later need, he wants to make use of this other human being, and treats him merely as a means. Suppose next that this Egoist saves some child from drowning, at a great risk to himself, but that his only aim is to be rewarded. Since this man treats these other people merely as a means, Kant's principle implies that, in keeping his promise and saving this child's life, this man acts wrongly. That is clearly false.

To avoid such conclusions, we might claim that

> (3) we do not treat someone merely as a means if, as we know, our acts will not harm this person.

But suppose that, in

> *Mutual Benefit*, *Green* marries *Gold*, a 90-year old billionaire, to whom Green gives various services, and in other ways treats well. Green's sole aim, as Gold knows, is to inherit some of Gold's wealth. Though Gold would prefer genuine affection from Green, he accepts a mutually advantageous arrangement on Green's egoistic terms.

Suppose next that Green regards Gold as a mere tool, whom she would treat in whatever way would best achieve her aims. Green's first plan was to forge Gold's will and then murder him, and she changed her plan to marrying Gold, and treating him well, only because that seemed a safer way to get some of Gold's wealth. According to (3), since Green knows that her acts will not harm Gold, she is not treating Gold merely as a means. That claim is implausible. Though Green knows that her acts will not harm Gold, this fact makes no difference to her decisions. She would have murdered Gold if that had seemed a safer plan. We should admit, I believe, that Green treats Gold merely as a means.

If we cannot appeal to (3), Kant's view implies that Green acts wrongly. Perhaps we should accept that conclusion. But when my Egoist keeps his promises, or risks his life to save some drowning child, we should not claim that these acts are wrong. Our claim should be only that, given this man's self-interested motives, his acts do not have what Kant calls *moral worth*. [198]

To avoid condemning such acts, we might again revise Kant's view. According to

> *the Third Mere Means Principle*: It is wrong to treat anyone merely as a means, or to come close to doing that, if our act will also be likely to harm this person. [199]

In moving to this principle, we would be giving up the view that, if we treat someone merely as a means, or we are close to doing that, these facts are enough to make our act wrong.

I have discussed two ways in which, on Kant's view, we ought to treat all rational beings, or persons. We ought to follow the Consent Principle, by treating everyone only in ways to which they could rationally consent. And it is wrong to treat anyone merely as as a means. On our latest version of this second claim, such acts are wrong only if they are also likely to harm this person.

We can next connect these parts of Kant's view. We do not treat someone merely

as a means, nor are we even close to doing that, if our treatment of this person is governed or guided in sufficiently important ways by some relevant moral belief or principle. Kant's own example is the Consent Principle. We treat people as ends, Kant claims, and not merely as a means, if we deliberately treat these people only in ways to which they could rationally consent. [200]

Return now to

> *Lifeboat*: I am stranded on one rock, and five people are stranded on another. Before the rising tide drowns all of us, you could use a lifeboat to save either me or the five.

Consider also

> *Tunnel*: A driverless, runaway train is headed for a tunnel, in which it would kill the same five people. As a bystander, you could save these people's lives by switching the points on the track, thereby redirecting this train on to another track and through another tunnel. Unfortunately, as you know, I am in this other tunnel.
>
> *Bridge*: The train is headed for the five, but there is no other track and tunnel. I am on a bridge above the track. Your only way to save the five would be to open, by remote control, the trap-door on which I am standing, so that I would fall in front of the train, thereby triggering its automatic brake.

In all three cases, if you save the five, I would die. But my death would be differently causally related to your saving of the five. In *Lifeboat*, you would let me die because, in the time available, you could not save both me and the five. In *Tunnel*, you would save the five by redirecting the train with the foreseen side-effect of thereby killing me. In *Bridge*, you would kill me as a means of saving the five. I and the five, we should suppose, are all of about the same age, none of us is responsible for the threats to our lives, nor are there any other morally relevant differences between us.

It might be claimed that, in *Bridge*, you would not really be *killing* me as a means of saving the five. You would be merely using my body as a means of stopping the train, and you would be delighted if I survived. On this view, we kill someone as a means only when this person's death is an essential part of what achieves our aim. That might have been true, for example, of some medieval king's second son, who wanted to be the legitimate or rightful heir to his father's throne. Only his elder brother's death would achieve that aim. In a wider sense, however, we kill or injure someone as a means when we act in some way that foreseeably kills or injures this person, as a means of achieving some aim. That is how I shall use the phrase 'kill or injure as a means'.

Most people would believe that, in *Lifeboat*, you either may or ought to save the five. Some people would believe that, in both *Tunnel* and *Bridge*, it would be wrong for you to save the five. On this view, we have a duty not to kill which outweighs, or has priority over, our duty to save people's lives. Many other people would believe that, though our duty not to kill usually has such priority, that is not true in cases like *Tunnel*. On these people's view, it is not wrong to redirect some unintended threatening process---such as some flood, avalanche, or runaway train---so that it kills fewer people. Of those who hold this view, most would believe that you *would* be acting wrongly if, in *Bridge*, you killed me as a *means* of stopping the train and saving the five. There are also some people who reject these distinctions, believing that in all these kinds of case we ought to save as many lives as possible. My aim here is not to resolve this disagreement, but only to ask what is implied by the Kantian principles that we have been considering.

In *Lifeboat*, I have claimed, I could rationally consent to your saving the five rather than me. [201] If the choice were mine, I would have sufficient reasons to save my own life, but I would also have sufficient reasons to save the five rather than myself. Since I could also rationally consent to your saving the five, the Consent Principle would not condemn this act.

Similar claims apply to *Tunnel*. As before, if the choice were mine, I would have sufficient reasons to save either myself or the five. It would make no relevant difference that I would here be saving the five by redirecting the train so that it would kill me instead. This way of dying, we can suppose, would be no worse for me. Since I could rationally save the five by redirecting the train, I could also rationally consent to *your* acting in this way. So the Consent Principle would not condemn this act.

Similar claims apply to *Bridge*, in which you could save the five only by killing me. If the choice were mine, I would have sufficient reasons to jump in front of the train, so that it would kill me rather than the five. And compared to killing myself as a side-effect of saving the five, in *Tunnel*, it would be no worse for me, in *Bridge*, if I killed myself as a means of saving the five. Since I could rationally kill myself as a means of saving the five, I could also rationally consent to your treating me in this way.

It might be objected that I could not rationally consent to your killing me as a means, because this act would be wrong. But if I consented to this act, it would not be wrong. So even if this act would be wrong without my consent, that would not give me any reason to refuse consent.

Suppose next that, as I know, you accept the Consent Principle, and you always act upon it, so that this principle governs your acts. If I had the time, I might then think:

> According to this principle, it is wrong to treat people in any way to which they could not rationally consent.
>
> I could rationally consent to your killing me as a means of saving the five.

Therefore

> Even if I would not in fact consent, the Consent Principle does not condemn this act.
>
> We do not treat people merely as a means if our treatment of them is governed by the Consent Principle.

Therefore

> Since your treatment of me would be governed by the Consent Principle, you would neither be treating me merely as a means, nor be close to doing that, so no version of the Mere Means Principle would condemn this act.

This argument, I believe, is sound. It might be wrong for you to kill me, without my consent, as a means of saving the five. But that is not implied by these Kantian principles.

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

It may seem that, in making these claims, I must be misunderstanding or misapplying the Mere Means Principle. On one widely accepted view, which I shall call

> *the Standard View*, if we harm people, without their consent, as a means of achieving some aim, we thereby treat these people merely as a means, in a way that makes our act wrong.

This view involves, I believe, three mistakes. When we harm people as a means, we may not be treating these *people* as a means. Even if we *are* treating these people as a means, we may not be treating them *merely* as a means. And even if we *are* treating them merely as a means, we may not be acting wrongly.

Suppose first that, in

> *Attempted Murder*, when Brown attacks me with a knife, trying to kill me, I save myself by kicking Brown in a way that predictably breaks his leg.

Though I am *harming* Brown as a means of stopping him from killing me, I am not

treating *Brown* as a means. Just as we do not *use* falling rain when we wear raincoats to protect ourselves from being drenched, we do not *use* the people who attack us when we protect ourselves from their attack. We can add that, though I ought to treat *Brown himself* as an end and not merely as a means, I ought to *harm* Brown *merely* as a means and not even in part as an end, or for the sake of harming Brown.

It might be objected that, since harming someone is a way of treating this person, harming someone as a means must be a way of treating this person as a means. But this objection overlooks the difference between *doing* something to someone as a means and using *this person*. As I have said, when my doctor hurts me to find out whether my rib is broken, she isn't thereby using *me*. She isn't treating *me* as a means, I suggested, because she is hurting me for my own good and with my consent. Though I might be benefiting Brown by preventing him from committing murder, that is not the best way to explain why, in harming Brown as a means, I would not be using Brown. We might instead suggest that, since I am merely protecting myself from Brown's attack, my aims would be more easily achieved if Brown wasn't even there. If I was using Brown, I *would* want him to be there.

Turn next to the cases in which, when we harm people as a means, we *do* also treat these *people* as a means. On the Standard View, if we impose harm on someone as a means of achieving some aim, that is enough to make it true that we are treating this person *merely* as a means. To test this view, consider

> *Third Earthquake*: You and your child are trapped in slowly collapsing wreckage, which threatens both your lives. You cannot save your child's life except by using *Black's* body as a shield, without her consent, in a way that would crush one of her toes. If you also caused Black to lose another toe, you would save your own life.

Suppose you believe that it would be wrong for you to save your life in this way. Only the saving of a child's life, you believe, could justify imposing such an injury on someone else. Acting on this belief, you save your child's life by causing Black to lose only one toe. Since your act harms Black, without her consent, as a means of achieving your aim, the Standard View implies that you are treating Black merely as a means. But that is not true. If you were treating Black merely as a means, you would save your own life as well as your child's, by causing Black to lose two toes. We are not treating someone merely as a means if we are letting ourselves die rather than imposing a small injury on this person.

The Standard View might be revised. It might be suggested that, though you are not treating Black merely as a means, that is because you are limiting the harm that you impose on Black, in a way that is worse for you, or less effectively achieves your aims. No such claim would apply to your act, in *Bridge*, if you killed me as a

means of saving the five. You would not be limiting the harm that you imposed on me. And you would have acted in the very same way even if you had regarded me as a mere means. That may seem enough to justify the charge that, in acting in this way in *Bridge*, you would be treating me merely as a means. On this suggestion,

> (C) we treat someone merely as a means if
>
> > (1) we harm this person, without his or her consent, as a means of achieving some aim,
>
> unless
>
> > (2) we limit the harm that we impose, in some way that would or might be significantly worse for us, or make our act significantly less effective in achieving our aims.

This view is also, I believe, mistaken. We have supposed that, in *Third Earthquake*, you decide not to save your life by causing Black to lose a second toe. Suppose next that, just before you act, the situation changes, since the collapsing wreckage now threatens only your child's life. When you save your child's life by causing Black to lose one toe, you are not now limiting the harm that you impose on Black, so (C) implies that you are treating Black merely as a means. That is an indefensible conclusion. Rather than causing Black to lose a second toe, you would have let yourself die. That is enough to make it true that you are not treating Black merely as a means. It is irrelevant that you cannot now act in this way.

For another example, suppose that I am a soldier in some just war, fighting my way with my platoon through some occupied city. Before attacking the enemy soldiers in any building, I risk my death from sniper fire so that I can shout to these people, giving them a chance to surrender. If these people refuse my offer, and I kill or injure them as a means of capturing some building, (C) rightly allows that I am not treating these people merely as a means, since I have risked my life for their sake. Suppose next that the enemy soldiers in some building have already been given a chance to surrender, and have refused this offer. According to (C), if I kill or injure these people, I am treating them merely as a means. That is not true. I would have risked my life to give these people a chance to surrender. It is irrelevant that, on this occasion, I do not act in this way, because these people have already been given this chance. My attitude to all enemy soldiers is the same, and I treat none of them merely as a means.

Similar claims apply to *Bridge*. Suppose that you use remote control to cause me to fall onto the track, so that my body would stop the runaway train. Your aim is to ensure that the five will be saved. You also try, however, to save my life by running to the track, so that you can jump in front of the train, thereby stopping it

before it reaches me. If your attempt succeeds, you would not be treating me merely as a means, since you would be killing yourself for my sake. It would make no relevant difference, I believe, if you failed to reach the track in time. Nor would it make such a difference if, though you would have sacrificed your life to avoid killing me, this was never possible. In both versions of *Bridge*, your act may be wrong. And if it is, what makes it wrong may be the fact that you would be *killing* me *as a means* of saving the five. But you would not be treating *me* as a *mere* means.

I have rejected the standard account of what is involved in treating people as a mere means. Some writers give other accounts. For example, O'Neill writes:

> if we coerce or deceive others. . . we do indeed use others, treating them as mere props or tools in our own projects. . . a maxim of deception or coercion treats another as mere means. . . [202]

Korsgaard similarly writes:

> Coercion and deception are the two ways of using others as mere means. [203]

But suppose that, in a variant of *Attempted Murder*, I stop Brown from killing me by threatening to shoot him, or by falsely telling him that the police will soon arrive. Though I would be coercing or deceiving Brown, I may not be treating Brown as a mere means. I may be coercing or deceiving Brown because these are the only ways in which, without harming Brown, I could stop him from killing me. Suppose next that, in

> *Desperate Plight*, you and I are in some diving bell which is caught on the ocean's floor. Though we cannot hope to be rescued in less than ten hours, we have enough oxygen to keep two people alive for only six or seven hours. So, as I know, unless one of us dies soon, we shall both die. I start acting in some way that will kill me and thereby save your life. When you try to stop me, I coerce you or deceive you so that your attempt fails.

Though I am coercing or deceiving you, I am not treating you as a mere means. As before, we are not treating someone as a mere means if we are sacrificing our life for this person's sake.

When O'Neill explains her claim that deception and coercion treat others as a mere means, she writes

> To treat something as a mere means is to treat it in ways that are appropriate to things. [204]

Deception and coercion are not, however, appropriate ways of treating things, since neither is even possible.

On Kant's view, Korsgaard also writes,

> Any attempt to control the actions and reactions of another by any means except an appeal to reason treats her as a mere means. . . [205]

This claim implies that whenever people in positions of authority tell us to do something---such as to show them our train ticket, or fill out a customs declaration, or fasten our safety-belts---they are treating us as a mere means. That is not true. Korsgaard also writes that, on Kant's view, we treat others as a mere means whenever 'we do something that only works because most other people don't do it'. [206] But when poor people feed themselves with the scraps that others throw away, they do not treat these other people as a mere means.

Suppose next that, in

> *Bad Samaritan*, while driving across some desert, I see you lying injured by the road, needing help. I ignore you, and drive on.

According to some writers, Kant would claim that I am here treating you merely as a means. That claim would be false. In ignoring you, I am not using you in any way, so I cannot be merely using you.

These writers might reply that, when Kant uses the phrase 'merely as a means'---or, more accurately, its German equivalent---Kant does not use this phrase in its ordinary sense. Kant often uses words in special senses. When I drive past you, ignoring your need for help, it might be true that, in Kant's special intended sense, I am treating you merely as a means. O'Neill and Korsgaard might similarly claim that all deception and coercion does, in Kant's special sense, treat people merely as a means.

We are sometimes justified in using words in something other than their ordinary senses. For example, it can be worth stretching the sense of 'painful', so that it applies to unpleasant sensations, such as nausea. By using 'painful' in this wider sense, we avoid the need to keep writing 'painful or unpleasant', and the distinction that we are ignoring seldom matters. Some unpleasant sensations are much worse to have than some pains. It is often a mistake, however, to use words in special senses. We may then make claims that are misleading and only seem to be important. For example, Rawls suggests that, if we accept his Contractualist moral theory, we should use 'right' to mean: in accordance with the principles that would be chosen by his imagined contractors. [207] That would make it trivial to claim that acting in accordance with these principles is right. Rawls also suggests that we could call these principles 'true' in the sense that they would be chosen by these

contractors. [208] That would make it trivial to claim that these chosen principles are true. [209]

If we believe that Kant uses 'merely as a means' in some special sense, we ought not to say that, on Kant's view, we must never treat people merely as a means. If that is what we say, our hearers may take us to be claiming that, on Kant's view, we must never treat people merely as a means. To avoid being misunderstood, we should use some other phrase. We might say that, on Kant's view, we must never treat people in certain ways, which we shall call treating people *shmerely as a means*. We could then explain what we use this new phrase to mean.

The phrase 'merely as a means' has, I believe, an ordinary sense that is both fairly clear, and morally significant. Though Kant may sometimes use this phrase in a special sense, [210] he also uses it, I believe, in the ordinary sense. It is not misleading to say that, according to Kant's Formula of Humanity, we must never treat people merely as a means. And this is the version of Kant's formula that is most worth discussing.

On my rough definition of this ordinary sense, we treat someone merely as a means if we both use this person in some way and regard her as a mere tool, someone whose well-being and moral claims we ignore, and whom we would treat in whatever way would best achieve our aims. We do *not* treat someone merely as a means, nor are we even close to doing that, if either (1) our treatment of this person is governed in a sufficiently important way by some relevant moral belief, or (2) we do or would relevantly choose to bear some great burden for this person's sake.

When people give other definitions, they are often trying to make Kant's claim cover a wider range of acts. That can sometimes be done, I have suggested, not by using 'merely as a means' in some special sense, but by revising Kant's claim so that it also condemns acts that are *close* to treating people merely as a means. And rather than stretching Kant's claim so that it covers other kinds of act, we should sometimes make other, similar claims. When Bad Samaritans ignore someone who needs urgent help, they do not treat this person as a mere means. But they do treat this person as a *mere thing*, something that has no importance, like a stone or heap of rags lying by the road. That, we could claim, is just as bad. And there are ways of treating people that are worse than treating them as a mere means. Though Hitler treated the Slavs in his conquered Eastern territories as a mere means, that is not how he treated the Jews.

32 Harming as a Means

We can now return to the question of whether, as Kant claims, it is wrong not only to

regard people merely as means, but also to *act* in ways that treat them merely as a means.

Kant's claim, as we have seen, is too strong. When my gangster buys his cup of coffee, he treats the coffee seller merely as a means, but though this man's attitude is wrong he is not acting wrongly. Nor does my Egoist act wrongly when he risks his life to save a drowning child, though he is using this child as a mere means of getting some reward. [211]

To meet such objections, as I have said, we can revise Kant's claim. According to

> *the Third Mere Means Principle*: It is wrong to act in any way that treats anyone merely as a means, or comes close to doing that, if our act will also be likely to harm this person.

But we ought, I believe, to reject this principle. Let us again compare

> *Lifeboat*, in which you could save either me or the five,
>
> *Tunnel*, in which you could redirect a runaway train so that it kills me rather than the five,

and

> *Bridge*, in which you could save the five only by killing me.

According to one view, in all three cases, you ought to save the five. It makes no difference whether, in saving the five, you would be killing me. When people's lives are threatened, we ought to do whatever would save the most lives.

According to a second view, you ought to save the five only in *Lifeboat*. We have a duty not to kill which outweighs our duty to save people's lives. On this view, it would be wrong for you to save the five in both *Tunnel* and *Bridge*, since these ways of saving the five would both kill me. As before, it makes no difference whether you would be killing me as a means.

According to a third view, you ought to save the five in *Lifeboat*, and you would be at least permitted to save the five in *Tunnel*, but it would be wrong for you to save the five in *Bridge*. This, I believe, is the most widely held of these three views. On this view, it *does* make a difference whether you would be killing me as a means.

If we accept this third view, we might appeal to

> *the Harmful Means Principle*: It is wrong to impose harm on someone as a means of achieving some aim, unless

(1) there is no better way to achieve this aim,

and

(2) given the goodness of this aim, the harm we impose is not disproportionate, or too great.

This principle does not tell us which harms would be too great. We would have to use our judgment here. On one view, there is an upper limit on the amount of harm that we could justifiably impose on someone as a means. According to Judith Thomson, for example, it would be wrong to kill or seriously injure one innocent person, however many other people's lives we could thereby save. [212] Most of us would accept a less extreme view. We would believe it to be right to kill one innocent person if that were the only way in which we could prevent some nuclear explosion that would kill as many as a million other people. But we may believe it to be wrong to kill one person as a means of saving only five, or only fifty other people. There would be cases in between in which this moral question would have no clear or determinate answer.

On what I have called the Standard View, if we harm someone, without this person's consent, as a means of achieving some aim, we thereby treat this person merely as a means. As I have argued, that may not be true. When I break Brown's leg to stop him from murdering me, I am *harming* Brown as a means of defending myself. But I am not treating *Brown himself* as a means, so I cannot be treating Brown merely as a means.

Return next to cases in which, if we impose harm on someone as a means, we may also be treating this person as a means. When we ask whether such an act would be wrong, we have two questions:

Q1: Might the wrongness of this act partly depend on whether we would be harming this person as a means of achieving some aim?

Q2: Might the wrongness of this act partly depend on whether we would also be treating this person *merely* as a means?

When we compare cases like *Bridge* and *Tunnel*, we may decide that the answer to Q1 is Yes. We may believe that, though you could justifiably redirect the runaway train so that it would kill me rather than the five, it would be wrong for your to save the five *by* killing me. I have *not* been arguing against this view.

The answer to Q2, I believe, is always or nearly always No. If you killed me in *Bridge* without my consent, you might not be treating me merely as a means, or be close to doing that. Your treatment of me might be governed in a sufficiently important way by some relevant moral principle, such as Kant's Consent Principle.

And it might be true that, if you had been closer to the train, you would have saved the five by killing yourself rather than me. But these facts would not, I believe, affect whether your act would be wrong. If it would be wrong for you to kill me as a means of saving the five, this act would be wrong *whether or not* you would also be treating me merely as a means. Even if you were *not* treating me merely as a means, and were not even close to doing that, these facts would not justify your act.

Turn next to cases in which we *could* justifiably impose harm on someone as a means. In *Third Earthquake*, you cannot save your child's life except by crushing Black's toe, without Black's consent. This act, I believe, would be justified. If someone crushed my toe to save their child's life, I would not (I hope) complain. Though some people would believe this act to be wrong, these people would accept that there are some lesser harms that we could justifiably impose on someone, if that was our only way to save someone else's life. On Thomson's view, for example, we could permissibly save someone's life by bruising someone else's leg, causing this other person 'a mild, short-lasting pain'. [213] So we can suppose that, in

> *Fourth Earthquake*, my gangster cannot save his child's life except by bruising Black's leg, without her consent, causing her a mild, short-lasting pain.

This gangster regards Black as a mere means. He would kill or gravely injure Black if that would help him to achieve any of his aims. So if this gangster saved his child by bruising Black's leg, he would both be imposing harm on Black and be treating Black merely as a means. According to Kant's Formula of Humanity, which includes the Mere Means Principle, it is wrong to act in any way that treats people merely as a means. According to the Third Mere Means Principle, it is wrong to impose harm on people in any way that also treats them merely as a means. These principles both imply that, if my gangster saved his child's life by bruising Black's leg, he would be acting wrongly.

That is an unacceptable conclusion. Though this gangster has the wrong attitude to Black, he could justifiably save his child's life by imposing this small harm on Black. This child has a moral claim to be saved; and her claim is not undermined, or overridden, by the wrongness of her father's attitude to Black. Similar claims apply to other cases. If you would be morally permitted to save your child in *Third Earthquake* by causing Black to lose one toe, my gangster would be morally permitted to save his child in the same way. [214]

It has been widely believed that, to explain the wrongness of harming some people as a means of benefiting others, we could appeal to Kant's claim that we must never treat people merely as a means. This belief, I have argued, is mistaken. If it would be wrong to impose certain harms on people *as a means* of achieving certain

good aims, these acts would be wrong even if we were *not* treating these people *merely* as a means. And if it would *not* be wrong to impose certain lesser harms on people as a means of achieving such aims, these acts would not be wrong even if we *were* treating these people merely as a means.

Kant's claim contains an important truth. It is wrong to *regard* anyone merely as a means. But the wrongness of our *acts* never or hardly ever depends on whether we are treating people merely as a means.

CHAPTER 10 RESPECT AND VALUE

33 Respect for Persons

In another comment on his Formula of Humanity, Kant writes

> every rational being. . . must always be regarded as an end. . . and is an object of respect. [215]

This requirement to respect all persons is one of Kant's greatest contributions to our moral thinking. But it does not tell how we ought to act.

Allen Wood suggests that

> (A) we must always treat people in ways that express respect for them. [216]

We can treat people rightly, however, without *expressing* our respect for them. Wood suggests that, whenever we treat people rightly, our acts could be taken to express respect for these people. [217] But on this suggestion (A) would tell us only that we must always treat people rightly. (A) would not help us to decide which acts are right, since we could not decide whether some act would express respect for people except by deciding whether this act would be right.

Some writers suggest that

> (B) it is wrong to treat people in ways that are incompatible with respect for them.

Some wrong acts are clearly incompatible with respect for persons. Kant's examples are: disgraceful or humiliating punishments, ridicule, defamation, and acts that display arrogance or contempt. [218] But Kant's formula is intended to cover all wrong acts, and most wrong acts do not treat people in such disrespectful ways.

All wrong acts, some writers suggest, are in a wider sense incompatible with respect for persons. On this suggestion, (B) would not be a useful claim. As before, to decide whether some act would be in this wider sense incompatible with respect for persons, we would first have to decide whether this act would be wrong. If this act would *not* be wrong, it would be compatible with respect for persons. As both Kant and Sidgwick warn, moral philosophers often make claims that seem to give us 'valuable information' but really tell us only that acts are wrong if they are wrong.

Kant also claims that

> (C) we must always respect *humanity*, or the 'rational nature' that makes us persons.

Wood calls (C) 'the most useful formulation' of Kant's supreme principle of morality. [219] Though (C) cannot directly solve all moral problems, this principle provides, Wood claims, 'the correct basis for deciding moral questions'. [220] To support this claim, Wood points out that in his last and longest book about morality, Kant often makes remarks that seem to appeal to (C). [221]

Kant's remarks do not, I believe, show (C) to be a useful principle. As Wood himself concedes, Kant's appeals to (C) are 'usually both brief and casual'. [222] Such remarks add little to Kant's view. For example, Kant writes that our duty to develop our talents 'is bound up with the end of humanity in our own person'. [223] Kant makes other claims that Wood rightly rejects. It would be wrong, Kant claims, for any of us to give ourselves sexual pleasure, or to hasten our deaths to avoid suffering, because such acts debase or defile humanity. [224] And when he condemns telling some lie even 'to achieve some really good end', Kant writes that any liar 'violates the dignity of humanity in his own person', so that he becomes a 'mere deceptive appearance of a human being', who has 'even less worth than if he were a mere thing'. [225] These are not the claims that make Kant the greatest moral philosopher since the ancient Greeks.

Wood suggests that, in making these claims, Kant misapplies (C). We can reject Kant's views about sex, suicide, and lying, Wood writes, 'because we justifiably believe that we know more about what respect for humanity requires in these matters'. It is 'an advantage' of this principle 'that both sides in profound moral disagreements can use it to articulate what they regard as their strongest arguments'. [226]

This assessment seems to me mistaken. When Kant claims that certain acts would violate or debase humanity, and we reject these claims, neither Kant nor we are giving our strongest arguments. Nor would (C) help us to decide, in difficult cases, which acts would be wrong.

34 Two Kinds of Value

When Kant explains the sense in which we must always treat rational beings as ends, he claims that such beings have *dignity*, by which he means a kind of supreme value. This claim raises one of the deepest questions in ethics: that of how what is good is related to what is right, or to what we ought morally to do. Kant also claims that, rather than following the ancient Greeks by first asking which ends are good and

then drawing conclusions about which acts are right, we ought to reverse this procedure. Rawls calls it a central feature of Kant's moral theory that 'the right' is, in this way, 'prior to the good'. [227] But Wood in contrast claims that, though Kant's Formula of Humanity 'takes the form of a rule or commandment, what it basically asserts is the existence of a substantive value.' [228] And Herman suggests that Kant's 'fundamental theoretical concept' is 'the Good', and that 'Kant's ethics is best understood as an ethics of value'. [229]

Before we consider Kant's claims about value, it will help to draw some more distinctions. Many things are good or bad in what I have called *reason-implying* senses. Such things have certain kinds of properties or features that would, in some situations, give us or others reasons to respond to these things in certain ways. [230]

Some of these good things have a kind of value that, as Scanlon and others say, is *to be promoted*. Two examples are happiness and the relief or prevention of suffering. When things have this kind of value, it is really these things, not their value, that we have reasons to promote.

What we can promote are events, in the wide sense of 'event' that also covers acts and states of affairs. Events can be good or bad either as an end or as a means to some end. On some views, acts can be good or bad only as a means. We ought, I believe, to reject such views. We act well, for example, if we bring up our children well, or we act as good friends or lovers, or we engage with some success in various other worthwhile activities, or we act rightly and treat people with respect. Such things might be worth doing, not merely as a means to happiness or other good ends, but partly or wholly for their own sake. So we should include acts among the events that might be good or bad as ends.

On what seems to me the best view about the goodness of events, which I shall call

> *Actualism*: Possible acts and other events would be good as ends when they have intrinsic properties or features that give us reasons to want them to be actual, by being done or occurring, and to make them actual if we can. Possible acts and other events would be good as a means when our making them actual would be an effective way of achieving some end. [231]

Similar claims apply to events that would be bad as ends, or bad as a means to some end. Events may be good as ends either for particular people or in the impartial-reason-implying sense, or both. As well as having reasons to try to cause or prevent good or bad events, we have reasons to have various other attitudes towards them, such as hope, gladness, fear, and regret. These are all attitudes towards the possibility or fact that such events are actual or real, being a part of the way things

go.

Since Actualism applies to all possible acts and all of their possible effects, this view covers everything whose goodness is directly relevant to any decision about what we should do. We have a reason to act in some way if and only if, or just when, this act would be in some way good either as an end, or as a means to some good end. Actualism does not, however, claim to cover the goodness of things that are not acts or other events.

According to some writers, this view can be widened to cover the goodness of some persisting things, such as people and works of art. Such things are claimed to be good when their nature gives us reasons to want them to exist, or continue to exist, and reasons to make that happen if we can. G. E. Moore even writes:

> when we assert that a thing is good, what we mean is that its existence or reality is good. [232]

But these claims are mistakes. Something's existence can be good though this thing itself is not good, and *vice versa*. There are many bad people, for example, whose continued existence would be good as an end. When some good person is dying a slow and painful death, the continued existence of this person may be bad as an end. And there would be nothing good in the continued existence of good works of art if no one could ever see them.

According to what Scanlon calls *teleological* theories, it is only acts and other events that have *intrinsic* value in the sense of being in themselves good. Scanlon rightly rejects this claim. There are other things that can be in themselves good, such as people, books, and arguments. Since these things are not events, we cannot want them to happen, or make them happen. But we can respond to them in other ways. We can have reasons to read good books, be convinced by good arguments, and try to become more like good people.

We can now turn to a kind of value which, as Scanlon and others say, is to be *respected* rather than promoted. As before, when things have such value, it is really these things, not their value, that we have reasons to respect. Though people are the best example of what can be claimed to have such value, we can start with some other examples. These can be things that are claimed to have symbolic, historical, or associational value, such as our nation's flag, the oldest living tree, icons and other religious paintings, and the bodies of dead people.

Understanding something's value, Scanlon writes, is in part 'a matter of knowing *how* to value it---knowing what kinds of actions and attitudes are called for.'[233] Many of these acts and attitudes can be loosely called ways of respecting or

honouring this thing. We might respect our nation's flag, the oldest tree, and some religious painting by refusing to use these things as a dishcloth, firewood, and the target in a game of darts. To respond appropriately to the value of many such things, we ought to protect them, so that they continue to exist. But that is not always true. We can respond appropriately to the value of dead people's bodies, not by trying to preserve them as the ancient Egyptians did, but by destroying them in some respectful way, such as burning them bedecked with flowers on some funeral pyre, rather than throwing them onto some rubbish dump.

The value of such things is quite different from the goodness of good ends, or good people. It is not a kind of *goodness*. Though some dead people's bodies would be good as *cadavers*, for use in teaching anatomy or surgery, and some other bodies would be good as corpses in some horror film, these are not the kind of value that all dead people's bodies can be claimed to have. And some religious paintings are not good. Though this kind of value is not a kind of goodness, and is not a value that is to be promoted, when we could respond to the value of such things by treating them in respectful ways, these *acts* would be good as ends, having the kind of value that is to be promoted. [234]

We can turn next to claims about the value of human life. Appreciating this value, Scanlon writes,

> is primarily a matter of seeing human lives as something to be respected, where this involves seeing reasons not to destroy them, reasons to protect them, and reasons to want them to go well. [235]

To see that we have such reasons, however, we don't need to see human lives as having a kind of value that is to be respected *rather* than promoted. When people's lives go well, that is both good for these people and impersonally good, in the reason-implying senses. Such happy and well-lived lives are good as ends. We have reasons to protect the living of such good lives, and to help these people in other ways to make their lives go well.

On some views, human life has a different kind of value. Suppose that you have begun to die a slow, painful, and undignified death, and you have nothing important left to do. You may have strong reasons to kill yourself, and other people may have strong reasons to help you to act in this way. Of those who appeal to the value of human life, some would believe that this act would be wrong. These people might agree that it would be both better for you, and impersonally better, if you died an earlier, natural death. But you ought not to kill yourself, these people believe, and other people ought not to help you, since such acts would fail to respect the value of human life. On this view, respecting the value of someone's life is not the same as, and may conflict with, doing what would both be best for this person and be what this person chooses.

Scanlon rejects this view. We have reasons not to end someone's life, he writes, only 'as long as the person whose life it is has reason to go on living or wants to live'. [236] Scanlon here denies that a person's life has the kind of value that we ought to respect in ways that conflict with this person's well-being and autonomy. This, I believe, is the right view about the value of human life. To defend the claim that suicide and assisting suicide would be, in such cases, wrong, we would need some other argument. [237]

It is not human life but the people who *live* these lives who should be claimed to have the kind of value that should be respected rather than promoted. We should respect this value, Scanlon claims, by treating people only in ways that could be justified to them. Kant similarly claims that, to respect people, we should treat them only in ways to which they could rationally consent.

35 Kantian Dignity

We can next consider Kant's claims about value. While making these claims, Kant distinguishes three kinds of end. What Kant calls *ends-to-be-produced* are the aims or outcomes that we could try to achieve or bring about. These are ends in the ordinary sense, as in the claim that the relief of suffering is a good end. Kant contrasts such ends with what he calls *existent* or already existing ends, of which his main examples are rational beings, or people. Kant's third kind of end he calls *ends-in-themselves*. Such things have what Kant calls *dignity*, which he defines as absolute, unconditional, and incomparable value or worth. [238] Such value is supreme, or unsurpassed, in the sense that nothing else has greater value.

According to some writers, Kant believes that such supreme value is had only by some existent ends, such as rational beings, whose value is of the kind that is to be respected rather than promoted. But there are several ends-to-be-produced which Kant claims to have supreme value, and to be ends that we ought to try to promote, or achieve.

One such end is having a *good will*. Our will is good, Kant claims, when we do our duty because it is our duty, and not with some other aim, such as avoiding punishment. Our having a good will can be taken to be either a mental state or disposition, or an activity which consists in good willing. [239] Regarded in either way, having a good will is something that, on Kant's view, we ought to try to achieve. In Kant's own words, 'the true vocation of reason must be to produce a will that is good.'[240]

Another end-to-be-produced with supreme goodness is what Kant calls the *Realm of Ends*. This is the possible state of affairs, or *possible world*, that we together would

produce if everyone had good wills and always acted rightly. [241]

A third such end is what Kant calls the *Highest* or *Greatest Good*. [242] This possible world is the Realm of Ends with the further feature that everyone would have all of the happiness that their virtue would make them deserve. [243] Kant claims that 'we ought to try to promote' this end, and that 'reason. . . commands us to contribute everything possible to its production.'[244]

There may be a fourth such end. Kant calls rational beings 'something whose existence in itself has absolute worth'. [245] And he writes that, if there were no rational beings, the Universe would be 'a mere waste, in vain, without a final purpose'. [246] These remarks suggest that, on Kant's view, the continued existence of rational beings is another end-to-be-produced with supreme value. [247]

We can now return to Kant's claim that rational beings or people are ends-in-themselves, who have dignity, or supreme value. As I have said, people are not ends-to-be-produced. And their value is of a different kind. On Kant's view, as Wood and Herman claim, 'even the worst human beings have dignity',[248] and a person whose will is good 'is of no greater value' than someone with an ordinary or a bad will. [249] This part of Kant's view is, I believe, a profound truth. But the value of the morally worst people is not a kind of goodness. Hitler and Stalin were not good. People have dignity or value in the quite different sense that, given their nature as rational beings, they must always be treated in certain helpful or respectful ways. A similar claim applies, I believe, to all sentient beings. Even the lowliest worm, if it can feel pain, has a kind of dignity, in an extended Kantian sense. A worm cannot be in itself good, but its nature makes it a being on which it would be wrong to inflict pointless pain.

I have been ignoring one complication. Kant sometimes uses 'humanity' to refer to rationality, or what he also calls 'rational nature'. So, when Kant claims that humanity is an end-in-itself with dignity, or supreme value, he might mean that rationality has such value. And though the value of rational beings is not a kind of goodness, their being rational might be claimed to be good. Herman writes that, in Kant's ethics, 'The domain of "the good" is rational activity and agency,' and that Kant 'grounds morality' on 'rationality as a value'. [250] Wood even calls Kant's claim about rationality's value 'the most fundamental proposition in Kant's entire ethical theory'.[251]

On Kant's view, like having a good will, rationality is in part an end-to-be-produced, or promoted. We ought to use our rationality, and we can try to become more rational by developing our rational abilities. Kant calls *dignity* a value that is 'infinitely far above' a lower kind of value, which he calls *price*. [252] Among the

things that have mere price Kant includes pleasure and the absence of pain. So, if Kant meant to claim that rationality or rational activity had dignity, Kant's view would imply that rationality has infinitely greater value than the relief of pain. Cardinal Newman claims that, though both sin and pain are bad, sin is infinitely worse, so that, if all mankind suffered extremest agony, that would be less bad than if one venial sin were committed. [253] Though this view is horrific, we can understand why it has been held, since we can see how sin might seem infinitely worse than pain. If rationality or rational activity had dignity in the sense of infinite value, and preventing pain had only finite value, Kant's view would have implications that would be even harder to accept. On this view, for example, we ought to increase our ability to play chess, or to solve crossword puzzles, rather than saving any number of other people from any amount of pain. That conclusion would be insane.

It might be objected that, even on this view, we ought to save these other people from pain, since that would help them to act rationally. But we might be saving these people from pain during surgical operations, by making them unconscious. That would not help them to act rationally.

It might next be claimed that rationality's value is of the kind that is to be respected rather than promoted. That is not Kant's view, since Kant often claims that we ought to try to develop and use our rational abilities. And this revised version of Kant's view would face a similar objection. We respect the value of persons, not by adding new people to the world, but by following various other moral requirements, such as the requirement not to kill or injure people. If rationality had similar value, as Thomas Hill points out, there would be similar requirements not to damage or impair people's rational abilities. And if rationality's value was infinitely far above all price, it would be wrong to 'trade' or 'sacrifice' any rational ability for the sake of anything with mere price, such as relief from pain. [254] So it would be wrong for us to damage our ability to play chess or solve crossword puzzles, even if that would be one effect of our saving any number of people from any amount of pain. That conclusion would also be insane.

Kant's view does not, I believe, have such implications. When Kant claims that humanity has dignity, he is seldom referring, I believe, to rationality. Kant distinguishes between (1) our capacity for acting morally and having a good will, and (2) our other rational capacities and abilities. We can call (2) our *non-moral rationality*. Just after defining dignity as a kind of absolute and incomparable value, Kant writes:

> morality, and humanity insofar as it is capable of morality, is that which alone has dignity. [255]

The word 'humanity' cannot here refer to non-moral rationality. In many other

passages, Kant distinguishes between ourselves and what he calls 'the humanity in our person'. These uses of 'humanity' mostly refer, I believe, not to our rationality, but either to our capacity for acting morally and having a good will, or to ourselves as what Kant calls *noumenal* beings. Though some of Kant's remarks suggest that non-moral rationality is an end-in-itself, with supreme value, he is not, I believe, committed to this view. Kant is 'the least exact of the great thinkers', [256] and his uses of 'humanity' are shifting and vague. Kant does condemn some vices, such as gluttony and drunkenness, on the ground that such vices interfere with our rational activities or abilities. [257] But Kant's main claims do not imply that it would be wrong for us to eat too much, or to make ourselves drunk, even if these were the only ways of saving any number of people from any amount of pain.

In his claims about value, Herman writes, Kant provides 'a radical critique of traditional conceptions'. [258] On Kant's view, 'past moral philosophy . . . mistakes the nature of the good'. [259]

Kant does not, I believe, provide such a critique. If Kant claimed that nothing has the kind of value that is to be promoted, he would be rejecting many earlier views. But as we have seen, Kant claims that such value is had by our having good wills, and by the Realm of Ends, and by Kant's Greatest Good, the possible state of affairs or world in which everyone would be virtuous and happy. On Kant's view, these are all ends-to-be-produced, which we ought to promote as much as we can. In his claims about which things have such value, Kant also follows earlier philosophers, many of whom claim that virtue and happiness are the two things that are good as ends.

Kant may not accept one widely held view about value, since he often ignores the reason-implying senses in which things can be non-morally good or bad. He claims for example, that the principle of prudence, or of doing what would promote our own happiness, is a merely *hypothetical imperative*, which applies to us only because we want to be happy. [260] Kant here ignores our non-moral reasons to want to be happy. In his account of practical reason, Kant describes morality and instrumental rationality, with little but a wasteland in between. Kant's ignoring of non-moral goodness, which I discuss in Appendix G, is not, however, a critique.

There is another widely held view that Kant may not accept. On this view, to be valuable is always to be in some way good. [261] When Kant claims that all rational beings have the kind of value that he calls dignity, he does not mean that all rational beings are good. As I have said, Kant means that all rational beings have a kind of value that is to be respected, since these beings ought to be treated only in certain ways. This value is a kind of *status*, or what Herman calls 'moral standing.' [262] Such value is ignored by many traditional views.

Kant, I believe, is right to claim that even the morally worst people have the same moral status as anyone else. And by calling this status *dignity* or *supreme value*, Kant expresses this claim in a helpfully persuasive way. But for the idea of moral status to be theoretically useful, it must draw some distinction, by singling out, among the members of some wider group, those who meet some further condition. In Roman law, to give one analogy, only those human beings who were not slaves had full legal status, and counted as persons. In democracies, only those persons who are adults have the status of being entitled to vote, and in many countries only those persons who are citizens have the status of being entitled to certain benefits. On Kant's view, in contrast, *all* rational beings or persons ought to be treated only in certain ways. We add little if we say that all rational beings or persons have the moral status of being entities who ought to be treated only in these ways.

Kant's claims about value are also, in one way, misleading. As I have said, when Kant claims that all rational beings have dignity, or supreme value, he does not mean that all such beings are good. But Kant claims that such supreme value is also had by morality, good wills, the possible worlds which are the Realm of Ends, and the Greatest Good. The value of these things, on Kant's view, *is* a kind of goodness. So, in his claims about value, Kant fails to distinguish between being supremely good and having a kind of moral status that is compatible with being, like Hitler and Stalin, very bad. It is easy, however, to add this distinction to Kant's view.

36 The Right and the Good

The Highest or Greatest Good, Kant claims, would be a world in which everyone was both wholly virtuous, or morally good, and had all of the happiness that their virtue would make them deserve. [263] Kant also writes:

> Everyone ought to strive to promote the Greatest Good. [264]
>
> the moral law commands me to make the greatest possible good in a world the final object of all my conduct. [265]

According to what we can call this

> *Formula of the Greatest Good*: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

This ideal world would be hard to achieve. So, in applying this formula, we should compare unideal but more achievable states of the world, and ask how we could get as close as possible to Kant's ideal.[266]

It would be best, Kant claims, if everyone's degree of happiness was *in proportion* to their degree of virtue, or worthiness to be happy. That would be true in the ideal world in which we would all be wholly virtuous and happy. Some writers suggest that, of the worlds that are not ideal, the best would be those in which this *proportionality condition* would be met. [267] But this seems unlikely to be Kant's view. Everyone's happiness might be in proportion to their virtue if no one was either virtuous or happy, or if everyone was both vicious and miserable. These worlds would clearly be much worse than worlds in which everyone had great virtue and great happiness, but some people had slightly less or slightly more happiness than they deserved. So we can assume that, on Kant's view, it would always be better if there was more virtue, and more deserved happiness, even if the proportionality condition would be less well met.

Kant claims, implausibly, that no one can affect how virtuous other people are. On this assumption, we can promote virtue only by increasing our own virtue. We can best do that by trying to have good wills, and doing whatever else we ought to do. We can best promote deserved happiness by trying to give happiness to people who are less happy than they deserve. It is often claimed that we cannot act in this way, since we cannot know how much happiness people deserve. We do not, however, need *knowledge*. It would be enough to have rational beliefs about which people are more likely to deserve more happiness. As Kant assumes, we often have such beliefs. [268] We could act on these beliefs by trying to make these people happier. So Kant's Formula of the Greatest Good gives us an aim that we could try to achieve.

We can next draw some more distinctions, and introduce some of Kant's other claims. Moral theories are in one sense

> *Act Consequentialist* if they claim that everyone ought always to do, or try to do, whatever would best achieve one or more common aims.

According to one such theory, *Hedonistic Act Utilitarianism* or

> *HAU:* Everyone ought always to produce, or try to produce, the greatest possible amount of happiness minus suffering.

These theories are *person-neutral* in the sense that they give the same common aims to everyone. According to most moral theories, and most people's moral beliefs, there are some common aims that everyone ought to try to achieve, such as the aim that people be saved from starving. But each of us ought also to try to achieve many *person-relative* moral aims. On such views, for example, rather than having the common aims that promises be kept and children be cared for, each of us ought to try to keep our own promises, and to care for our own children. A third group of

views do not give us any common moral aims. That is true, for example, of the view that our only duties are to obey the Ten Commandments.

Some moral theories are wholly or partly *value-based*, in the sense that they appeal to claims about what is good or bad, in some significant, substantive sense. According to what we can call *Value-based Act Consequentialism*, or

> *VAC*: Everyone ought always to do, or try to do, whatever would make things go best.

On this version of HAU, for example, everyone ought to produce, or try to produce, the greatest net sum of happiness because that is how we could make things go best.

As well as making claims about what is good and what we ought morally to do, some moral theories make claims about how the concept *good* is related to the moral version of the concept *ought*. According to some theories, the concept *good* is fundamental, and can be used to define this version of the concept *ought*. According to some other theories, it is the concept *ought* that is fundamental, and can be used to define the concept *good*. According to a third group of theories, neither of these concepts can be defined in terms of the other. The best theories, I believe, are of this third kind. Because these are the only theories that use *good* and *ought* in senses that are independent, these are the only theories that can make true substantive claims about the relations between what is good and what we ought morally to do.

As one example of the first kind of theory, we can take G. E. Moore's *Principia Ethica*. Moore claims that, when we say that

> we *ought* to do something, we mean that this act would do the most good, by making things go best. [269]

We can call this the *goodness-promoting* sense of 'ought'. Moore also claims

> M1: Everyone ought always to do what would make things go best.

This claim may seem to be a version of Value-based Act Consequentialism. But if Moore is using 'ought' in his goodness-promoting sense, M1 is a concealed tautology, one of whose open forms would be

> M2: Everyone would always do what would make things go best if everyone always did what would make things go best.

Everyone could accept this claim, whatever their moral beliefs. Moore's *Principia* does not put forward a substantive moral view. [270]

Kant's view is the opposite of Moore's, since Kant claims that we should define *good* in terms of *ought*. In Kant's words,

> the concepts of *good* and *evil* must not be determined before the moral law. . . but only after it. . . and by means of it. [271]

Surprisingly, Kant also writes:

> All imperatives are expressed by an 'ought'. . . and say that. . . some act would be good. [272]

Kant may here seem to be doing just what he claims that we must not do, by defining *ought* in terms of *good*. Kant similarly calls certain acts 'practically necessary, that is, good.' [273] But these remarks do not use 'good' in any of its ordinary senses. In these ordinary senses, for example, some act may be good, though some other act would be even better. In these and other passages, Kant does not distinguish between some act's being good and this act's being practically necessary, or what we ought to do. And it is these latter words that better express what Kant has in mind. So I suggest that, when Kant calls some act 'good', he means that this act is what we ought to do. Kant would then be following his requirement that *good* be defined in terms of *ought,* since he would be using 'good' in an *ought-based* sense.

When Kant calls some end or outcome 'good' or 'best', he seems often to be using a similar ought-based sense. For example, when Kant claims

> K1: Good wills are supremely good, [274]

he seems in part to mean

> K2: Everyone ought to try to have a good will.

But Kant may also mean that we ought to try to have such wills *because* such wills are supremely good. This use of 'good' would not be ought-based. In this respect Kant's moral theory may be, as Herman claims, an ethics of value. But Kant would not be doing what he claims that we must not do, by deriving the content of the moral law from his beliefs about what is good. From the claim that good wills are supremely good we may be able to derive K2. But we cannot draw any other conclusions about what we ought to do.

The ancient Greeks, Kant claims, did make this mistake, since they tried to derive the moral law from their beliefs about the *Summum Bonum*, or the *Greatest Good*. [275] As we have seen, however, Kant himself describes an ideal world which he calls the Highest or Greatest Good, and he claims that everyone ought always to try to produce this world. Is Kant here making what he calls the 'fundamental error' of

the ancient Greeks? Is he deriving his beliefs about what we ought to do from his beliefs about the Greatest Good?

It may seem so. As we have seen, Kant claims

> K3: Everyone ought always to strive to promote the Greatest Good.

This may seem to be another version of Value-based Act Consequentialism. Kant may seem to be claiming that everyone ought always to try to produce the world that would be the best, or be the greatest good. And he makes other such remarks, as when he writes, of every human being, 'his duty at each instant is to do all the good in his power. ' [276]

This is not, I believe, the best way to interpret K3. Kant, I suggest, uses 'the Greatest Good' in an ought-based sense, to mean 'what everyone ought always to strive to promote'. If this is what Kant means, K3 could be restated as

> K4: Everyone ought always to strive to promote the world that everyone ought always to strive to promote.

This claim may seem to be a mere tautology, which everyone could accept. But that is not so. K4 implies that we should accept some version of Act Consequentialism, since K4 implies that there is some world that everyone ought always to strive to promote. Many people would reject that claim.

K4 does not, however, imply a *value-based* version of Act Consequentialism. And when Kant claims K3, he may also be using 'the Greatest Good' to refer to the possible world that he elsewhere claims to *be* the Greatest Good. K3 could then be more fully stated as

> K5: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

This is the clearest statement of this part of Kant's view, and this claim does not even use the words 'good' or 'best. So Kant's version of Act Consequentialism is *not* significantly value-based.

37 Promoting the Good

Nor is Kant's view clearly *Act* Consequentialist. Kant's Formula of the Greatest Good might be claimed to be the only principle we need, because we ought always to try directly to promote Kant's ideal world. But that is not Kant's view. Kant claims that we ought to follow certain other formulas, such as his Formulas of

Humanity and of Universal Law. So we can next ask how Kant's claims about the Greatest Good are related to his other formulas.

We can assume, Kant writes, that

> the laws of morality lead by their fulfilment to the highest end. [277]

He also writes:

> the strictest observance of the moral laws is to be thought of as the cause of the ushering in of the Greatest Good (as end). [278]

In these and other passages, Kant assumes

> K6: It is by following the moral law, as described by Kant's other formulas, that everyone could best promote the Greatest Good.

If everyone followed the moral law, and had good wills, everyone would thereby promote one element in Kant's ideal world, universal virtue, since such universal virtue would *consist* in everyone's following the moral law and having good wills. But this is not all that Kant means. When Kant claims that, if everyone followed the moral law, this would *lead to* or be the *cause of the ushering in of* the Greatest Good, Kant must be referring to the other element in this ideal world, universal deserved happiness. So Kant seems to assume

> K7: It is by following the moral law that everyone could best give everyone the happiness that their virtue would make them deserve.

Though everyone's following the moral law would make the world much closer to Kant's ideal, this would not be enough, Kant claims, fully to achieve this aim, since we would not be able to give everyone all of the happiness that they would deserve. Some good people, for example, would die young. But we can hope that our souls are immortal, and that after our deaths God will give everyone the rest of the happiness that they deserve.

We may doubt that Kant could have assumed K7. Kant seems to have believed that we ought to follow certain strict rules, such as rules forbidding lying, stealing, and breaking promises. It may seem unlikely that Kant could have believed that following such rules would most effectively promote deserved happiness.

That is *not*, however, unlikely. It was widely assumed, when Kant wrote, that

> (A) it is by following the rules of common sense morality, rather than by trying directly to promote everyone's happiness, that everyone could best promote everyone's happiness.

This assumption is also fairly plausible, as Sidgwick later argued. In trying to predict which acts would produce most happiness, people would make serious mistakes. And they would often deceive themselves in their own favour. It is easy to believe, for example, that our need for the property that we could steal is greater than the owner's need. If everyone was always trying to maximize happiness, that would also undermine or weaken various valuable social practices or institutions, such as the practice of trust-involving promises. And it would be in several ways bad if everyone had the motives of those who always try to maximize happiness. To be able always to act in this way, most of us would have to lose too many of the motives---such as strong love for particular people---on which much of our happiness depends.

We can next draw some distinctions that many earlier thinkers did not draw. I shall now use 'Consequentialist' to refer only to value-based views, and I shall use 'best' as short for 'best or expectably-best'. If we suppose that everyone will try to follow some set of rules, some possible rules would be

> *optimific* in the sense that, if these are the rules that everyone tries to follow, things would go best.

For the reasons just given, Sidgwick believed that the rules of common sense morality are fairly close to being optimific. According to one version of *Rule Consequentialism*, or

> *RC:* Everyone ought always to try to follow the optimific rules.

According to one version of *Act Consequentialism*, or

> *AC*: Everyone ought always to try do what would make things go best.

Of the people who accept either of these views, most now assume that these views conflict, so that we must choose between them. These people believe that

> (B) in some cases, breaking some optimific rule would be likely or certain to make things go best.

As an Act Consequentialist, Sidgwick claims that, in such cases, we ought break this optimific rule. According to most Rule Consequentialists, we ought instead to follow the optimific rules even when, by acting in this way, we would be likely or even certain to make things go worse.

There have been some people, however, who reject (B). These people believe that

> (C) it is by trying to follow the optimific rules that everyone would always be most likely to make things go best.

Moore came close to accepting (C). In trying to do the most good, Moore claims, we ought always to try to follow certain optimific common sense rules. [279] If (C) were true, these two forms of Consequentialism would not conflict but coincide, and we could accept them both. According to what we can call *Act-and-Rule Consequentialism*, or

> *ARC*: Everyone ought always to try to follow the optimific rules, since that is how everyone would be most likely to do what would make things go best.

In asking whether (C) is true, so that these forms of Consequentialism coincide, we must appeal to some view about how we ought to assess the effects of our acts. According to what we can call

> *the Marginalist View*: To decide how much good some act would do, we should ask what *difference* this act would make. The good that some act would do is the amount by which, if this act were done, things would go better than they would have gone if this act had not been done.

When we consider some kinds of case, this view can seem implausible. One example are cases in which some good result would be fully achieved if some number of people act in some way. If *more* than this number of people act in this way, the Marginalist View may imply that none of these people does any good. Suppose that, in

> *Rescue*, a hundred miners are trapped underground, with flood-waters rising. These miners lives will all be saved if four people join some rescue mission.

To make the causal relationships clear, we can suppose that, if four people stand on some platform, these people's weight will together be enough to raise each miner to the surface. On the Marginalist View, if five people join this mission, none of these people will save anyone's life. It is true of each of these five people that, if this person hadn't joined this mission, and stood on this platform, that would have made no difference, since the other four people would have saved all of the hundred miners' lives. According to Marginalists, none of these people does any good.

That conclusion may seem absurd. If none of these people saves anyone's life, how did a hundred lives get saved? Some writers claim that, to avoid such absurd conclusions, we should appeal to the effects of what people *together* do. According to one such view, which we can call

> *the Share of the Total View*: When some group of people together produce some good effect, the good that each person does is this person's share of the total good.

This view implies that, if five people join our rescue mission, thereby together saving a hundred lives, each person should be counted as saving twenty lives. It is irrelevant that, if any of these five people had not joined this mission, that would have made no difference. On this view, in deciding which of our possible acts would do the most good, we should ignore the effects of each act when considered on its own.

When Hume discusses our obligations not to steal and to respect other property rights, he asserts a similar but vaguer view. Justice and fidelity, Hume claims, 'are absolutely necessary to the well-being of mankind'. But the benefits of justice are 'not the consequence of every single act', since any particular just act, when 'considered in itself', may have effects that are 'extremely hurtful'. The benefits of justice arise only 'from *the whole scheme*' or 'the observance of the general rule'. [280] Hume therefore claims that, to produce these benefits, we must follow strict rules, making no exceptions even when breaking some rule would when 'considered in itself' have good effects. Such rules must be strict, or inflexible, because it is 'impossible to separate the good from the ill'.

On Hume's view, which we can call

> *the Whole Scheme View*: To decide how much good some act would do, we should not ask how much difference this act *by itself* would make. Each of our acts would do the most good if this act is one of a set of acts that would *together* do the most good.

If Act Consequentialists reject the Marginalist View and accept the Whole Scheme View, they might accept Hume's claim that we ought to follow certain strict rules, such as 'Never steal', since they might believe that this is how each of our acts would do the most good. These Act Consequentialists would then also be *Rule* Consequentialists. If the Whole Scheme View were true, so would be the claim that

> (C) it is by trying to follow the optimific rules that everyone would be most likely to make things go best.

On these assumptions, these two forms of Consequentialism would not conflict but coincide.

When Kant defends another strict rule, 'Never lie', he makes similar claims. In a notorious article, Kant condemns lying even to a would-be murderer who asks where his intended victim is. [281] It is often assumed that, in claiming that we must never lie, Kant states a view that could not possibly be Act Consequentialist. That is not so. Kant writes that, in telling a lie,

> I bring it about, as far as I can, that statements. . . in general are not believed,

> and so too that all rights which are based on contracts come to nothing and lose their force, and this is a wrong inflicted upon humanity in general.

And he writes

> Thus a lie. . . always harms another, even if not another individual, nevertheless humanity generally, inasmuch as it makes the source of right unusable. [282]

In these passages Kant condemns all lies by appealing to the harm that these acts bring about. As before, these claims might be made by those Act Consequentialists who reject the Marginalist View and accept the Whole Scheme View. Kant may have believed, like Hume, that each of our acts would do most good if we always followed certain strict rules.

Return next to Kant's claim that everyone's deserved happiness would be best promoted by 'the strictest observance of the moral laws'. Kant often makes such claims. For example, he writes:

> to promote the happiness of others is an end, the means to which I can furnish in no other way than through my own perfection. . .[283]

What Kant calls 'our own perfection' chiefly consists in our having good wills and acting rightly. So Kant here claims that acting rightly is the only way---or, as he may mean, the best way---to promote the happiness of others.

Kant also writes:

> If there is to be a Greatest Good, then happiness and the worthiness thereof must be combined. Now in what does this worthiness consist? In the practical agreement of our actions with the idea of universal happiness. If we conduct ourselves in such a way that, if everyone else so conducted themselves, the greatest happiness would arise, then we have so conducted ourselves as to be worthy of happiness. [284]

Kant here claims that, to be virtuous and act rightly, we must act in the ways which are such that, if everyone acted in these ways, that would produce universal happiness. This claim states one version of a Consequentialist theory: *Hedonistic Rule Utilitarianism*. If the Whole Scheme View and (C) were true, Kant's claim would also state a version of Hedonistic Act Utilitarianism, since these views would coincide.

These claims, however, have only historical importance, since we ought to reject both the Whole Scheme View and (C). Suppose again that, in

> *Rescue*, a hundred miners are trapped underground, with flood-waters rising. These miners will all be saved if four people join some rescue mission. I know that four other people have already joined this mission. I could either join this mission as well, or go elsewhere and save the life of some other single person.

On the Whole Scheme View, I ought to join this mission, since my act will then be one of a set of acts that will together do the most good, by saving a hundred people. That is clearly the wrong conclusion. I ought to save the single person, since one more person's life would then be saved. At least in most cases, we ought to accept the Marginalist View. When we ask which is the act that would do the most good, we ought to ask what *difference* this act would make. Since we ought to accept the Marginalist View, we could not be Act-and-Rule Consequentialists. Consequentialists have to choose between these forms of their view.

According to what I have called Kant's

> Formula of the Greatest Good: Everyone ought always to strive to promote a world of universal virtue and deserved happiness.

As I have argued, Kant seems to assume

> K6: It is by following the moral law, as described by Kant's other formulas, that everyone could best promote this ideal world.

On these assumptions, Kant's moral theory has the unity or harmony that Kant claims to be one of the goals of pure reason. Kant's Formula of the Greatest Good describes a single ultimate end or aim that everyone ought always to try to achieve, and Kant's other formulas describe the moral law whose being followed by everyone would best achieve this aim.

In deciding whether we ought to accept these claims, we would have two questions:

> Q1: Ought we always to strive to promote a world of universal virtue and deserved happiness?
>
> Q2: Is it by following Kant's other formulas that we can best promote this ideal world?

We cannot yet try to answer Q2, since we have not yet considered what is implied Kant's other main formula, his Formula of Universal Law.

Though we might try to answer Q1, I shall not do that. I shall, however, discuss one of Kant's assumptions about his ideal world. It is sometimes said that Kant's

claims about the Greatest Good add nothing to the rest of his moral theory. Kant claims elsewhere that we have two ends that are also duties, our own virtue and the happiness of others. [285] But in describing his ideal world, Kant adds that happiness is good only when it is *deserved*. On Kant's view, it would be bad if people had more happiness, or less suffering, than they deserve. [286] These claims about desert cannot be plausibly derived from, or claimed to be supported by, Kant's other formulas. [287] Nor does Kant try to support these claims in this way. He simply asserts these claims, or takes them to be obvious, as when he writes:

> Reason does not approve happiness. . . except insofar as it is united with worthiness to be happy, that is, with moral conduct. [288]

Kant's claims about desert are, I believe, false. And as I shall now argue, Kant came close to seeing that.

CHAPTER 11 FREE WILL AND DESERT

38 The Freedom that Morality Requires

According to *determinists*, all events are causally inevitable, so that, whenever we act in some way, it would have been causally impossible for us to have acted differently. Kant claims that, if determinism were true, morality would be undermined, since we wouldn't have the kind of freedom that morality requires. [289] And Kant believes that, in one way, determinism is true. But determinism is not, he claims, the whole truth. Kant distinguishes between the spatio-temporal *phenomenal* world, or reality as it appears to us to be, and the world of *noumena*, or things-in-themselves, which is reality as it really is. In this noumenal world, Kant argues, there is neither space nor time. It is conceivable that, as well as being phenomenal beings in the spatio-temporal world, we are also noumenal beings in this other world. Though our acts are partly events which occur in time in the spatio-temporal world, these acts might have undetermined origins in the timeless noumenal world. That, Kant claims, would give us the freedom that morality requires.

Kant also argues that we have such freedom. Kant's argument can be stated as follows:

> (A) Our acts cannot be wrong unless we ought to have acted differently.
>
> (B) 'Ought' implies 'can'. We ought to have acted differently only if we could have acted differently.

Therefore

> (C) Our acts cannot be wrong unless we could have acted differently.
>
> (D) If our acts were merely events in the spatio-temporal world, these acts would be causally determined, so it would never be true that we could have acted differently.

Therefore

> (E) If our acts were merely such events, none of our acts could be wrong, so morality would be an illusion.
>
> (F) Morality is not an illusion. We ought to act in certain ways, and some of

our acts are wrong.

Therefore

(G) Our acts are not merely events in the spatio-temporal world. [290]

In considering this argument, we might first object that, if (E) is true, we could not know that (F) was true unless we knew that (G) was true. If morality is an illusion unless our acts are not merely events in the spatio-temporal world, and we don't know whether our acts are merely such events, how could we know that morality is not an illusion? [291] But there might be ways in which, without first knowing that (G) was true, we could rationally believe that morality is not an illusion. This belief might, for example, be implied by some set of religious beliefs that we could rationally accept, and claim to know, as revealed truths.

We should also accept Kant's argument for (C). As Kant assumes, 'ought' implies 'can'. If we could not possibly act in some way---such as saving someone's life by running faster than a cheetah---it cannot be true that we *ought* to act in this way. For some act of ours to be wrong, because we ought to have acted differently, it must be true that we *could* have acted differently. There are, however, conflicting views about the sense in which this must be true. These are conflicting views about the kind of freedom that morality requires.

Suppose that, while I am standing in some field during a thunderstorm, a bolt of lightning narrowly misses me. If I say that I could have been killed, I might be using 'could' in a *categorical* sense. I might mean that, even with conditions just as they actually were, it would have been causally possible for this bolt of lightning to have hit me. If we assume determinism, that is not true, since it was causally inevitable that this lightning struck the ground just where it did. I may instead be using 'could' in a different, *hypothetical* or *iffy* sense. When I say that I could have been killed, I may mean only that, *if* conditions had been in some way slightly different---if, for example, I had been standing a few yards to the West---I would have been killed. Even if we assume determinism, that claim would be true.

We ought to have acted differently, Kant assumes, only if we could have done so in the categorical sense. It must be true that, even given our actual state of mind, it would have been causally possible for us to have chosen to act differently, and to have done so. If it was causally inevitable that we chose and acted as we did, it would not be relevantly true that we could have acted differently. On this view, as (E) claims, determinism is *incompatible* with the kind of freedom that morality requires.

As many writers argue, however, we ought to reject this *incompatibilist* view. Return to the case in which I say, 'You ought to have helped that blind man cross the

street', and you say, 'I couldn't have done that'. If I ask 'Why not?', it would not be enough for you to reply, 'Because I didn't want to'. Perhaps you could not have acted differently, in the relevant sense, if you were in the grip of some irresistible desire, or were insane. But most of us are not in these or other such ways unfree. In most cases, for it to be relevantly true we *could* have acted differently, it need only be true that

> (H) we *would* have acted differently *if* we had wanted to, and had chosen to do so.

We can call this the *hypothetical, motivational sense* of 'could'. This sense of 'could' is compatible with determinism. You could have helped the blind man cross the street in the sense that you would have done so *if* you had chosen to do so. It is irrelevant whether, given your actual desires and other mental states, it was causally inevitable that you did not choose to act in this way.

Someone might now object:

> If all of our decisions and acts are causally inevitable, we would have acted differently only if we had miraculously defied, or broken, the laws of nature. It is pointless to ask whether we ought to have acted in some way that would have required such a miracle.

Such questions, however, can be well worth asking. What we do often depends on our beliefs about what we ought to do. And if we come to believe that some act of ours was wrong, or irrational, because we ought to have acted differently, this belief may lead us to try to change ourselves, or our situation, so that we do not act wrongly, or irrationally, in this kind of way again. These changes in us or our situation may affect what we later do. It does not matter that, for us to have acted differently in the *past*, we would have had to perform some miracle. If we come to believe that we ought to have acted differently, this change in our beliefs may cause it to be true that in similar cases, without any miracle, we *do* in the *future* act differently. [292] That is enough to make it worth asking whether we ought to have acted differently.

Kant calls this *compatibilist* view 'a wretched subterfuge'. On this view, he claims, we would have only the 'freedom of a turnspit': a mechanical device that, when wound up, turns all by itself. But Kant's objections to compatibilism seem to depend in part on his failure to draw another distinction.

According to *fatalism*, it is inevitable that we shall later act in certain ways, *whatever* we decide to do. All of our different possible decisions would merely be different ways in which we would end up doing the same things. On this view, there is no point in our trying to make good decisions, since that would make no difference to

what we later do. Since it is clear that most of our acts *do* depend on our decisions, fatalism is believable only when it is restricted to certain particular acts. According to the Ancient Greek myth, for example, Oedipus was fated, whatever he decided, to kill his father and marry his mother. For this to be true, some Greek god would have had to be ready to intervene, to ensure that Oedipus's decisions would not have prevented his later acting in these two ways.

Determinism is a quite different view. On this view, what we shall later do will depend on what we decide to do. Though our decisions will be causally inevitable, we often don't know in advance, and could not possibly always know, what we shall later decide to do. And if we make better decisions, and act upon them, things will be likely to go better. These facts are enough to give us reasons to try to make good decisions. If we believed that there was no point in trying to make good decisions, we would be mistakenly slipping back into fatalism, by assuming that our decisions would make no difference to what happens.

Kant sometimes makes this mistake, as when he writes:

> unless we think of our will as free this imperative is impossible and absurd and what is left for us is only to await and observe what sort of decisions God will effect in us by means of natural causes, but not what we can and ought to do of ourselves, as authors. [293]

These remarks imply that, if determinism is true, there would be no point in our trying to decide what we ought to do. We would have to be *passive*, waiting to see what sort of decisions we shall be caused to make. That is not so. Even if determinism is true, we can be *active*, by trying to make and to act upon good decisions. If we are in some burning building, for example, we might try to decide how we can escape. If we merely wait and see what decision we shall later be caused to make, we shall be likely to make a worse decision, and be more likely to die.

Kant elsewhere suggests a different, compatibilist view. He writes:

> the practical concept of freedom has nothing to do with the speculative concept. . . For I can be quite indifferent as to the origin of my state in which I am now to act, I ask only what I now have to do, and then freedom is a necessary practical proposition. [294]

Kant seems here to see that, when we are deciding what to do, we can ignore the speculative or theoretical question of whether determinism is true. If we don't yet know what we shall decide, we are free in the sense that nothing will stop us from acting in certain ways, *if* we decide to do so. For practical purposes, it is only this compatibilist kind of freedom that we need. It is irrelevant whether, given our

actual state of mind, some other decision would have been causally impossible.

Though Kant sometimes suggests that, for practical purposes, the freedom that we need is compatible with determinism, his dominant view is clearly incompatibilist. Kant even claims that noumenal causeless freedom is the keystone of his entire philosophy. He would not have made that claim if he had accepted this compatibilist view.

According to the argument that we have been discussing, more briefly stated:

> (A) to (E): If our acts were merely events in time, these acts would be causally determined, and morality would be an illusion, since we would not have the kind of freedom that morality requires.
>
> (F) Morality is not an illusion.

Therefore

> (G) Our acts are not merely events in time.

We ought, I have claimed, to reject the reasoning that is summed up in (A) to (E). For some act of ours to be wrong, because we ought to have acted differently, it must be true that we *could* have acted differently. But the relevant sense of 'could' is the hypothetical, motivational sense. And this sense of 'could' is compatible with determinism. Even if our acts are causally determined, we could have the kind of freedom that morality requires.

39 Why We Cannot Deserve to Suffer

There is, however, another kind of compatibilism that Kant rightly rejects. Some of Kant's claims suggest this argument:

> (I) For it to be true that some act of ours was wrong, we must be morally responsible for this wrong act in some way that could make us deserve to suffer.
>
> (J) If our acts were merely events in time, we could never be responsible for these acts in this suffering-deserving way.

Therefore

> (E) If our acts were merely events in time, none of our acts could be wrong, so

morality would be an illusion.

(F) Morality is not an illusion.

Therefore

(G) Our acts are not merely events in time.

Premise (I) may seem plausible. There are some people whom no one believes to be morally responsible for their acts in some way that could make them deserve to suffer. That is true, for example, of young children, and some people who are insane. As well as believing that these people are not in this way responsible for their acts, we may believe that, for this reason, they cannot act wrongly.

There is a better way to explain why these people cannot act wrongly. Young children and these insane people cannot have or act upon beliefs about which acts are wrong. But ordinary sane adults can have and act on such beliefs. That is enough to justify our belief that most people are moral agents, whose acts can be right or wrong. So we should reject Kant's assumption that, for us to be moral agents, we must be responsible for our acts in some way that could make us deserve to suffer. We can coherently believe both that our acts can be right or wrong, and that no one could deserve to suffer.

According to premise (J), if our acts were merely events in time, we could not be responsible for our acts in this suffering-deserving way. This part of Kant's view is, I believe, a profound truth. We can be morally responsible in several other ways, or senses, but no one could ever be responsible, I believe, in any way that could make them deserve to suffer. Nor, I believe, could anyone deserve to be less happy.

Of Kant's reasons for assuming (J), one is his belief that

(K) if our acts were merely events in time, these acts would be causally determined,

and that

(L) if our acts were causally determined, we could never be responsible for these acts in some way that could make us deserve to suffer.

The kind of freedom that morality requires is, I have claimed, compatible with determinism. We could have acted differently, in the relevant sense, when nothing stopped us from acting differently except our desires or other motives. As Kant assumes, however, this kind of freedom is *not* enough to justify the belief that we can deserve to suffer for what we did. Kant here rightly rejects what we can call

compatibilism about desert.

Of the other people who reject this view, some would reject Kant's claim that, if our acts were merely events in time, these acts would all be causally determined. Most physicists now believe that determinism is not true, since events that involve sub-atomic particles are partly uncaused, or random. Such claims may not apply, however, to our decisions to act, and to other mental events. Most neuroscientists believe that mental events consist in, or causally depend upon, physical events in our brains which *are* fully causally determined, because these events occur on too large a scale to be affected by random events at the level of sub-atomic particles. But some people reject this view, believing that some of our decisions are not fully causally determined. Of those who have this belief, some appeal to randomness at the sub-atomic level. Others are *interactionist dualists*, who believe that mental events do not either consist in, or fully causally depend upon, physical events in our brains.

To justify the belief that we can deserve to suffer, it is not enough to defend the claim that our decisions to act in certain ways are not fully caused. If that is all we claim about any such decision, this would be, in Kant's phrase,

> tantamount to handing it over to blind chance. [295]

On this view, we would have the freedom not of a turnspit, whose movement is causally inevitable, but of a sub-atomic particle, whose movement is random. We could not deserve to suffer when and because some of the matter in our brains moved or changed in certain random ways. Nor would it help if, as some dualists claim, our decisions are non-physical events that are partly random.

Many people have claimed that, though *most* events must be either fully caused or partly random, that may not be true of our decisions and acts. These people try to describe some third possibility. Some of these people appeal to our rationality. When we claim that someone acted *for some reason*, these people suggest, we are not claiming that this person's act was fully caused, nor are we claiming that this act was partly random. Our ability to act for reasons may thus seem to provide a third alternative.

When someone acts for some reason, however, we can ask *why* this person acted for this reason. In some cases, the answer is given by some further reason. My reason for telling some lie, for example, may have been to conceal my identity, and my reason for concealing my identity may have been to avoid being accused of some crime. But we shall soon reach the beginning of any such chain of motivating reasons. My ultimate reason for telling my lie may have been to avoid being punished for my crime. When we reach someone's ultimate reason for acting in some way, we can ask why this person acted for this reason, rather than acting in some other way for some other reason. If I had a self-interested reason to try to

avoid being punished, and a moral reason not to tell this lie, why did one of these reasons weigh more heavily with me, so that I chose to act as I did? *This* event did not occur for some further motivating reason. So the suggested third alternative here disappears. This event was either fully caused or partly random. And there is always such an event at the origin of any chain of motivating reasons. Since our decisions to act as we do all involve such events, there is no coherent third alternative.

To avoid this argument, some people claim that acts can be caused by *agents* in a way that does not involve any *event*. Such believers in *agent-causation* partly accept Kant's view that, if our acts were merely events in time, we could not have any kind of freedom that could make it true that we can deserve to suffer because of what we did. But these writers believe that, as agents, we are fully part of the spatio-temporal world, so they cannot intelligibly claim that the causing of acts by agents are *not* events.

Kant makes some other relevant claims. To be responsible for our acts, Kant assumes, we must be responsible for our own character. In his words:

> The human being must make or have made *himself* into whatever he is. . . in a moral sense, good or evil. Either condition must be an effect of his free choice. . . [296]

And Kant writes of

> a man's character, which he himself creates,

and of

> a person who is his own originator.

Aristotle similarly writes:

> thus it was open at the beginning to the unjust and the self-indulgent man not to become like that, and so they are voluntarily as they are: but when they have become so, it is no longer possible for them not to be so. [297]

But Aristotle does not ask what could have happened 'at the beginning', when someone chose to make himself unjust or self-indulgent. Kant asks that question, and rightly claims that, if we are merely beings in the spatio-temporal world, we cannot have freely created our own character, thereby freely choosing to be either good or evil.

With the claims just quoted, and some other similar claims, Kant suggests another argument for his belief that our acts are not merely events in time. This argument is, in part:

> (M) What we decide to do depends on our character, and on certain other facts about what we are like, or *how we are*.

Therefore

> (N) To be responsible for our acts in some way that could make us deserve to suffer, we must be responsible for being in the relevant ways how we are.
>
> (O) If our acts were merely events in time, we could not be responsible for being how we are unless we acted *earlier* in ways that made us how we are.
>
> (P) To have been responsible for these earlier acts, we must have been responsible for how we *then* were, by having acted even earlier in ways that made us how we then were.
>
> (P) To have been responsible for these earlier acts, we must have been responsible for how we then were, by having acted even earlier in ways that made us how we then were.
>
> (P) To have been responsible for these earlier acts *etc. . . . and so on to infinity*.
>
> (Q) We could *not* have been responsible for such an infinite series of character-forming acts.

Therefore

> (J) If our acts are merely events in time, we cannot have chosen our own character, or be responsible for our acts in any way that could make us deserve to suffer. [298]

This part of Kant's argument is valid, and has, I believe, true premises. So we ought to accept (J).

Kant's argument continues:

> (R) We *are* responsible for our acts in a way that can make us deserve to suffer.

Therefore

> (S) Our acts are not merely events in time. We are responsible for our acts because, in the timeless noumenal world, we freely choose to give ourselves

our character, and to act as we do.

When other writers try to describe some third alternative to some act's being fully caused, or partly random, it is a decisive objection to such claims that they are incomprehensible. Compared with such claims, Kant's appeal to our noumenal timeless freedom is in one way easier to defend. We should not expect, Kant claims, to understand this noumenal timeless world. All we can expect to understand is the spatio-temporal phenomenal world. In Kant's words, though such noumenal freedom is incomprehensible, we can at least 'comprehend its incomprehensibility'.

This is not, I believe, a sufficient defence of Kant's view. We can vaguely understand how some part of reality might be timeless. And we can make some sense of the idea that all the features of the spatio-temporal world may, in some non-temporal way, depend on something that vaguely resembles a decision. Such claims may make some sense when applied to God. But some of Kant's claims about our timeless freedom are not even vaguely intelligible. On Kant's view, for example, though everything that happens in the spatio-temporal world is fully causally determined, everything that happens is also in part jointly brought about by a vast number of free and separate decisions, made timelessly, by all of the rational beings who ever live. It is inconceivable that so many free decisions, some of them good and others bad, could all select and bring about parts of the same single wholly determined sequence of events which is the entire history of the spatio-temporal world. And since these decisions would in part determine which rational beings ever exist, these beings must somehow bring it about that they themselves exist. It is not enough to say that we can at least understand why such claims are incomprehensible.

According to the argument that we are now discussing:

(J) If our acts were merely events in time, we could never deserve to suffer.

(R) We can deserve to suffer.

Therefore

(S) Our acts are not merely events in time.

We ought, I have claimed, to reject this argument's conclusion. Our acts *are* merely events in time. Since this argument is valid, and we ought to reject its conclusion, we must reject one of its premises.

Some people would reject (J). There are people who believe that, though our wrong acts are merely events in time, and are causally inevitable, we could deserve to suffer in Hell. On such views, to deserve to suffer, we don't have to have any

kind of contra-causal freedom, or to be in any way responsible for our own character, or for being as we are.

Of those who make such claims, some admit that they cannot understand how such claims could be true. God's justice, these people claim, is incomprehensible. Compared with Kant's claim that we should not expect to understand the timeless noumenal world, it is less plausible to claim that we should not expect to understand how we could deserve to suffer. We have no reason to expect such moral truths to be incomprehensible.

Rather than rejecting (J), we ought, I believe, to reject (R). Kant rightly claims that

(J) if our acts were merely events in time, we could not deserve to suffer.

We can add

(T) Our acts *are* merely events in time.

Therefore

(U) We cannot deserve to suffer.

Kant, I have said, came close to seeing the truth of (U). Kant believed that

(V) we could not deserve to suffer if our acts were either all causally inevitable, or were subject to blind chance, and we were not responsible for our own character.

These things *would* be true, Kant believed, if our acts were merely events in time. If Kant had lost his belief in our noumenal freedom, and come to believe that all our acts *are* merely events in time, he might have continued to believe (V), and drawn the conclusion that we cannot deserve to suffer. But I cannot claim to know that Kant would have drawn this conclusion. Kant might instead have ceased to believe (V), concluding that we *can* deserve to suffer even if our acts *are* causally inevitable or subject to blind chance, and we are not responsible for being as we are. I can merely hope that Kant would have continued to believe (V), and would have therefore seen that we cannot deserve to suffer.

Of those who believe that we can deserve to suffer, some would give this counter-argument:

(W) God makes some people suffer in Hell.

(X) God is just.

Therefore

(R) We can deserve to suffer.

But we don't, I believe, know that (W) is true. If we believe in a just God, we must accept either

(Y) God acts justly in making wrongdoers suffer in Hell, though it is unintelligible how such acts can be just,

or

(Z) God does not make anyone suffer in Hell.

Of these two claims, we would have more reason, I believe, to accept (Z). If God does not make anyone suffer in Hell, it may be surprising that so many people have believed that God *does* act in this way. But we can understand how these people might have come to have this false belief, and we cannot understand how a just God could make anyone suffer in Hell.

We can deserve many things, such as gratitude, praise, and the kind of blame that is merely moral dispraise. But no one could ever deserve to suffer. For similar reasons, I believe, no one could deserve to be less happy. When people treat us or others wrongly, we can justifiably be indignant. And we can have reasons to want these people to understand the wrongness of their acts, even though that would make them feel very badly about what they have done. But these reasons are like our reasons to want people to grieve when those whom they love have died. We cannot justifiably have ill will towards these wrong-doers, wishing things to go badly for them. Nor can we justifiably cease to have good will towards them, by ceasing to wish things to go well for them. We could at most be justified in ceasing to like these people, and trying, in morally acceptable ways, to have nothing to do with them. [299]

If Kant had seen that no one could deserve to suffer, or to be less happy, his ideal would still have been a world in which we were all virtuous and happy. But he would have changed his view about less than ideal worlds, since he would have ceased to believe that it would be bad if some people suffered less, or were happier, than they deserved.

Though Kant makes various other claims about his ideal world, these are not the most valuable parts of Kant's moral theory. Many other writers claim that the two greatest goods are virtue and happiness. And Kant says little to defend his assumption that, if we follow his other formulas, we shall be doing what will best promote his ideal world. What is most valuable are some of the parts of Kant's theory that are not in these ways Consequentialist. We have considered Kant's

Formula of Humanity, and his related claims that to treat people as ends, we must treat them only in ways to which they could rationally consent, and must never treat them merely as a means. We can now turn to Kant's other main statement of his supreme principle, the Formula of Universal Law. Though many people have discussed this formula, none, I believe, has fully seen what Herman calls the 'untapped theoretical power and fertility of this alternative to Consequentialist reasoning'. [300]

PART THREE THEORIES

CHAPTER 12 UNIVERSAL LAWS

40 The Impossibility Formula

Whether our acts are right or wrong, Kant claims, depends on our *maxims*, by which Kant usually means our policies and their underlying aims. Some of Kant's examples are: ''Increase my wealth by every safe means', [301] 'Let no insult pass unavenged', [302] 'Make lying promises when that would benefit me', 'Give no help to those who are in need', [303] and 'the maxim of self-love, or one's own happiness'. [304]

According to one of Kant's versions of his Formula of Universal Law, which we can call

> *the Impossibility Formula*: It is wrong to act on any maxim that could not be a universal law. [305]

This formula needs to be explained. In one passage, Kant refers to a maxim's being 'a universal permissive law'. [306] This may suggest that Kant means

> (A) It is wrong to act on any maxim if we could not all be permitted to act upon it.

But Kant never appeals to (A). And as I explain in a note, (A) would not be a useful claim. [307]

Some writers suggest that Kant means

> (B) It is wrong to act on any maxim that we could not all *accept*, in the sense of deciding to act upon it.

On this suggestion, Kant's formula would be unreliable. If (B) condemned acting on any maxim that it would be inconceivable, or logically impossible, for all of us to accept, this formula would fail to condemn most wrong acts. We can easily conceive or imagine worlds in which everyone accepts bad maxims, such as the maxim 'Deceive and coerce other people whenever that would benefit me'. Such worlds might be *causally* impossible, because there are some good people who

would be psychologically unable to accept these bad maxims. But there are also some bad people who would be psychologically unable to accept some good maxims. So if (B) appealed to such causal impossibility, this formula would mistakenly condemn acting on these good maxims. We might appeal to some other kind of impossibility. But as these remarks suggest, (B) is implausible. We have no reason to believe that whether maxims are good or bad, and whether it is wrong to act upon them, depends on whether everyone could accept them.

Some writers suggest that Kant means

> (C) It is wrong to act on some maxim if it would be impossible for everyone to act upon it.

The word 'everyone' here refers to all of the people who could act on this maxim. The maxim 'Care for my children', for example, applies only to parents.

This formula would also be unreliable, since (C) condemns many morally required or permissible acts. There are many good maxims on which some people could not act, because they do not have the opportunity or ability to act in these ways. Some parents cannot care for their children, because they are in prison, or are mentally ill. But caring for our children is not wrong. To avoid this objection, (C) might condemn acting on any maxim that could not be acted on by everyone who has both the opportunity and the ability to act upon it. But no maxim would fail this test. And (C) is also implausible, since we have no reason to believe that whether maxims are good or bad, and whether it would be *wrong* to act upon them, depends on whether everyone *could* act upon them.

Some writers suggest that Kant means

> (D) It is wrong to act on some maxim if it would be impossible for everyone who could act upon it to act *successfully*, in the sense that they would achieve their aims. [308]

This formula would be no better. There are many maxims on which it would be permissible or good to act, though we could not all successfully act upon them. Some examples are: 'Become a doctor or a lawyer', 'Adopt an orphan', 'Give more to charity than the average person gives', and 'Be the last person to use any fire-escape, or to leave any sinking ship'. If we all tried to achieve these aims, some of us would fail. (D) is also implausible. We have no reason to believe that, if we could not all successfully act on some maxim, it would be wrong for anyone to act upon it. It is not wrong to make attempts some of which we know will fail.

We have been trying to understand Kant's claim that it is wrong to act on maxims that could not be universal laws. (A) to (D) are the most straightforward ways to interpret this claim. But as well as being either unhelpful or both unreliable and

implausible, these are not claims to which, when Kant applies his formula, he himself appeals. Though Kant's *stated* Impossibility Formula is

> (E) It is wrong to act on any maxim that could not be a universal law,

Kant's *actual* formula is

> (F) It is wrong to act on any maxim of which it is true that, if everyone accepted and acted on this maxim, or everyone believed that it was permissible to act upon it, that would make it impossible for anyone successfully to act upon it. [309]

Could this formula help us to decide which acts are wrong?

Consider first the maxim 'Kill or injure other people when that would benefit me'. As Herman points out, if we all accepted and acted on this maxim, that would not make it impossible for any such act to succeed. [310] So (F) does not condemn such acts. Nor does (F) condemn self-interested coercion. If we all tried to coerce other people whenever that would benefit ourselves, some of these acts would succeed.

Turn next to lying. Herman writes that (F)

> seems adequate for maxims of deception. . . Universal deception would be held by Kant to make speech and thus deception impossible. [311]

Korsgaard similarly writes:

> lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive. . . [312]

But no one acts on the maxim 'Always lie'. Many liars act on the maxim 'Lie when that would benefit me'. Kant's formula condemns such acts only if, in a world of self-interested liars, it would be impossible for any such lie to succeed. That would not be impossible. Even in such a world, it would often be in our interests to tell others the truth. And when it would be in our interests to deceive someone, there would often be no point in lying, since this person would not believe our lie. So, even if we were all self-interested liars, many of our statements would be true. Most of us would know this fact. And since we could not always tell which statements by others were lies, some lies would be believed, and would achieve the liar's aim.

To explain why theft is wrong, Kant writes:

> Were it to be a general rule to take away his belongings from everyone, *mine*

> and *thine* would be altogether at an end. For anything I might take from another, a third party would take from me. [313]

As before, however, no one acts on the maxim 'Always steal'. Many thieves act on the maxim 'Steal when that would benefit me'. If this maxim were universally accepted and acted upon, that would not produce a world in such acts would never succeed. There would still be property, which would not always be successfully protected. Thieves would sometimes achieve their aims.

When Kant discusses the maxim 'Let no insult pass unavenged', he claims that, if this maxim were universal, it would be 'inconsistent with itself', and would not 'harmonize with itself'. [314] But if everyone acted on this maxim, that would not make it true that no one could succeed. It might even be true that every insult was avenged, so that *everyone* would succeed.

Kant's actual formula, we have found, fails to condemn many of the acts that are most clearly wrong. This formula does not condemn self-interested killing, injuring, coercing, lying, and stealing.

These failures may suggest that Kant's formula condemns nothing. But we have still to consider Kant's best example: that of someone who makes a lying promise so that he can borrow money that he does not intend to repay. This man acts on the maxim 'Make lying promises when that would benefit me'. [315] Kant claims that, if everyone accepted this maxim, and believed that lying promises are permissible, that would make it impossible for any such promise to succeed. In his words:

> the universality of a law that everyone . . . could promise whatever he pleases with the intention of not keeping it would make the promise . . . impossible, since no one would believe what was promised him but would laugh at all such expressions as vain pretenses. [316]

In assessing this claim, as Rawls suggests, we should ask what would be true after some period that was long enough for everyone's acceptance of the lying-promiser's maxim to have its full effects.[317] Kant seems right to claim that, in such a world, no one would be able to benefit themselves by making any lying promise. Not only would such promises not be believed; the social practice of morally motivated, trust-involving promises would have ceased to exist. Kant's formula therefore condemns such lying promises.[318] And most of these acts, we can assume, are wrong.

Now that we have found one kind of wrong act that Kant's formula condemns, we can ask whether this formula is plausible. Kant's formula is, in part:

> (G) It is wrong to act on any maxim of which it is true that, if everyone believed such acts to be permissible, that would make it impossible for any such act to succeed.

This claim condemns those acts whose success depends on other people's refraining from such acts, because they believe such acts to be wrong. And (G) may seem to condemn these acts for a good reason. Lying promisers act wrongly, we might suggest, because if everyone believed such acts to be permissible, that would undermine a valuable social practice.

Kant's claims are not restricted, however, to *valuable* social practices. The soldiers in Hitler's armies, for example, were required to swear oaths of unconditional obedience. Kant condemns lying promises with the claim that, if everyone believed that lying promises were permissible, the practice of making promises would be a 'vain pretense', or sham. Some of these German soldiers rightly believed that it was morally permissible for them, despite having sworn this oath, to disobey all immoral commands. We could similarly claim that if all these soldiers had believed such disobedience to be permissible, the practice of swearing oaths of unconditional obedience would have been a vain pretense or sham. Kant's remarks seem to imply that such disobedience would be wrong. But as Kant himself claims, everyone ought to disobey immoral commands.

For another test of (G), we can suppose that, during the Second World War, some non-Jewish German civilian knows that German Jews are being rounded up and killed. This person successfully acts on the maxim 'Tell lies to the police when that would save some Jewish person's life'. Suppose next that, if everyone had been known to believe that such lies were permissible, that would have made it impossible for anyone to save people's lives in this way. German policemen would have been required to search every building, ignoring anyone's claims that this building contained no Jews. On these assumptions, (G) would have condemned this person's life-saving acts.

Kant might have accepted this conclusion, given his claim that it would be wrong to lie even to a would-be murderer who asks where his intended victim is. [319] But such life-saving lies would be clearly justified. And when applied to this example, (G) is implausible. It would be no objection to this way of saving people's lives that, if everyone believed such acts to be permissible, that would make them impossible.

This imagined case is like Kant's case of a lying promiser. Kant's promiser achieves his aim because there are many people who can be trusted not to make lying promises, given their belief that such promises are wrong. Kant claims that, if everyone was known to believe that such promises are not wrong, that would have made it impossible for anyone to act successfully on this lying promiser's maxim. If that is true, Kant's formula implies that this person's lying promises are wrong.

Similar claims apply to my example. My German civilian achieves her aim because there are many people who can be trusted not to lie to the police, given their belief that such lies are wrong. I have supposed that, if everyone was known to believe that such lies are not wrong, that would have made it impossible for anyone to act successfully on this person's life-saving maxim. If that is true, Kant's formula mistakenly implies that this person's life-saving lies were wrong. The important difference between these acts is in what they are intended to achieve; and this difference is ignored by (G).

As this and other such cases show, (G) is unacceptable. [320] As well as failing to condemn nearly all of the acts that are most clearly wrong, (G) condemns some acts that are clearly right. And though (G) correctly condemns lying promises, it condemns these acts for a bad reason.

Kant's formula is also, in part,

> (H) It is wrong to act on any maxim whose being universally accepted and acted upon would make it impossible for anyone successfully to act upon it.

This formula, some writers claim, condemns acting on several good maxims, such as 'Refuse to accept bribes' and 'Give generously to the poor'. If these maxims were universally acted upon, that would soon make it impossible for anyone to act successfully on these maxims, since no one would offer any bribes, and there would cease to be any poor people. So Kant's formula mistakenly implies that it would be wrong both to refuse bribes and to give generously to the poor.

Korsgaard partly answers this objection. When people act on the maxim of giving to the poor, their aim, Korsgaard suggests, is to abolish poverty. If all rich people acted on these people's maxim, that might abolish poverty, thereby making it impossible for anyone later to act on this maxim. But (H) would not mistakenly condemn these people's acts, because by giving to the poor these people would *achieve* their aim. [321]

These claims do not apply, however, to some rich people. When these people act on the maxim 'Give generously to the poor', their aim is not to abolish poverty but to be admired for their generosity. If all rich people acted on this maxim, their acts might abolish poverty, thereby making it impossible for any of these people to act on their maxim in a way that would achieve their aim. (H) would then mistakenly condemn these people's acts. When these people give large sums to the poor, their acts have no moral worth, but they are not acting wrongly. [322]

Consider next those men who accepted codes of honour, like the code that led the Russian poet Pushkin to fight his fatal duel in the snow. Suppose that Pushkin had

accepted the maxim 'Fight duels to show my courage, but always fire into the sky'. If all these men had accepted and acted on this maxim, the practice of duelling would have become farcical, and would not have survived. That would have made it impossible for Pushkin to act on his maxim in a way that would achieve his aim, so (H) would have condemned Pushkin's acting on this maxim. (H) may seem to give the right answer here, since duelling is wrong. But (H) would *not* have condemned acting on the maxim 'Fight duels to show my courage, and always shoot to kill.' And acting on this second maxim would have been much worse. As this comparison suggests, (H) would have condemned Pushkin's act for a bad reason. It would have been no objection to Pushkin's maxim that, if this maxim were universally accepted, the practice of duelling would not survive. As before, Kant's formula mistakenly ignores the question of whether some social practice is good, and ought to be supported.

For another example, consider the maxim, 'Have no children, so as to have more time and energy to work for the future of humanity.' If everyone acted on this maxim, that would make it impossible for anyone successfully to act upon it, since humanity would have no future. So (H) mistakenly condemns such acts.

O'Neill proposes a weaker version of (H). Kant's formula, O'Neill suggests, could become

> (I) It is wrong to act on any maxim whose being successfully acted on by some people would prevent some other people from successfully acting on it. [323]

This formula condemns deception and coercion, O'Neill claims, since those who deceive or coerce others thereby 'guarantee that their victims cannot act on the maxims they act on.' [324] But this claim is false. Of those who have been deceived or coerced, most can deceive or coerce other people. O'Neill also claims that, while we are deceiving or coercing people, we 'undercut their agency', thereby preventing them 'for at least some time' from acting successfully in the same way as us. [325] But this claim is also false. Two people can simultaneously deceive each other. And there can be mutual simultaneous coercion. Two wrestlers might simultaneously use force to keep each other on the ground. And I might coerce you by making one credible threat, while you are coercing me by making another. That is how hostile nations with nuclear weapons might deter each other from using these weapons.

O'Neill could reply that, to show that (I) condemns deception and coercion, it is enough to claim that *some* deceivers and coercers prevent *some* of their victims from deceiving or coercing others. This weaker claim is true. O'Neill similarly claims that, if we acted on maxims of 'severe injury', some of us would disable some of our victims, thereby preventing these people from severely injuring others. So (I) condemns some wrong acts. But (I) condemns these acts for a bad reason. What is wrong with deceiving, coercing, and severely injuring others isn't that, by acting in

these ways, we prevent some other people from successfully doing the same.

(I), moreover, mistakenly condemns many good or morally permissible acts. There are many good or permissible maxims of which it is true that, if some people successfully acted on them, that would prevent some other people from doing the same. As O'Neill points out, (I) implies that we act wrongly if we play competitive games with the aim of winning. [326] Though some English schoolboys were told to accept this view, it seems too severe. And there would be nothing wrong with acting on the maxim 'Become a doctor', even if, by applying and being admitted to some medical school, we prevented someone else from being admitted to any medical school. Or consider the maxims 'Discover what killed all the dinosaurs', 'When traveling with others, always carry the heaviest load', and 'Find someone with whom I can happily live my life'. It is not wrong to try to make some discovery, or to carry the heaviest load, even though, if we succeed, we shall make it impossible for some other people to do these things. Nor is it wrong to live happily with the only person with whom someone else could have happily lived.

Korsgaard proposes another version of Kant's Impossibility Formula. What this formula forbids, she suggests, are acts whose success 'depends upon their being exceptional.' This test, she adds, 'reveals unfairness'. [327] But that is not, I believe, true. And this version of Kant's formula also mistakenly condemns many permissible acts. Some poor people get their food by searching through the rubbish that others throw away. That method must be exceptional, but is not wrong, or unfair. It was not wrong for romantic poets to give themselves the experience of being the only human being in some wilderness. Nor is it wrong, or unfair, to use tennis courts when they are least crowded, [328] pay the debts on our credit cards before interest is charged, [329] buy only second-hand books, or give surprise parties. [330]

Though there are other ways in which we might interpret or revise Kant's Impossibility Formula, these possibilities are not worth considering. Of the interpretations and revisions that we have considered, none contains a good idea. There is no useful sense in which we could claim it to be wrong to act on maxims that could not even *be* universal laws.

41 The Law of Nature and Moral Belief Formulas

Kant proposes another, better formula. According to Kant's main statement of his

> *Formula of Universal Law*: It is wrong to act on maxims that we could not *will* to be universal laws. [331]

Kant remarks that, when maxims fail this test, we have unstrict duties not to act

upon them. Such duties are *unstrict* in the sense that we are sometimes morally permitted to act on such maxims. We should ignore this remark, as Kant often does. Kant claims that our *strict* duties can be derived from his Impossibility Formula. As we have seen, that is not true. So we should ask whether Kant's Formula of Universal Law can do better, by correctly implying that some kinds of act are always wrong. As Herman points out, it would not be enough if Kant's formula implied that, though it would be wrong to have a policy of killing others for our own convenience, such acts are *sometimes* permitted. [332]

When we apply Kant's formula, we suppose or imagine that we have the power to *will*, or choose, that certain things be true. We are doing a *thought-experiment*, which involves comparing different possible states of the world, or what we can call different *possible worlds*. Like the thought-experiments of some scientists, our thoughts about these possible worlds may lead us to conclusions which also apply to the actual world.

When Kant asks whether we could will it to be true that some maxim is a universal law, he sometimes asks whether we could *consistently* will this to be true. He asks, for example, whether our will would *conflict* with itself, or would *contradict* itself. In other passages, Kant seem to ask what we could *rationally* will, or choose. Kant's formula is more likely to succeed if we use 'could will' in this second, wider sense. On some views, this will make no difference, since our choices fail to be rational only when they are inconsistent, or conflict with each other. But as I have argued, for our choices to be rational, we must also respond well to reasons or apparent reasons. We could not rationally choose or will it to be true that some maxim is a universal law if we are aware of facts that give us clearly decisive reasons not to make this choice.

In willing that some maxim be a universal law, what would we be willing? Kant sometimes claims that, when we apply his formula, we should ask whether we could will that our maxim be a 'universal law of nature', in the sense that everyone would accept and act on this maxim. [333] On this version of Kant's formula, which we can call

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

As before, the word 'everyone' refers to all of the people to whom some maxim applies. The maxim 'Give up smoking', for example, applies only to smokers.

In some other passages, Kant appeals to what we can call

> *the Permissibility Formula*: It is wrong for us to act on some maxim unless we

> could rationally will it to be true that everyone is morally permitted to act on this maxim. [334]

When Kant applies this formula, he assumes that, if everyone were permitted to act on some maxim, at least some people would be more likely to act upon it. This effect would be produced, not by these people's *being* permitted to act on this maxim, but by their *believing* that such acts are permitted. So Kant must also be appealing to what we can call

> *the Moral Belief Formula*: It is wrong for us to act on some maxim unless we could rationally will it to be true that everyone believes that such acts are morally permitted. [335]

Given their similarity, it is not worth using both these formulas. And unlike the Permissibility Formula, as I explain in a note, the Moral Belief Formula can be plausibly used on its own. [336] So we can ignore the Permissibility Formula.

Kant remarks that he is proposing, not a 'new principle', but only a more precise statement of the principle that 'common human reason. . . has always before its eyes'. [337] This remark understates Kant's originality. But Kant's Law of Nature and Moral Belief Formulas develop the ideas that are expressed in two familiar questions: 'What if everyone did that?' and 'What if everyone thought like you?'

When we apply these formulas, we must appeal to some beliefs about rationality and reasons. We might appeal to what Kant himself believed. But that would be difficult, since Kant did not clearly state these beliefs. And we are asking whether Kant's formulas can help us to decide which acts are wrong, and help to explain why these acts are wrong. In asking these questions, we should try to appeal to true beliefs about rationality and reasons. We should therefore appeal to our own beliefs, since we are then appealing to what we believe to be the truest or best view. Though we know that we might be mistaken, we cannot appeal to what *is* true rather than what we believe to be true.

There are, however, some beliefs to which we should not appeal. First, we should not appeal to our beliefs about which acts are wrong. I am calling these our *deontic beliefs*. Nor should we appeal to the *deontic reasons* that an act's wrongness might provide. When we apply Kant's Law of Nature Formula, it would be pointless to claim both that

> (1) it is wrong to act on a certain maxim because we could not rationally will it to be true that everyone acts on this maxim,

and that

> (2) we could not rationally will it to be true that everyone acts on this maxim because such acts are wrong.

If we combined these claims, that would be like pulling on our boot laces in an attempt to hold ourselves in mid air. To vary the metaphor, we would be going round in a circle, getting nowhere. Kant does not make this mistake. When Kant claims that we could not rationally will it to be true that everyone acts on some bad maxim, he never appeals to his beliefs that such acts are wrong and that we could not rationally will it to be true that everyone acts wrongly. Kant knew that, if he appealed to such beliefs, his Law of Nature Formula would achieve nothing, since this formula could not then help us to reach true beliefs about which acts are wrong, nor could it support these beliefs.

Similar remarks apply to Kant's Moral Belief Formula. It would be pointless to claim both that

> (3) it is wrong to act on a certain maxim because we could not rationally will it to be true that everyone believes such acts to be permitted,

and that

> (4) we could not rationally will it to be true that everyone believes such acts to be permitted because such acts are wrong.

When we ask whether we could rationally will that everyone *believes* some kind of act to be wrong, we should not appeal to our beliefs about whether such acts *are* wrong. As before, when Kant applies this formula, he follows this *Deontic Beliefs Restriction*, making no appeal to such beliefs.

There is another belief to which we should not appeal. Many wrong acts benefit the agent in ways that impose much greater burdens on others. On some views, such acts are irrational, since we are rationally required to give great weight to everyone else's well-being. If we accept such a view, we should ignore it when we apply Kant's formulas. The main idea behind Kant's Law of Nature Formula is that, even if wrong-doers could rationally act on certain bad maxims, they could not rationally will it to be true that *everyone* acts on their maxims. When we apply this idea, it would be irrelevant to claim that, because these people are rationally required to give great weight to other people's well-being, they could not even rationally will it to be true that *they themselves* act on their maxims.

As before, Kant does not make such claims. When Kant discusses a rich and self-reliant man who has the maxim of not helping others who are in need, Kant does not appeal to the belief that this man is rationally required to give such help. As Rawls and Herman suggest, when we apply Kant's formulas to people who act on such maxims, we should suppose that these people's maxims and acts are both rational. [338]

We can add that, if we combine Kant's formulas with less controversial and more widely accepted assumptions about rationality and reasons, these formulas would, if they succeed, achieve more.

42 The Agent's Maxim

Whether some act is wrong, Kant's formulas assume, depends on the agent's maxim. Of the maxims that Kant discusses, most involve some *policy*, which could be acted on in several cases. Two maxims may be different, though they involve the same policy, because they involve different underlying motives or aims. Two merchants, for example, may both act on the policy 'Never cheat my customers'. But these merchants act on different maxims if one of them never cheats his customers because he believes this to be his duty, while the other's motive is to preserve his reputation and his profits.

Kant's appeal to the agent's maxim raises various problems. Let us call some maxim

> *universal* when everyone both acts on this maxim whenever they can, and believes such acts to be permitted.

Suppose that I wrongly steal some wallet from some woman dressed in white who is eating strawberries while reading the last page of Spinoza's *Ethics*. My maxim is to act in precisely this way, whenever I can. I could rationally will it to be true that this maxim is universal, because it would be most unlikely that anyone else would ever be able to act in precisely this way, so this maxim's being universal would be most unlikely to make any difference. Since I could rationally will this maxim to be universal, Kant's formulas mistakenly permit my act. [339] Similar claims apply to other highly specific maxims. When wrong-doers act on such maxims, they could rationally will that their maxims be universal, because they would know that other such acts would be rare, and would therefore make little difference. Kant's formulas would mistakenly permit these wrong acts. We can call this the *Rarity Objection*.

This objection can be partly answered. Just as it is a factual question what someone believes, or wants, or intends, it is a factual question on which maxim someone is acting. And real people seldom act on such highly specific maxims. When we describe someone's maxim, as O'Neill and others claim, we should not include any details whose absence would have made no difference to this person's decision to do whatever he is doing. [340] In a realistic version of my example, I would have stolen from my victim even if if she had been dressed in red, or had been eating blueberries, or had been reading the first page of *Right Ho Jeeves!* My real maxim

would be something like 'Steal when that would benefit me'. This may *not* be a maxim that I could rationally will to be universal. Kant's formulas would then correctly imply that my act is wrong.

These remarks do not fully answer the Rarity Objection. Even if actual wrong-doers never acted on such highly specific maxims, we can imagine such people. Kant's formulas ought to be able to condemn these imagined people's acts. [341] And as we shall see, this objection applies to some actual cases.

Kant's appeal to the agent's maxim raises other, more serious problems. Consider some man who often acts on

the Egoistic Maxim: Do whatever would be best for me.

This man could not rationally will it to be true either that everyone always acts on this maxim, or that everyone believes that all such acts are morally permitted. Egoists could not rationally choose to live in a world of Egoists, since that would be much worse for them than a world in which everyone accepts various moral maxims. Since this Egoist could not rationally will that his maxim be universal, Kant's formulas imply that, whenever he acts on his maxim, his act is wrong. This man acts wrongly not only when he steals and lies, but also when, for self-interested reasons, he pays his debts, keeps his promises, and saves a drowning child, because he hopes to get some reward. These are unacceptable conclusions. When this Egoist acts in these ways, his acts have no moral worth. But these acts are not wrong.

It might be claimed that, when this man acts in any of these ways, *what* he is doing is not wrong, but *his doing* of it is. Kant suggests a similar distinction when he claims that, to fulfil some *duties of virtue*, we must not only act rightly, but also act with the right motive. On Kant's view, Rawls claims, even if we do not kill ourselves, we may have failed to fulfil our duty not to kill ourselves. To fulfil this duty, we must refrain from killing ourselves for the right reason. [342] Kant similarly claims that to fulfil a duty of gratitude, we must feel grateful. [343]

These distinctions cannot answer this objection to Kant's formulas. My Egoist may never fulfil his duties of virtue, since he may never have the right motive. As Kant claims, however, we also have many *duties of justice*, which we can fulfil by doing what is morally required, whatever our motive. One example is our duty to pay our debts. Kant's prudent merchant would do his duty if he acted on the maxim 'Pay my debts', even if this merchant's only motive was to preserve his reputation and his profits. Kant's formula gives the right answer here, since this merchant would be acting on a maxim that he could rationally will to be universal. But when

my Egoist pays his debts, he is acting on his Egoistic maxim, which he could *not* rationally will to be universal. So Kant's formulas mistakenly imply that, when this man pays his debts, he is *not* doing his duty, but is acting wrongly.

Return now to the drowning child. Suppose that, because this child has fallen into some fastly flowing river near some deep waterfall, any attempt to save this child would be too risky to be anyone's duty. If some good person saved this child, despite these risks, this person would be heroically acting beyond the call of duty. My Egoist decides that it would be worth taking these risks, since he could then hope to get a greater reward. Acting on his maxim, he dives into the river. On the suggestion we are now considering, if this man saves this child's life at this great risk to his own life, what he is doing is not wrong, but his doing of it is. That is clearly false. This man is not failing to fulfil any duty, or acting wrongly in any sense.

Turn next to prudent acts which affect no one else. When this Egoist takes some medicine, or puts on warmer clothing, he may be acting on his maxim 'Do whatever would be best for me'. Since this man could not will that this maxim be universal, Kant's formulas again mistakenly imply that he is acting wrongly. Nor could we claim that, though *what* he is doing is not wrong, his *doing* of it is. There is no sense in which, when this man puts on warmer clothing, his acting in this way is wrong.

Some writers suggest that we should not apply Kant's formulas to maxims that are as general as 'Do whatever would be best for me'. But Kant often discusses this Egoistic maxim, which he calls 'the maxim of self-love, or one's own happiness'. [344] And if we claimed that such maxims are too general, we would be ignoring many people's actual maxims. Kant discusses the maxim 'Make a lying promise when that would benefit me'. There are other, similar maxims, such maxims of stealing, cheating, or breaking the law whenever that would be best for ourselves. Since these maxims all involve the same more general policy, they are unnecessary clutter, and could all be replaced by the single maxim 'Do whatever would be best for me'. When many actual people act on this maxim, or policy, it may be simply false to claim that these people also accept, and are acting upon, on any other, less general policy.

For examples of a different kind, we can turn to conscientious people who have false moral beliefs. One example could be Kant himself during the period in which, as some of his remarks suggest and we can here suppose, Kant accepted the maxim 'Never lie'. This maxim is condemned by Kant's formulas. Kant could not have rationally willed it to be true that no one ever tells a lie, not even to a would-be murderer who asks where his intended victim is. Nor could he have rationally willed it to be true that everyone believes these life-saving lies to be wrong. [345] So Kant's formulas imply that, whenever Kant acted on this maxim by telling anyone

the truth, his act was wrong. He acted wrongly even when he told someone the correct time of day. That is clearly false. Similar claims would apply to people who accept the maxims 'Never steal' and 'Never break the law'. These people could not rationally will it to be true that no one ever steals or breaks the law, not even when such acts were the only way to save some innocent person's life. So Kant's formulas imply that, whenever these people act on these maxims, by returning someone's property or keeping some law, they act wrongly. These implications are also clearly false.

Our problem can be redescribed as follows. Some maxims are *wholly bad*, or *wholly good*, in the sense that it is always wrong, or always right, to act upon them. Two examples are the maxims 'Torture others for my own amusement' and 'Prevent pointless suffering'. When applied to such maxims, Kant's formulas succeed. But many maxims are

> *morally mixed* in the sense that, if we always acted on these maxims, some of our acts would be wrong, but other acts would be permissible or even morally required.

Two examples are the Egoistic maxim and Kant's maxim 'Never lie'. In proposing his formulas, Kant overlooks such mixed maxims. Kant's formulas assume that acting on some maxim is either always wrong, or never wrong. When applied to mixed maxims, Kant's formulas fail, since these formulas condemn some acts that are clearly permissible or morally required. When my Egoist prudently pays his debts, and Kant tells most people the truth, they are not acting wrongly, as Kant's formulas mistakenly imply. We can call this the *Mixed Maxims Objection*.

After considering this and other objections to Kant's Formula of Universal Law, in either its law of nature or moral belief versions, some writers conclude that we cannot use Kant's formula to help us to decide which acts are wrong. Wood claims that, when used as such a criterion, Kant's formula is 'radically defective' and 'pretty worthless'. [346] Herman claims that, despite a 'sad history of attempts. . . no one has been able to make it work'. [347] O'Neill suggests that, in some cases, Kant's formula may 'give either unacceptable guidance or none at all'. [348] Hill doubts whether, when used on its own, Kant's formula can provide 'even a loose and partial action guide'. [349]

Because these people believe that Kant's formula cannot provide a criterion of wrongness, some of them suggest that Kant was not trying to provide such a criterion. Kant's formula, Herman suggests, may be intended only to show that there is a 'deliberative presumption' against acting in certain ways for certain reasons. [350] O'Neill suggests that Kant's formula may be intended to provide a test,

not of which acts are wrong, but only of which acts have moral worth. [351]

Kant, I believe, had more ambitious aims. Our acts are in one sense right or wrong when, in Kant's words, these acts *conform with duty* or are *contrary to duty*. This is the sense of 'right' and 'wrong' with which Kant's formula is concerned. While discussing or applying his formula, Kant writes:

> to inform myself in the shortest and yet infallible way. . . whether a lying promise is in conformity with duty, I ask myself: would I indeed be content that my maxim. . . should hold as a universal law? [352]

> someone feels sick of life. . . but [asks] himself whether it would not be contrary to his duty to himself to take his own life. [353]

> he still has enough conscience to ask himself, is it not forbidden and contrary to duty? [354]

> he asks himself whether his maxim of neglecting his natural gifts. . . is consistent with what one calls duty. [355]

Kant also claims that his formula

> determines quite precisely what is to be done. . . with respect to all duty in general', [356]

and that

> common human reason, with this compass in hand, knows very well how to distinguish in every case what is good and what is evil, what conforms with duty or is contrary to duty. [357]

These last claims are overstatements. But so, I believe, are the claims that, as a criterion of wrongness, Kant's formula is worthless, and cannot be made to work. Kant's formula *can* be made to work. When revised in some wholly Kantian ways, this formula is, I shall argue, remarkably successful.

In asking how we should revise our two versions of Kant's formula, we can first restate the Mixed Maxims Objection. To judge whether some act is wrong, we must know all of the *morally relevant* facts. It is not enough to know, for example, that some man moved one of his fingers, or that, in moving this finger, this man pulled the trigger of some gun, or that he thereby killed someone. We must know some other facts, such as whether this man was intending to kill this other person, and, if so, whether he was acting in self-defense, and, if so, whether he was

defending himself while attacking someone else.

Of the maxims that Kant discusses, as I have said, most involve some *policy* which could be acted on in several cases. Kant's formula assumes that, to judge whether someone's act is wrong, it is enough to know on which policy this person is acting. That is sometimes true. It would be enough to know that someone is acting on the policy 'Torture others for my own amusement'. But in many other cases Kant's assumption fails. If all we know is that my Egoist is acting on the policy 'Do whatever would be best for me', we cannot possibly decide whether this man is acting wrongly. We don't know whether this man is killing someone, saving someone's life, stealing, paying some debt, or putting on warmer clothing. And if all we know is that Kant has acted on the policy 'Never lie', we don't know whether Kant has told some would-be murderer where his intended victim is, or has merely told someone the correct time of day. As these examples show, if all we know is the policy on which someone is acting, we often don't know all of the morally relevant facts.

There is another problem. When we ask whether some act is wrong, or contrary to duty, Kant's formula often makes the answer depend on morally *irrelevant* facts. When my Egoist risks his life to save some drowning child, it is irrelevant that he is acting on the policy of doing whatever would be best for himself. When Kant told someone the correct time, it was irrelevant that he was acting on the policy 'Never lie'. These facts at most give us reasons to believe that in some *other* cases this Egoist would and Kant might act wrongly.

For Kant's formula to succeed, it would have to be true that there are no maxims or policies on which it would be sometimes but not always wrong to act. That is obviously false. So Kant's formula should not appeal to the agent's maxim, in the sense of 'maxim' that can refer to policies.

Some writers suggest that, rather than appealing to the agent's actual maxim, Kant's formula should appeal to the possible maxims on which the agent might have acted. In its law of nature version, Kant's formula might then become

> LN2: We act wrongly unless what we are doing is something that we could have done while acting on some maxim on which we could rationally will everyone to act. [358]

This formula avoids the Mixed Maxims Objection. When my Egoist saves the drowning child, and Kant tells most people the truth, they could have been acting on maxims on which they could rationally will everyone to act. But if we appeal to LN2, we lose our partial answer to the Rarity Objection. Return to the case in which

I wrongly steal from a white-dress-wearing strawberry-eating woman. What I am doing is something that I could have done while acting on a maxim of stealing from white-dress-wearing strawberry-eating women, whenever I can. I could rationally will it to be true that everyone acts on this maxim, since such acts would at most be very rare. So LN2 mistakenly permits my act. Similar claims apply to other cases. When people act wrongly, there is always some possible maxim on which these people *might* have acted which they could have rationally willed to be universal. So LN2 fails to condemn all wrong acts.

To avoid this objection, we can revise Kant's formulas in a simpler way. Kant's Law of Nature Formula can become

> LN3: We act wrongly unless we are doing something that we could rationally will everyone to do, in similar circumstances, if they can.

Kant's Moral Belief Formula can become

> MB2: We act wrongly unless we could rationally will it to be true that everyone believes such acts to be morally permitted.

These formulas avoid the Mixed Maxims Objection. When my Egoist saves the drowning child, and Kant tells someone the correct time, they could rationally will it to be true both that everyone acts in these ways, and that everyone believes such acts to be permitted. So these formulas do not mistakenly condemn these acts.

These revised formulas also avoid the Rarity Objection. When we apply these formulas to someone's act, we must describe this person's act in the morally relevant way. Suppose that, being a whimsical kleptomaniac, I really *am* acting on the maxim of stealing from white-dress-wearing strawberry-eating women, whenever I can. This maxim does not provide the morally relevant description of my act. It is irrelevant that I am stealing from someone who is a woman, and who is wearing white and eating strawberries. The relevant facts may be that I am stealing from someone who is no richer than me, merely for my own amusement. In applying these revised formulas, we should ask whether I could rationally will it to be true that everyone acts in this way, and that everyone believes such acts to be permitted. If the answer is No, as we can plausibly claim, these revised formulas would rightly condemn my act.

In many cases, to give the morally relevant description of some act, it is enough to describe what the agent is, or would be, *intentionally doing*. We must describe this person's immediate aims, or what this person is directly trying to achieve. We should also describe the effects which this person believes that his or her acts might have. What people *intentionally do* is not the same as what they *intend*. To give Sidgwick's example, if some Russian revolutionary in the late 19th Century blows up

the train on which the Czar is travelling, this man may be intending only to kill the Czar. But what this man is intentionally doing is blowing up this train knowing that, as well as killing the Czar, he will kill many other people. [359]

When we describe people's acts, we are usually describing what these people are intentionally doing. [360] It is sometimes unclear what is the morally relevant description of some act. It may be unclear, for example, how much we ought to include in our list of some act's foreseeable effects, or what we ought to describe as separate acts or as parts of a single complex act. And to decide whether some act is wrong, we sometimes need to know not only *what* someone is intentionally doing, but also *why* this person does what he or she is doing. To illustrate both these points, we can suppose that some sadist saves someone's life so that he can then kill this person in a more painful way. It may not be enough to claim that what this sadist is intentionally doing is saving someone's life.

When it is unclear whether some fact is morally relevant, it often does no harm to include this fact in our description of some act. But when we apply certain moral principles to some act, it can be important not to include morally irrelevant facts. To apply both LN3 and MB2, as I have said, we must give the right description of what people are doing. Similar claims apply to other moral principles, such as principles about the wrongness of lying, stealing, and breaking promises. It is sometimes unclear which acts should be regarded as being of these kinds. But we need not answer these questions here. My main claim is that, in many cases, the agent's maxim does *not* give us the morally relevant description of some act.

On my proposed revisions of Kant's formulas, we no longer use Kant's concept of a maxim. It might be suggested that we could use the word 'maxim' in a narrower sense, which does not cover the policy on which someone is acting, but refers only to what this person is doing. Kant sometimes uses 'maxim' in this way, as when he discusses the maxim 'Kill myself to avoid suffering'. This maxim is not a policy, since we could act on it only once. [361] But this narrower sense of 'maxim' would add nothing to the morally relevant descriptions of people's acts.

We can now add one more objection to Kant's use of the concept of a maxim. When people act, there is often *no* policy on which these people are acting. If we used the word 'maxim' to refer only to policies, we would have to admit that there are many *maximless* acts. To be able to cover such acts, Kant's formulas must often use the word 'maxim' to refer, not to some policy, but to what someone is doing, on the morally relevant description of this person's act. Since Kant's formulas must often be applied directly to people's acts, it is hard to see why these formulas should ever refer to people's policies *rather* than their acts.

It might be objected that, if we revise Kant's formulas by dropping the concept of a

maxim, we are no longer discussing Kant's view. This claim is true, but no objection. We are asking whether Kant's formulas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise these formulas in ways that are clearly needed, we are developing a Kantian moral theory. And Kant's use of the concept of a maxim is not, I believe, a valuable part of Kant's own theory. In ceasing to use this concept, we are not losing anything worth keeping.

Some people might question that last claim. Kant's appeal to the agent's maxim, O'Neill writes, is not 'a detachable or dispensable part of Kant's theory', since this feature of Kant's view enables us to claim that, when some wrong-doer wills that his bad maxim be universal, there is a contradiction in this person's will. We can thereby argue that wrong-doing involves 'failures to have coherent intentions'. [362] But as Kant points out, wrong-doers do not in fact will that their maxims be universal, so 'there is really no contradiction' in these people's wills. [363]

O'Neill also suggests that, by appealing to the agent's maxim, Kant answers the question of what are the morally relevant descriptions of people's acts. [364] But as we have seen and O'Neill elsewhere claims, [365] that is not so. If all we know is that my Egoist has acted on his maxim, we cannot possibly decide whether this man's act was wrong.

It may next be objected that, if we revise Kant's formulas so that they do not refer to maxims, we lose another valuable part of Kant's view. Kant defines a maxim as a subjective *principle* of action, and he asks whether we could will this principle to be a universal *law*. Our revisions of Kant's formulas do not refer to principles or laws. But MB2 could be restated as

> MB3: We act wrongly unless we could rationally will it to be true that everyone accepts some moral principle that permits such acts.

This revision keeps Kant's concern with principles and moral laws.

Return now to O'Neill's suggestion that, by applying Kant's formula to the agent's maxim, we can at least decide whether some act has moral worth. This suggestion has some plausibility, since an act's moral worth may depend on the agent's motive or underlying aim, which may be included in this person's maxim. When applied to my Egoist, O'Neill's suggestion rightly implies that this man's acts never have moral worth. As this man's maxim reveals, he never acts in some way because he believes this act to be his duty, nor does he act for any other moral motive.

When we turn to some other maxims, however, O'Neill's suggestion fails. Suppose that, when acting on his maxim 'Never lie', Kant tells someone the truth, at what he knows to be some great cost to himself, because he believes correctly that he has a

duty to tell this person the truth. If Kant is doing his duty, at such a cost, and his motive is to do his duty, that is more than enough to give his act moral worth. It would be irrelevant that Kant is acting on a maxim that he could not rationally will to be universal. Similar claims apply whenever people do their duty, because they truly believe their act to be their duty. It is irrelevant whether these people are acting on some maxim that they could not rationally will to be universal. Like an act's wrongness, an act's moral worth does not depend on the agent's maxim, in the sense of the policy on which this person acts.

We ought, I conclude, to revise Kant's formulas so that they do not refer to such maxims. After learning from the works of great philosophers, we should try to make some more progress. By standing on the shoulders of giants, we may be able to see further than they could.

CHAPTER 13 WHAT IF EVERYONE DID THAT?

43 Each-We Dilemmas

Though I have claimed that we ought to revise Kant's formulas, I shall go on discussing Kant's own formulas. It is worth showing that we have other reasons to revise these formulas, and many of my claims would also apply to our revised versions.

When we apply Kant's Law of Nature Formula, we ask whether we could rationally will it to be true that everyone acts on some maxim. To answer this question, we must know what the alternative would be. We might be able rationally to will that everyone acts on some bad maxim, such as 'Pay less than my fair share', if the alternative would be that everyone *except us* acts in this way. Another alternative might be that everyone continues to do whatever they are now doing. But Kant's formula would then mistakenly permit us to act on many bad maxims. If many people are already acting on some bad maxim, it would often make too little difference if this maxim were acted on by everyone. On the best version of Kant's formula, which seems to be what Kant has in mind, we should ask whether we could rationally will it to be true that some maxim is acted on by everyone rather than by *no one*. [366]

We also need to know on which *other* maxim everyone would act. We could rationally will it to be true that everyone acts on some bad maxim, if the alternative would be that everyone acted on some other even worse maxim. So we should ask whether there is some other maxim that is better, in the sense that we have stronger reasons to will it to be true that everyone acts upon it.

Kant's Law of Nature Formula works best when it is applied to maxims or acts of which three things are true:

> it would be possible for many people to act on this maxim, or in this way,
>
> whatever the number of people who act in this way, the effects of each act would be similar,
>
> these effects would be roughly equally distributed between different people.

In discussing such cases, I shall use 'we' to refer to all of the people in some group;

and I shall use 'he' and 'himself' in the senses that also apply to women. We are often members of some group of whom it is true that

> if *each* rather than none of us does what would be in a certain way *better*, *we* would be doing what would be, in this same way, *worse*.

We can call such cases *each-we dilemmas*.

It will be enough to consider cases in which each person's act would benefit one or more people. One large class of each-we dilemmas are the *self-benefiting* dilemmas that are often regrettably called *prisoner's dilemmas*. In such cases, we are members of some group of whom it is true that

> (1) each of us could either benefit himself or give some greater benefit to others,
>
> (2) these greater benefits would be roughly equally distributed between all these people,

and

> (3) what each person does would have no significant effects on what the other people do.

In such cases, if each of us benefits himself, each of us is doing what is certain to be better for himself, whatever the other people do. But if all rather than none of us act in this way, *we* are doing what is certain to be worse for all of us. None of us will get the greater benefits. These cases are *each-we* dilemmas in the sense that

> if *each* rather than none of us does what would be *better* for himself, *we* shall be doing what would be *worse* for each of us.

Put the other way around,

> if *we* do what would be *better* for each, *each* would be doing what would be *worse* for himself. [367]

These claims are not about what are misleadingly called *repeated prisoner's dilemmas*, which are much less important, as I explain in a note. [368]

Though each-we dilemmas are often overlooked, they are very common. More exactly, there are few such cases that involve only two people, or only a few people; but there are many cases that involve many people. [369]

Many such cases can be called *contributor's dilemmas*. These involve *public goods*: outcomes that benefit even those people who do not help to produce them. Some

examples are clean air, national defence, and law and order. [370] In many of these cases, if everyone contributed to such public goods, that would be better for everyone than if no one did. But it would be better for each person if he himself did not contribute. He would avoid the costs to himself, and he would be no less likely to receive the greater benefits from others. In many of these cases, the public good is that we avoid outcomes that would be bad for everyone, and the contributions that are needed are not financial, but some form of self-restraint.

There are countless actual cases of this kind. In *fisherman's dilemmas*, for example, if each fisherman uses larger nets, he will catch more fish, whatever the other fishermen do. But if all the fishermen use larger nets, the fish stocks will decline, so that, before long, they will all catch fewer fish. It would still be true, however, that it would be better for each fisherman if he uses larger nets, with the result that they all catch even fewer fish. Some other cases involve the many acts that together cause pollution, congestion, deforestation, over-grazing, soil-erosion, droughts, and overpopulation.

These cases are often overlooked because, in many such cases, there are some people to whom these claims do not apply. There may, for example, be some fishermen who are so skilful that, even when there is overfishing, these people still catch as many fish. When that is true, however, the other fishermen would still face an each-we dilemma. In my description of these cases 'everyone' means 'all the members of some group'. Claims (1) to (3) can apply to some group of people even though there are some people in the same community who, though acting in similar ways, are not members of this group.

Many each-we dilemmas do not involve choices between benefiting ourselves or giving greater benefits to others. Such cases can arise whenever people have different and partly conflicting aims. It can be true that, if each rather than none of us does what will best achieve our own aim, everyone's aims will be worse achieved. Some of these may be morally required aims. According to common sense morality, which we can call *M*, we have special obligations to give certain benefits to those people to whom we are related in certain ways. These are people such as our children, parents, pupils, patients, clients, colleagues, customers, or those whom we represent. We can call these our *M-related people*. If we ought to give some kinds of priority to the well-being of these people, we can face each-we dilemmas. In *parent's dilemmas*, for example, each of us can either benefit our own children, or give greater benefits to the children of others. If each rather than none of us gives priority to benefiting our own children, that will be worse for all our children. Many such dilemmas ride on the back of self-benefiting dilemmas. When poor fishermen all catch fewer fish, for example, that may be worse not only for them but also for their malnourished children, who would be even worse fed.

Each-we dilemmas raise both practical and theoretical problems. In some cases, the

practical problem has been at least partly solved. Some solutions are *political*, involving changes in our situation. In the case of many public goods, for example, failures to contribute have been made to be either impossible, or worse for each person, by taxation that is either unavoidable, or enforced by penalties for non-payment. In many other cases, however, political solutions cannot be achieved, or are too costly. In some of these cases, we have achieved solutions that are *psychological*, in the sense that, without a change in our situation, all or most of us choose to give the greater benefits to others. Such solutions often depend on our having and acting upon certain moral beliefs. We may contribute to some public goods, despite the costs to ourselves, because we believe that we ought to contribute.

Of these *moral* solutions to each-we dilemmas, two are especially relevant here. We might be Act Consequentialists, who believe that we ought always to give the greater benefits to others, since we shall thereby do more good. If we all acted on this moral belief, we would all contribute to such public goods. But these solutions are seldom achieved, since there are few people who are both Act Consequentialists and often act on their moral beliefs.

There are also Kantian solutions. If no one contributed to such public goods, that would be much worse for all of us than if everyone contributed. We could not rationally will it to be true that everyone rather than no one acts on the maxim 'Don't contribute'. So, if we were all conscientious Kantians who always acted on Kant's Law of Nature Formula, we would all contribute to these public goods.

When we have achieved some moral solution to some contributor's dilemma, common sense morality requires everyone to go on contributing. In such cases, there are often some *free riders*: people who benefit from these public goods, without making any contribution. Each free rider benefits himself in a way that imposes a greater total burden on others. Common sense morality condemns such acts as unfair. And these are some of the cases in which we can best think and say 'What if everyone did that?'

In *unsolved* each-we dilemmas, things are in one way different. When no one is contributing to some merely possible public good, no one is free-riding, or failing to do their fair share. But Kant's Law of Nature Formula still implies that, in failing to contribute, everyone acts wrongly. These are the cases for which this formula might have been especially designed. If everyone is failing to contribute, we could not say to each other, 'What if everyone did that?' Everyone *is* doing that. But we can ask our question in another way. Compared with a world in which everyone contributes, so that everyone gets these public goods, we could not rationally will it to be true that no one contributes, so that no one gets these goods. So Kant's formula requires us all to contribute.

When applied to such cases, Kant's formula conflicts with, and may lead us to revise,

some widely held and at least partly mistaken moral beliefs. In unsolved each-we dilemmas, most of us believe that we are either permitted or required to give the smaller benefits to ourselves, or to some of our M-related people, rather than giving the greater benefits to others. According to Kant's Law of Nature Formula, such acts are wrong. None of us could rationally will it to be true that all rather than none of us continue to act in these ways, since that would be worse for all of us, or worse for all of our M-related people.

As well as conflicting with some widely held beliefs, Kant's formula challenges these beliefs in an especially forceful way. Though Act Consequentialists would also claim that everyone ought to give the greater benefits to others, the Kantian argument for this conclusion is harder to reject. In unsolved each-we dilemmas, each of us is trying to benefit ourselves, or our children, parents, pupils, patients, or other M-related people. When judged at the *individual* level, each of us succeeds, since each of us *is* doing what is better for himself, or for his children, parents, pupils, patients, etc. But *we* are doing what is *worse* for all these people. *We* are failing, or doing worse, even in our own terms, since we are making it true that everyone's morally required aims will be worse achieved. In these cases, in acting on common sense moral principles, we are acting in ways that are *directly collectively self-defeating*. If we were Rational Egoists, that would be no objection to our view, since this form of Egoism is a theory about *individual* rationality and reasons. But moral principles or theories are intended to answer questions about what *all* of us ought to do. So such principles or theories clearly fail, and condemn themselves, when they are directly self-defeating at the collective level. [371]

Kant comes close to giving such an argument. When Kant discusses the limits on our duty to benefit others, he writes,

> a maxim of promoting the happiness of others with a sacrifice of one's own happiness. . . would conflict with itself if it were made into a universal law. [372]

Kant must mean 'with a *greater* sacrifice of one's own happiness'. His point must be that, if everyone promoted the happiness of others at a greater cost to their own happiness, everyone would lose more happiness than they gained. If the effects of such acts would be roughly equally distributed between different people, that would be true. This would be how this maxim would 'conflict with itself'. A similar point applies to a maxim of promoting one's own happiness at a greater cost to the happiness of others. On similar assumptions, if this maxim were a universal law, it would also conflict with itself. There would be only one maxim that could be made universal without conflicting with itself, or being collectively self-defeating. This would be the maxim of doing whatever would, on the whole, best promote everyone's happiness. [373]

Kant's formula has even greater value when it is applied to one kind of unsolved

each-we dilemma. In many cases,

> (4) each of us could benefit ourselves or our M-related people in ways that would impose a greater total sum of burdens on others. But these burdens would be spread over very many people. So each act would impose burdens on each of these other people that would be trivial, and would often be imperceptible.

These claims are true in most of the contributor's dilemmas mentioned above. When we know that our acts would impose only such trivial or imperceptible burdens on each of many other people, our ordinary concern for others would not be aroused. Even if we were conscientious Act Consequentialists, we would be likely to ignore such effects. But when many of us act in these ways, the combined effects may be very great and very bad. One example is the way in which, by using fossil fuels, we are recklessly and selfishly overheating the Earth's atmosphere. In such cases, Kant's Law of Nature Formula can act like a moral microscope, getting us to see what we are doing. We could not rationally will it to be true that we together inflict such damage on ourselves, our children, and our children's children. [374]

44 The Threshold Objection

We can now turn to some cases in which Kant's formulas do less well. According to Kant's

> *Law of Nature Formula*: It is wrong to act on some maxim unless we could rationally will it to be true that everyone acts upon it.

In some cases, however, whether some act is wrong may depend on how many people act in this way. When that is true, Kant's formula may fail, by condemning acts that are right, or permitting acts that are wrong.

In discussing such cases, it will be enough to consider acts whose rightness depends in part on their predictable effects. There are many maxims of which it is true that

> (5) if too many people acted on this maxim, these people's acts would have bad effects, but when fewer people act on this maxim the effects are neutral or good.

It may then be true that

> (6) though such acts would be wrong if too many people acted on this maxim, when fewer people act on this maxim such acts are permissible, and may even be morally required.

In such cases,

> (7) most of us could not rationally will it to be true that everyone acts on these maxims.

Kant's formula may mistakenly condemn such acts when they are permissible or even morally required.

One example is the maxim 'Have no children, so as to devote my life to philosophy'. If Kant acted on this maxim, he did not act wrongly. But he could not have rationally willed it to be true that everyone acts on this maxim, so Kant's formula seems to imply that Kant's deliberate failure to have children would have been wrong. [375] Consider next the maxims: 'Consume food without producing any', 'Become a dentist', and 'Live in Iceland, to absorb the spirit of the Nordic Sagas'. [376] It is not wrong, in the world as it is, to act on these maxims. But since we could not rationally will it to be true that everyone acts on these maxims, Kant's formula seems to imply that such acts are wrong. Other examples are: 'Don't take the first slice', 'Don't speak until others have spoken', and 'When you meet another car on a narrow road, stop and wait until the other car has passed'. We could not rationally will it to be true that everyone acts on these maxims. In such a world, cakes would never get eaten, conversations would never get started, and some people's journeys would never end. But acting on these maxims is not, in the actual world, wrong.

Since this problem is raised by acts that are wrong only if the number of such acts is above some rough threshold, we can call this the *Threshold Objection*.

Thomas Pogge suggests that, to answer this objection to Kant's view, we should turn from Kant's Law of Nature Formula to his Moral Belief Formula. [377] Though we could not rationally will it to be true that everyone *acts* on such maxims, we *could* rationally will it to be true that everyone believes such acts to be morally permitted. Even if everyone had these beliefs, there is no danger that too many people would choose to act in these ways. Most people already believe that they are permitted to act on the maxims that I have just mentioned. But enough people are having children and producing food. Nor are there too many dentists or inhabitants of Iceland, or too many polite people who always let other people eat, speak, or go first. Since we could rationally will it to be true that everyone believes such acts to be permitted, Kant's Moral Belief Formula permits these acts.

These claims are not, I believe, a sufficient answer to this objection. If none of us had children, we would be ending human history. If none of us produced food, we would be ending history more brutally, by letting ourselves and our children starve to death. These are not merely consequences that we could not rationally will. If we all acted in these ways, we would be acting wrongly. Nor could we rationally will it to be true that everyone falsely believes that these acts would not be wrong.

It is not enough to say that, even if we all had these false beliefs, there is no danger that too many of us would act in these ways. We always have some reason to want ourselves and others not to have false moral beliefs, and these are not cases in which we have any contrary reason.

Pogge suggests another answer to this objection. Many maxims are *conditional*, in the sense that we intend to act in some way only when our acts would have certain effects. Such maxims would not apply when our acts would not have these intended effects, or would have certain other, bad effects. Our maxims may be implicitly conditional in such ways even if we have not had conscious thoughts about these conditions. It is enough that, if these conditions were not met, we would not act on these maxims, and would not have changed our mind.

Of the actual maxims that Kant's Law of Nature Formula may seem mistakenly to condemn, most are at least implicitly conditional. If we intend to produce no food, that intention would not apply if we were starving. Our maxim is something like 'Produce no food as long as enough other people are producing food.' We could rationally will it to be true that everyone acts on this maxim, so Kant's formula does not imply that, in failing to produce food, we are acting wrongly.

We can also assume that, of those who accept the maxim 'Become a dentist', most intend to act on this maxim only if they could thereby earn a living. Perhaps we could rationally will it to be true that everyone accepts this conditional maxim, since we would know that, in the case of most people, this maxim's condition would not be met. But Kant's Law of Nature Formula would here make our moral reasoning take a rather strange form. And we have some reason *not* to will it to be true that everyone accepts this maxim. That would be to will a world whose entire population wanted to become dentists, so that most people had the disappointment of an unfulfilled ambition because there was no room for them in the dental profession. It would be more plausible to follow Pogge's first suggestion, by turning to Kant's Moral Belief Formula. Anyone is permitted to act on this conditional maxim, we might claim, because everyone could rationally will it to be true that everyone believes such acts to be permitted. That is a better way to explain why, in a world with teeth to be filled, becoming a dentist is not wrong.

We have not yet fully answered the Threshold Objection. Though most people's maxims take such conditional forms, there are some exceptions. Kant may have believed that, since most other people could be relied upon to have children, it was permissible for him to abstain. [378] But of those who choose to have no children, some act on maxims that are unconditional. And moral principles ought to apply successfully to cases that are merely imaginary, when it is clear enough what such cases would involve. We can imagine fanatical, unconditional maxims whose universal acceptance would lead us all to become childless underemployed Icelandic dentists who starved themselves to death. Since we could not rationally will it to be

true that everyone acts on these unconditional maxims, or believes such acts to be permitted, Kant's formulas mistakenly condemn our acting on these maxims even when we know that, because few people are acting on these maxims, our acts will have good effects.

This is not, however, a new objection. Like the Egoist's maxim 'Do whatever would be best for me' and Kant's maxim 'Never lie', these are *mixed maxims*, on which it would be sometimes but not always wrong to act. To answer this objection, I have claimed, we should make Kant's formulas apply, not to maxims in the sense that can refer to policies, but to the morally relevant description of what people are doing. On our revised version of Kant's Law of Nature Formula,

> LN3: We act wrongly unless we are doing something that we could rationally will everyone to do, in similar circumstances, if they can.

Suppose that, in acting on these unconditional maxims, we would be having no children, or producing no food, in circumstances in which we knew that there were not too many people who were acting in these ways. We could rationally will it to be true that everyone acts in these ways, in similar circumstances, if they can. In such a world, there would not be too many people who acted in these ways. So LN3 would not mistakenly imply that these acts would be wrong.

45 The Ideal World Objections

There is another kind of case in which an act's wrongness may depend on the number of people who act in this way. It may be true that

> (8) if enough people acted in some way, these people's acts would have good effects, but when fewer people act in this way the effects would or might be very bad.

It may then be true that

> (9) we ought to act in this way if enough people are doing that, but in other cases such acts are wrong.

Kant's Law of Nature Formula, many writers claim, requires some such acts even when they are clearly wrong.

Consider first the maxim 'Never use violence'. Kant's formula, it is sometimes claimed, requires us to act on this maxim, since there is no other conflicting maxim on which we could rationally will everyone to act. If that were true, Kant's formula would require us never to use violence.

Pacifism has considerable intuitive appeal. And many people (one of them my father) have been pacifists on Kantian grounds. But like Kant's belief that we must never lie, pacifism is too simple. Return to the time of the Second World War. If everyone outside Germany had been pacifists, that would have allowed Hitler to dominate the world, with effects that would have been likely to be even worse than this terrible war. If Kant's Law of Nature Formula implied that it was wrong to fight against Hitler's armies, that would count against this formula.

Suppose next that, in

> *Mistake*, several people's lives are in danger. You and I must choose between two ways of acting. The possible outcomes are these:

		I: do A	I: do B
You	do A	we save everyone	we save no one
	do B	we save no one	we save some people

We ought both to do A, since that is our only way to save everyone. But suppose that, because you misunderstand our situation, you do B. Despite knowing that you have made this mistake, I do A, with the result that we save no one. I know that, by doing A, I shall prevent us from saving some people whom we would have saved if I had done B. But as a Kantian, I believe that I ought to do A, since that is the only thing that I could rationally will us both to do. [379]

If Kant's formula implied that I ought to do A, despite knowing that you have done B, that implication would be wholly unacceptable. While pacifism has some plausibility, it would be absurd to claim that I ought here to do A, thereby letting some people die whom we could have saved.

These examples illustrate another objection to Kant's Law of Nature Formula. Kant's 'standard of conduct', Korsgaard writes,

> is designed for an ideal state of affairs: we are always to act as if we were living in the Kingdom of Ends, regardless of possible disastrous results. [380]

Korsgaard takes this problem to be raised by the fact that some people act wrongly. But as *Mistake* shows, this objection to Kant's formula is not raised only by deliberate

wrong-doing. Though this case is artificially simple, there are many actual cases of this kind. It is often true that, if we did what we could rationally will everyone to do, as Kant's formula is claimed to require, our acts would predictably have bad effects of a kind that would make them wrong. Discussing such cases, Hill writes:

> The problem is that acting in this world by rules designed for another can prove disastrous. [381]

According to what we can call this

> *Ideal World Objection*: Kant's formula mistakenly requires us to act in certain ways even when, because some other people are *not* acting in these ways, our acts would make things go very badly, and for no good reason.

In discussing this objection, it will be enough to consider cases in which, as in *Mistake*, it would be best if all of the relevant people acted in the same way. [382] Consider this maxim:

> M1: Do whatever I could rationally will everyone to do.

According to the Ideal World Objection, compared with willing that everyone acts on M1, we could not rationally will that no one does. If this claim were true, Kant's formula would require us to act on M1 even when, as in *Mistake*, our acts would predictably have very bad effects.

This claim is not, however, true. Here is a better maxim:

> M2: Do whatever I could rationally will everyone to do, unless some other people haven't acted in this way, in which case do whatever I could rationally will that, in these circumstances, people do.

I could rationally will it to be true that everyone acts on M2. In *Mistake*, we would both act on M2 if we both did A, since that is how we could save everyone's lives. But I know that you haven't acted in this way, since you have mistakenly done B. Given your mistake, I could not rationally will that I do A, thereby preventing us from saving anyone. To follow M2, I must do B, thereby enabling us to save at least some people. Since Kant's formula permits me to act on M2 rather than M1, this formula permits me to respond to your mistake in what is obviously the right way.

Return next to the pacifist maxim 'Never use violence'. According to the Ideal World Objection, Kant's formula requires us to act on this maxim, since there is no other conflicting maxim on which we could rationally will everyone to act. As before, that is not so. Here is a better maxim:

> Never use violence, unless some other people have used aggressive violence,

> in which case use restrained violence when that is my only possible way to defend myself or others.

Everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which no one ever uses violence. So Kant's formula does not require us to be pacifists, but permits us to use restrained violence to resist aggression.

Similar claims apply to all such cases. Kant's formula never requires anyone to act on unconditional maxims like M1 or the pacifist maxim. Everyone could rationally will it to be true that everyone acts on conditional maxims like M2 or the maxim of resisting aggression. In acting on such maxims, as Kant's formula permits, we could respond in the best ways to the wrong acts or mistakes of other people.

There is, however, another problem. Kant's Law of Nature Formula merely *permits* us to act on these better maxims. Consider this maxim:

> Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can.

As before, everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which no one ever uses violence. But in the real world some people have used aggressive violence. Since this maxim passes Kant's test, Kant's formula permits the rest of us to act upon it, by killing as many people as we can. Consider next:

> Keep my promises, and help those who are in need, unless some other people haven't acted in these ways, in which case copy them.

This maxim also passes Kant's test. Everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which everyone kept their promises and helped those who were in need. [383] In the real world, however, some people haven't acted in these ways. Since this maxim passes Kant's test, Kant's formula mistakenly permits the rest of us to copy these other people, by breaking all our promises and never helping those who are in need.

To state this problem in a simpler way, we can turn to

> M3: Do what everyone could rationally will everyone to do, unless some other people haven't acted in these ways, in which case do whatever I like.

Since everyone could rationally will it to be true that everyone acts on M3, this maxim passes Kant's test. We know that, in the real world, some people haven't

acted on M3, since these people haven't done what everyone could rationally will them to do. So, in permitting us to act on M3, Kant's formula permits the rest of us to do whatever we like.

According to the Ideal World Objection, Kant's formula sometimes requires us to act as if we were in an ideal world even when, in the real world, such acts would have disastrous effects, and would be clearly wrong. We can answer that objection by applying Kant's formula to conditional maxims, as we often need to do for other reasons. But we have now found that, when applied to such maxims, Kant's formula requires too little. According to this

> *New Ideal World Objection*: Once a few people have failed to do what we could rationally will everyone to do, Kant's formula ceases to imply that any act is wrong.

If this objection cannot be answered, it would be at least as damaging.

Similar claims apply to some other moral principles or theories. According to one version of *Rule Consequentialism*, or

> RC: Everyone ought to follow the rules whose being followed by everyone would make things go best. [384]

We *follow* some rule when we succeed in doing what this rule requires us to do. It is often objected that RC requires us to follow these *ideal rules* even when we know that, because some other people are not following these rules, our acts will have disastrous effects. This objection can be answered. Consider

> R1: Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever, given the acts of others, would make things go best.

This is one of the ideal rules, since everyone's following R1 would make things go best. [385] So RC does *not* require us to follow those ideal rules whose being followed by only some people would have disastrous effects. But consider

> R2: Follow the rules whose being followed by everyone would make things go best, unless some other people have not followed these rules, in which case do whatever you like.

Since R2 is *also* one of the ideal rules, RC permits us to follow this rule. We know that, in the real world, some people have not followed the ideal rules. So in permitting us to follow R2, RC permits the rest of us to do whatever we like. Similar objections apply to most other versions of Rule Consequentialism, such as those theories which appeal to the rules whose being *accepted* by everyone, or by

most people, would make things go best. [386] And similar objections apply to some Contractualist moral theories.

To answer this new objection to Kant's Law of Nature Formula, we should again revise this formula. When we apply this formula to some maxim, it is not enough to ask whether we could rationally will it to be true that *everyone* acts upon it. Kant's formula could become:

> *LN4*: It is wrong for us to act on some maxim unless we could rationally will it to be true that this maxim be acted on by everyone, and by *any other number* of people, rather than by no one.

For some maxim to pass this wider test, we must be able rationally to will that this maxim be acted on, not only by *everyone* rather than by no one, but also by *most* people rather than by no one, by *many* people rather than by no one, by a *few* people rather than by no one, and by any other number of people rather than by no one. We must be able rationally to will that, *whatever* the number of people who *don't* act on this maxim, *everyone else* does. [387]

If we widen Kant's formula in this way, it condemns the bad maxims that we have discussed. One example is:

> Do not use violence, unless some other people have used aggressive violence, in which case kill as many people as I can.

Though we could rationally will it to be true that *everyone* acts on this maxim, we could not rationally will that any other number of people act upon it. If anyone uses aggressive violence, everyone else would act on this maxim by killing as many people as they can.

When we consider many maxims and acts, this revision of Kant's formula would make no difference. There are many acts that are right whatever the number of people who act in this way. In such cases there are unconditional maxims on which we could rationally will any number of people to act. Some examples are the maxims 'Help those who are in need' and 'Never injure others merely for my own convenience'. As we have seen, however, when we consider some other kinds of act, what we could rationally will is that people act on conditional maxims which tell us to take into account the acts of others. Some such maxims could take this form:

> Do A, unless the number or proportion of A-doers is or will be below some threshold, in which case do B, or below some other threshold, in which case do C.

Some of these thresholds could be defined as the numbers or proportions of A-doers below which acts of kind A would cease to have certain good effects, or would start to have certain bad effects.

Similar claims apply to Rule Consequentialism. The formula stated above could become

> RC2: Everyone ought to follow the rules whose being followed by any number of people rather than by no one would make things go best.

Some of these rules could take such conditional forms. These rules would tell us to act in the ways that would make things go best, given the number or proportion of people who are following these rules. [388] Similar claims would apply to those versions of RC which appeal to what would happen if people *accepted* certain rules.

This revision makes Rule Consequentialism in some ways closer to Act Consequentialism. That is most importantly true when we ask what proportion of their income or wealth the world's rich people ought to give to the more than a billion people who now live on around $2 a day. When applied to this question, most versions of Rule Consequentialism are not very demanding. These theories appeal to claims about what would be true if *all* or *most* people accepted or followed certain principles. Things might go best if all or most rich people gave to the poor some fairly modest proportion of their wealth or income, such as one fifth, or even one tenth. That would make a great difference, since the richest nations now give less than one per cent. If we revise Rule Consequentialism by changing 'all' or 'most' to 'any number of people', and we appeal to conditional rules of the kind just mentioned, Rule Consequentialism would often be much more demanding. If most rich people are not giving what it would be best for the rich to give, the best rule would require the others to give a great deal. [389]

In revising Kant's Law of Nature Formula in this way, we give up the idea expressed in the question 'What if everyone did that?' But this idea can be successfully applied only to certain kinds of case. In each-we dilemmas, if we are free-riders who fail to contribute to some public good, we can be rightly challenged with the question 'What if everyone did that?' But in many other cases, it is enough to reply 'Most people won't'. [390]

Kant's Moral Belief Formula appeals to a different idea, which might be successfully applied to all kinds of case. Though we cannot plausibly assume that everyone ought to act on the same maxims, or in the same ways, we *can* plausibly assume that everyone ought to have the same moral beliefs. [391] So when people object to one of our moral beliefs, saying 'What if everyone thought like you?', it is *not* enough

simply to reply 'Most people won't'. If we could not rationally will it to be true that everyone believes some kind of act to be permitted, this fact might, as Kant assumes, show such acts to be wrong. [392]

We can now turn to some simpler and more fundamental questions.

CHAPTER 14 IMPARTIALITY

46 The Golden Rule

When describing how his Formula of Universal Law explains our duty to benefit others, Kant writes

> I want everyone else to be beneficent toward me; hence I ought also to be beneficent toward everyone else. [393]

This may remind us of

> *The Golden Rule*: We ought to treat others as we would want others to treat us.

This rule expresses what may be the most widely accepted fundamental moral idea, which was independently discovered in at least three of the world's earliest civilisations. [394] Though Kant calls his formula 'the supreme principle of morality', he dismisses the Golden Rule as 'trivial' and unfit to be a universal law. [395] Does this rule deserve Kant's contempt?

In rejecting the Golden Rule, Kant writes:

> It cannot be a universal law, because it does not contain the ground of duties toward oneself, nor that of duties of love toward others (for many a man would gladly agree that others should not benefit him if only he might be excused from benefiting them); and finally it does not contain the ground of duties owed to others, for a criminal would argue on this ground against the judge who punishes him.

According to one of Kant's objections, the Golden Rule does not imply that we have duties to benefit others. Many people, Kant claims, would gladly agree never to be benefited by others.

This objection backfires. These people ought to benefit or help others, the Golden Rule implies, if they themselves would want to be helped. Kant does not deny that these people would want to be helped. He makes the different claim that these people would agree not to be helped *if they would thereby be excused* from helping others. To state this claim in Kantian terms, these people would will it to be true

that the maxim of not helping others be a universal law. That does not imply that, according to the Golden Rule, these people have no duty to help others. It is *Kant's* formula, not the Golden Rule, that permits us to act on maxims that we could will to be universal laws.

Kant's objection might be revised. He might ask us to consider people who do *not* want to be helped by others, whether or not they would thereby be excused from helping others. Kant might then claim that, since these people do not want to be helped, the Golden Rule fails to imply that they have a duty to help others.

As before, however, this objection would apply to Kant's own formula. According to this formula, these people ought to help others if they could not will it to be true that the maxim of not helping others be a universal law. If these people do not even want to be helped, they could more easily will that this maxim be such a law. No one could will such a law, Kant claims, because such a person would thereby 'rob himself of all hope of the assistance that he wishes for himself.' [396] This claim does not apply to people who *don't* wish to be helped.

Kant might reply that, in not wishing or wanting to be helped, these people would be irrational. And he might then argue that, when applied to such people, his formula does better than the Golden Rule. Kant might claim that, since the Golden Rule appeals to these people's desires, which are irrational, this rule fails to imply that these people have a duty to help others. In contrast, because these people could not *rationally* will it to be true that they would never be helped, Kant's formula does imply that they have this duty.

This objection to the Golden Rule has no force. We can first explain why, in most of its stated versions, this rule does not appeal to how we would *will* that others treat us. We are not absolute monarchs or dictators, who can successfully will it to be true that other people act in some way. Since we do not have such power over others, we can only want or wish it to be true that other people act in some way. Kant's formula asks us to imagine or suppose that we have the power to choose, or will it to be true, that other people act in some way. The Golden Rule could take the same form. This rule need not appeal to our desires, but could appeal to how, if we had the choice, we would will that we ourselves be treated---or how we would be *willing* to be treated. Some familiar statements of the Golden Rule, such as 'Do as you *would be* done by', already take this form.

The Golden Rule can also appeal to what we would *rationally* choose, or will. It is true that, as commonly stated, this rule does not use the concept *rational*. But of Kant's many statements of his formula, only two use this concept, and none explicitly appeal to what we could rationally will. Given some of Kant's other claims, Kant clearly intends us to ask what we could rationally will or choose. The Golden Rule could take the same form. This rule could be stated as

> G2: We ought to treat others only in ways in which we would rationally be willing to be treated by others.

When we apply the Golden Rule, it is sometimes enough to ask whether we would be willing, in the actual world, to be treated in some way. Torturers, for example, would not be willing to be tortured. But when considering many kinds of act, we must ask how we would be willing to be treated in some merely imaginary case. When we could feed someone who is starving, for example, it is not enough to ask whether we would be willing to be given no food. If we have just eaten well, and have a well-stocked kitchen, our answer to that question might be Yes. We should ask whether, even if we were starving, we would be willing to be given no food.

Consider next some white racist who, in the worst period of racial discrimination in the Southern USA, excludes black people from his hotel. This man might claim to be obeying the Golden Rule. He might say:

> We ought to treat others only as we would be willing to be treated by others. I admit to my hotel anyone who is not black. I would be willing to be treated in this way. I *am* treated in this way. Since I am not black, I am admitted to every hotel.

This speech misunderstands the Golden Rule. On this rule, this man ought to treat black people only as he would be willing to be treated *if he were going to be in their position.* He must imagine either that (1) all hotels are owned by black people who exclude white people, or that (2) he himself is black. Though (1) would be merely a change in his circumstances, (2) would be a change in him. When we apply the Golden Rule to many other cases, the imagined change would have to be in ourselves, since we must imagine being relevantly *like* the people whom our acts would affect, by having these people's desires, attitudes, and other physical or psychological features. For example, for some man to imagine being treated as he treats women, he may have to imagine that he is a woman. Similar claims apply to sado-masochists.

In a fuller statement, then, the Golden Rule could be

> G3: We ought to treat others only in ways in which we would rationally be willing to be treated, if we were going to be in these other people's positions, and would be relevantly like them.

The phrase 'would be willing' can be misleading. In applying G3, we should not ask how, if we were in these other people's positions, we would *then* be willing to be treated. We should ask how we would *now* be willing to be treated later, if we were later going to be in these people's positions. (If I similarly said 'Would you want your organs to be used after you are dead?', I would be asking you, not to

predict your *post mortem* desires, but to make a decision now.)

Kant gives another objection to the Golden Rule. By appealing to this rule, Kant claims, 'a criminal could argue against the judge punishing him'. Kant must be assuming here that this criminal could say: 'Since you would not want to be punished, you ought not to punish me.' This objection takes the Golden Rule to be

> G4: We ought to treat *each* other person as we would rationally be willing to be treated, if we were going to be in this person's position, and we would be relevantly like this person.

Kant would be right to reject *this* rule. Suppose that, in

> *Case One,* I could save either Blue's life, or Brown's.

By appealing to G4, Blue could argue that I ought to save her life. I would not be willing to be left to die if I were going to be in Blue's position. Brown could similarly argue that I ought to save her life. So G4 mistakenly implies that, whatever I do, I shall be acting wrongly, by failing to treat either Blue or Brown as I ought to do. Suppose next that, in

> *Case Two*, I have a small loaf of bread, and meet two starving people.

By appealing to G4, each person could argue that I ought to give her my whole loaf.

When Jesus appealed to the Golden Rule, was he appealing to G4? Was he intending to imply that it would be wrong for me to share my loaf between these people? The answer is clearly No. The Golden Rule should be taken to mean, not G4, but

> G5: We ought to treat *other people* as we would rationally be willing to be treated if were going to be in the positions of *all* of these people, and would be relevantly like them.

In this better form, however, this rule is harder to apply. How are we to imagine that we shall be in the positions of two or more people?

Several suggestions have been made. Suppose that, in

> *Case Three*, I could either save Green's life, or save Grey from going blind.

On Nagel's proposal, I should imagine that, like an amoeba, I shall later divide and become two people, one in Green's position and the other in Grey's. [397] On Richard Hare's proposal, I should imagine that I shall later live lives that would be just like those of Green and Grey, not simultaneously, but one after the other. [398] On John

Harsanyi's proposal, I should imagine that I shall have an equal chance of being in either Green's position or in Grey's. On Rawls's proposal, I should imagine that I shall be in one of these people's positions, but with no knowledge of the probabilities. [399]

When we apply the Golden Rule to certain questions, it might make a difference which of these proposals we adopt. But in most cases these proposals would have the same implications. In *Case Three* for example, in whichever of these ways I imagine that I shall be in the positions of Green and Grey, I would not be willing to be saved from blindness in one of these positions rather than being saved from death in the other.

Of those who have appealed to the Golden Rule, many may not have considered the difference between G4 and G5. But if these people had compared these claims, and seen what they imply, they would have regarded G5 as better stating the moral idea that they had in mind.

Return now to Kant's claim that, by appealing to the Golden Rule, a criminal could argue that his judge ought not to punish him. On the better reading of the Golden Rule, as expressed in G5, judges could reject this argument. [400] These judges should ask how they would rationally be willing to be treated if they were going to be, not only in some criminal's position, but also in the positions of all of the other people whom their decision might affect. These other people include the possible victims of the crimes that would be more likely to be committed if this criminal is not punished, either because this criminal would be free and able to commit some other crime, or because he and other potential criminals would be less likely to be deterred. Since this is how judges ought to apply the Golden Rule, this rule does not mistakenly imply that no one should be punished.

According to Kant's remaining objection in the passage quoted above, the Golden Rule cannot be a universal law because this rule does not cover our duties to ourselves. We might reply that, since this rule applies only to our treatment of other people, it does not claim to cover our duties to ourselves. As Kant elsewhere suggests, however, this feature of the Golden Rule may make it misdescribe some of our duties to others. [401] Suppose that, in

> *Case Four*, I could either save my own life or save Grey from going blind.

If the Golden Rule tells me only how I ought to treat *other people,* this rule might mistakenly imply that I ought to save Grey from blindness at the cost of my life. This might be what I would be willing to have done if I were going to be only in Grey's position.

To meet this objection, this rule could become

> G6: We ought to treat *everyone* as we would rationally be willing to be treated if we were going to be in all of these people's positions, and would be relevantly like them.

The word 'everyone' here refers to all of the people whom our acts might affect. In many cases, *we* are one of these people. On this version of the Golden Rule, when applied to *Case Four*, I ought to do what I would be willing to have done if I were going to be, not only in Grey's position, but also in mine. As in *Case Three,* I would not be willing to be saved from blindness in one of these positions rather than being saved from death in the other. This revision better states the Golden Rule's assumption that everyone matters equally. It is not surprising that, in most statements of this rule, we are told only to treat *others* as we would be willing that we ourselves be treated. There is little danger that we shall ignore our own well-being. But this reference to others is, in a way, misleading, since *we* are among the people whose well-being we ought to consider in the impartial way that this rule requires.[402]

Kant's contempt for the Golden Rule is not, I have argued, justified. But Kant's Formula of Universal Law might still be, as Kant believed, a better principle. Is that so?

These principles often have the same implications. And as candidates for the supreme principle of morality, both meet the most obvious requirements. Both principles succeed in most of the cases in which Kant's Impossibility Formula so spectacularly fails. Most of us could not rationally will it to be true that everyone acts on maxims of self-interested killing, injuring, coercing, lying, and stealing. Nor would we be willing to be treated in these ways if we were going to be in the positions of the affected people.

Kant's Formula of Universal Law is in two ways similar to the Golden Rule. In their best forms, both principles appeal to claims about what it would be rational for people to choose. And both principles assume that everyone matters equally, and has equal moral claims. The 'intuitive idea' behind Kant's formula, O'Neill writes, is that 'we should not single ourselves out for special consideration or treatment'.[403]

These principles mainly differ in the ways in which they make our moral thinking more impartial. Both principles tell us to carry out certain thought-experiments, by asking questions about some imagined cases. To apply the Golden Rule, we ask 'What if that was done to me?' To apply the law of nature and moral belief versions of Kant's formula, we ask 'What if everyone did that?' and 'What if everyone believed such acts to be permissible?'

When we apply the Golden Rule, our thought-experiment is fairly simple. As when making many ordinary decisions, we ask what would happen in the actual world if we acted, on one occasion, in each of certain possible ways. We don't even need to decide what are the morally relevant descriptions of these particular possible acts. But we try to think about these possibilities, not only from our own point of view, but also from the points of view of all of the other people whom our act might affect. We ask what would rationally be willing to do, and have done to us, if we were going to be in all of these people's positions, and would be relevantly like them.

Kant's thought-experiments are in several ways harder. When we apply Kant's Law of Nature Formula, we must first decide what is the maxim on which we would be acting. In my revised version of this formula, we must decide what is the morally relevant description of our act. We then compare two possible worlds, or two ways in which the future history of our world might go. We ask what would happen both if everyone acted on some maxim, and if no one did, because everyone acted on some other maxim. Similarly, when we apply Kant's Moral Belief Formula, we ask what would happen both if everyone had some moral belief, and if no one did, because everyone had some other moral belief. These four possible worlds may all be very different from the actual world, and it would often be hard to predict what these worlds would be like. We may also have to consider various other possible maxims on which everyone might act. In another way, however, Kant's formulas are easier to apply than the Golden Rule. When we ask in which of these worlds we could rationally choose to live, we think about these worlds only from our own point of view.

Kant's formulas and the Golden Rule can be usefully compared with two other principles. According to another old idea, we should make our moral reasoning impartial in a different and simpler way. We should ask what it would be rational for us to choose, or prefer, neither from our own point of view, nor from the points of view of those other people whom our acts might affect, but from the imagined point of view of some detached observer, who is not involved in the case we are considering. On a variant of this idea, we ask what it would be rational for us to choose, or prefer, when we imagine some other relevantly similar case, in which everyone involved would be strangers to us. We can call this the *Impartial Observer Formula*.

We can also achieve impartiality by applying Kant's Consent Principle. By asking whether everyone could rationally consent to some possible act, we give equal weight to everyone's reasons for refusing consent.

There are various objections to the Golden Rule. It can be difficult to imagine that we shall be in other people's positions and shall be relevantly like these other people. And what we must try to imagine would often be deeply impossible. But that is not, as some writers claim, a decisive objection. Some thought-experiments are

useful even though they ask us to imagine something that is deeply impossible. Einstein usefully asked what he would see if he were travelling at the speed of light. Though we could not possibly *be* the horse whom we are whipping, or the trapped and starved animal whose fur we are wearing, we can imagine such things well enough for moral purposes.

Another objection to the Golden Rule has more force. As Rawls points out, if we imagine that we shall be in the positions of all of the people whom our acts might affect, we shall be led to ignore the fact that, in the real world, our acts would affect different people. One person's burdens cannot be compensated by benefits to other people. In ignoring this 'separateness of persons', we are ignoring facts that may give us decisive reasons to accept principles of distributive justice. [404]

In these and some other ways, the Golden Rule is theoretically inferior to both the Impartial Observer Formula and Kant's Consent Principle. But this rule may be, for practical purposes, the best of these three principles. By requiring us to imagine ourselves in other people's positions, the Golden Rule may provide what is psychologically the most effective way of making us more impartial, and morally motivating us. That may be why this rule has been the world's mostly widely accepted fundamental moral idea.

Of these four ways of making us more impartial, Kant's Formula of Universal Law is, I shall argue, the least successful. This formula fails to condemn many wrong acts. As we shall see, however, these problems have a Kantian solution.

47 The Rarity and High Stakes Objections

When people act wrongly, they may be doing something that cannot often be done. Some of these people could rationally will it to be true that everyone acts like them, since such acts would be too rare to have significant effects on them. I have called this the *Rarity Objection*. Consider, for example,

> *Unjust Punishment:* Unless *White* goes to the police and confesses, *Black* will be convicted and punished for some crime that White committed. Though White knows this fact, he does nothing.

Suppose that White acts on the maxim 'Let others be punished for my crimes'. To apply Kant's Law of Nature Formula, we ask whether White could rationally will it to be true that everyone acts on this maxim. In answering this question, for the reasons that I gave above, we cannot appeal to our belief that White's act would be wrong. Nor can we appeal to the *deontic* reason that the wrongness of this act might provide. If we appeal only to other, non-deontic reasons, we may have to admit that White could rationally will it to be true that everyone acts on his maxim.

We can suppose that, if White lets Black be punished for White's crime, White would avoid many years in prison. If everyone else acted on White's maxim when it applied to them, that would increase the risk that White would later be punished for someone else's crime. But this extra risk would be small, and would be clearly outweighed by the certain benefit to White of avoiding these many years in prison. Kant's formula therefore permits White to let Black be punished for White's crime, though this act is clearly wrong. Nor does Kant's Moral Belief Formula condemn this act, since White could rationally will it to be true that everyone believes such acts to be morally permitted.

For another example, consider

> *Murderous Theft*: While traveling across some desert, *Grey* and *Blue* have both been bitten by some snake. Blue has prudently brought some drug that is an antidote to this snake's lethal poison. Grey cannot save his life except by stealing Blue's drug, with the foreseen result that Blue dies.

Grey knows, we can assume, that no one else would discover that he stole Blue's drug, nor would his life be ruined by remorse. Since Grey is young, he can expect that his act would give him many more years of life worth living. Blue can also expect such a life, and is much younger. On these assumptions, all plausible moral views imply that it would be wrong for Grey to save his life by stealing Blue's drug.

Suppose first that, if Grey stole this drug, he would be acting on the maxim 'Steal when that is my only way to save my life'. Grey could rationally will it to be true that everyone acts on this maxim, whenever it applies to them. It is unlikely that, in such a world, anyone else would treat Grey in this way; and this risk would be clearly outweighed by the certain benefit to Grey if he saves his life. On these assumptions, this case also illustrates the Rarity Objection, since Kant's formulas would permit Grey's murderous theft.

Suppose instead that, in stealing Blue's drug, Grey would be acting on the Egoistic maxim

> E: Do whatever would be best for me.

Could Grey rationally will it to be true that everyone rather than no one acts on this maxim? That depends on the alternative. As I have said, we could not rationally will it to be true that everyone acts on some maxim if there is some other, significantly better maxim on which everyone could act. One such maxim might be

> E2: Do whatever would be best for me, except when such acts would impose much greater burdens on others.

If everyone always acted on E rather than E2, that would be much worse for most

people. That is why, as I have claimed, the Egoistic maxim usually fails Kant's test. Most egoists could not rationally choose to live in a world of egoists.

Grey, however, is one of the exceptions. Grey knows that, if everyone acted on E rather than E2, he would often bear burdens that would be imposed on him by the egoistic acts of others. But we can plausibly suppose that, even in such a world, the rest of Grey's life would be worth living. If that is so, Grey could rationally will it to be true that everyone acts on E rather than E2. If everyone acted on E2, Grey would not steal Blue's drug, and would die. If we ignore deontic reasons, we must agree that Grey has sufficient reasons to prefer, not the partly moral world in which he would die, but the egoistic world in which, by stealing Blue's drug, Grey would save his own life. So Kant's Law of Nature Formula mistakenly permits Grey's murderous theft. For similar reasons, so does Kant's Moral Belief Formula.

These claims illustrate a different objection to Kant's formulas. These formulas fail here, not because few other people could act on Grey's egoistic maxim, but because Grey's wrong act gives him a benefit that is unusually great. We can call this the *High Stakes Objection*.

There are some ways in which we might try to answer this objection. For example, we might repeat Rawls's claim that, in asking whether we could rationally choose to live in a world in which everyone acts on some maxim, we should suppose that this maxim has already been acted on for a long enough time for such acts to have had their full effects. We might then argue that Grey could not rationally choose the world in which everyone always acted on the Egoistic maxim, since there is a risk that, in this world, Grey would already be dead, having been earlier killed by some other egoist. This somewhat puzzling argument would not, however, be enough to defend Kant's Law of Nature Formula. We are comparing this formula with three other principles: Kant's Consent Principle, the Impartial Observer Formula, and the Golden Rule. And when applied to the kinds of case that we are now considering, these three other principles clearly do much better.

The chief difference is this. Since Blue is much younger than Grey, Blue's death would be, for her, a much greater loss. In applying these other principles, we take into account Blue's much greater loss. Blue would not have sufficient reasons to consent to Grey's stealing Blue's drug and thereby causing Blue's death. Any rational impartial observer, given the choice, would choose that Grey does not treat Blue in this way. And Grey could not rationally choose that he be treated in this way, if he were going to be, not only in his own position, but also in Blue's. Because these three principles make our moral reasoning impartial, they all rightly condemn Grey's murderous theft.

When we apply Kant's Law of Nature Formula, in contrast, we ignore Blue's well-being, since we think about this case only from Grey's point of view. We ask

whether Grey could rationally will it to be true that he saves his life, and lives in a world of egoists. For Kant's formula to condemn Grey's act, the answer must be No. We must claim that Grey could not rationally choose the world in which he saves his life, because he has decisive non-deontic reasons to prefer the world in which he dies. Compared with the claims to which we can appeal when we apply our other three principles, this claim is much harder to defend.

48 The Non-Reversibility Objection

There is another, similar, but practically more important objection to Kant's formulas. The Golden Rule makes us more impartial by requiring us to treat everyone as we would be willing to be treated if we were going to be in the positions of all these people, and would be relevantly like them. Kant's Law of Nature Formula makes us more impartial in a less direct way. When we apply this formula, rather than asking 'What if that was done to me?' we ask 'What if everyone did that?'

This question has some value. When we act wrongly, as Kant points out, we often make unfair exceptions for ourselves, doing things that we would not want or will other people to do. [405] Kant's Law of Nature Formula rightly condemns such acts. And as I have claimed, this formula is especially helpful when we are considering each-we dilemmas.

Kant's question is not, however, enough. In many cases, if we act wrongly, we would benefit ourselves in ways that would impose much greater burdens on others. The Golden Rule condemns such acts, since we would not be willing to have other people do such things to us. But when we apply Kant's formula to our acting on some maxim, we don't ask whether we could rationally will it to be true that *other* people do these things to *us*. We ask whether we could rationally will it to be true that *everyone* does these things to *others*. And we may know that, even if everyone did these things to others, *no one* would do these things to *us*. When that is true, we *could* rationally will it to be true that everyone acts like us, since we would then get the benefits from our own wrong acts, and the similar wrong acts of others would never impose the greater burdens on us. Kant's formula mistakenly permits such acts. In the simplest cases of this kind, our wrong acts are *not reversible*, since we are doing to others what they could not possibly do to us. So we can call this the *Non-Reversibility Objection*.

Unlike the Rarity and High Stakes Objections, this objection applies to many actual cases. Return first to our white racist. This man cannot claim to be following the Golden Rule. But he might claim to be following Kant's formulas. He might say:

> When I exclude blacks from my hotel, I could rationally will it to be true that everyone acts in this way. Everyone *does* act in this way. Every hotel owner excludes blacks. And I could rationally will it to be true that everyone believes such acts to be right. If the blacks believed that my acts are right, that would be fine with me.

If this man made these claims, would he have misunderstood Kant's formulas? I am not asking whether he would have misunderstood Kant's moral theory. Kant was in some ways remarkably egalitarian, and there is much in Kant's views that would condemn such racist attitudes and acts. [406] My question is only what is implied by Kant's Law of Nature and Moral Belief Formulas.

When Kant illustrates his formulas, he considers maxims on which most people do not act, and on which, he assumes, no one would want everyone to act. When he imagines some wrong-doer asking 'Could I will that my maxim be a universal law?', Kant assumes that this person's maxim *isn't* such a law. [407] But in some cases, like that of this white racist, this assumption fails. This man's maxim is already a universal law. When this man acts on the maxim 'Exclude blacks from my hotel', he is doing what, in his social world, all hotel owners do.

When wrong-doers act on such maxims, it may not help to ask 'What if everyone did that?' Kant's Law of Nature Formula permits such people's acts if they could rationally will it to be true that they and others continue to act as they are now doing. If it is bad for these wrong-doers that they and others are acting in some way--as might be true, for example, in some state of anarchy, or a war of all against all---these people could not rationally will the continuation of the existing state of affairs, or *status quo*. Kant's formula would then rightly condemn these people's acts. In many cases, however, the *status quo* is good for the people who are acting wrongly. And this state of affairs may be good for these people partly *because* their bad maxim is universal, or widely acted upon. Those to whom some maxim applies may be some powerful and privileged group, who are acting in ways that preserve their advantages over other people. Kant's Law of Nature Formula permits such people's acts if they could rationally will it to be true that they keep their privileged positions.

As before, in trying to argue that these people could *not* rationally choose to keep their privileged positions, we should not appeal to the wrongness of these people's acts, since Kant's formula would then achieve nothing. Nor could we usefully claim that these people are rationally required to give great weight to everyone else's well-being. Kant, rightly, does not appeal to such claims. For Kant's formula to support the view that these people's acts are wrong, we must be able to claim that, for other reasons, these people could not rationally will it to be true that they keep their advantages over other people. At least in the case of many of these people, we could not plausibly defend this claim.

Nor would it help to turn to Kant's Moral Belief Formula. Just as these people could rationally will it to be true that everyone in their position acts like them, they could rationally will it to be true that everyone believes such acts to be morally permitted. These people would have no relevant reason to prefer that everyone believes their acts to be wrong.

Consider, for example, those men who benefit themselves by treating women as inferior, denying women various rights and privileges, and giving less weight to women's well-being. Such acts are wrong, Kant's formulas imply, if these men could not rationally will it to be true either that everyone acts like them, or that everyone believes such acts to be justified. These claims do not provide a good objection to these men's acts. For most of history, most people---including most women---have treated women as inferior, and believed such treatment to be justified. Since we cannot appeal to the wrongness of such treatment, we would have to admit that many men could have rationally willed that they keep their privileged position.

Turn next to slave-owners. For Kant's formulas to condemn slavery, we would have to argue that slave-owners could not have rationally willed it to be true either that they keep their slaves, or that everyone, including the slaves, believes slavery to be justified. Since we cannot appeal to the wrongness of slavery, these claims might be hard to defend. It would be much better to appeal to Kant's Consent Principle, or to the Golden Rule. Women and slaves could not rationally consent to being treated as inferior, or as mere property. Nor would men or slave-owners be willing to be treated in these ways, if they were going to be in the positions of women or slaves.

Similar claims apply to many of the ways in which powerful people benefit themselves by oppressing or exploiting those who are weak. Kant's formulas condemn these people's acts only if they could not rationally will it to be true either that they and others continue to profit in these ways, or that everyone believes such exploitation to be justified. Since we cannot appeal to the unjustifiability of such exploitation, we could not plausibly defend these claims.

For one last example, we can return to global inequality. On any plausible moral view, those who control much the greatest shares of the world's resources ought to transfer much of their wealth or income to the poorest people in the world. Most rich people transfer nothing. To argue that Kant's formulas condemn these people's acts, we would have to claim that these rich people could not rationally will it to be true either that they and others continue to give nothing to the poor, or that everyone believes that, in giving nothing, the rich are acting rightly. Since we cannot relevantly appeal to the wrongness of these people's acts, or to altruistic rational requirements, we could not plausibly defend these claims. These rich people could rationally will it to be true that they continue to act as they do, and that

everyone believes their acts to be morally justified.

When Korsgaard discusses Kant's Formula of Universal Law, she writes:

> the kind of case around which the view is framed, and which it handles best, is the temptation to make oneself an exception, selfishness, meanness, advantage-taking, and disregard for the rights of others. It is this sort of thing, not violent crimes born of despair or illness, that serves as Kant's model of immoral conduct. I do not think we can fault him on this, for this and not the other is the sort of evil that most people are tempted by in their ordinary lives. [408]

Kant's formula does not, I have argued, best handle selfishness, meanness, and advantage-taking. In both its law of nature and moral belief versions, Kant's formula fails to condemn many of the acts with which some people take advantage of others---as when men, the rich, and the powerful take advantage of women, the poor, and the weak. And since Kant presents his formula as the supreme principle of morality, we *can* fault this formula for its failure to condemn such acts. These kinds of selfishness and advantage-taking are precisely the sorts of evil that men, the rich, and the powerful are tempted by, and often commit, in their ordinary lives.

49 A Kantian Solution

It might be claimed that, in presenting these objections to Kant's Formula of Universal Law, I have misinterpreted this formula. Nagel suggests that, when we ask whether we could rationally will it to be true that everyone acts on our maxim, Kant intends us to imagine that we are going to be in everyone else's positions, and that we shall be relevantly like all these other people. [409] This suggestion makes Kant's formula like a greatly inflated version of the Golden Rule, which requires us to try to imagine that we shall be in the positions of billions of other people.

None of Kant's claims about his formula support Nagel's interpretation. [410] And there are contrary passages, such as Kant's discussion of the rich and self-reliant man who has the maxim of not helping others who are in need. When Kant claims that this man could not rationally will that his maxim be a universal law, he writes:

> many cases could occur in which. . . by such a law of nature arisen from his own will, he would rob *himself* of all hope of the assistance that he wishes for *himself*. [411]

If Kant intended this man to imagine that he was going to be in the positions of the other people who need help, he would surely say that here.

Nagel defends his interpretation with the claim that, if Kant did not intend us to

imagine that we were going to be in everyone else's positions, Kant's formula would be open to serious objections. But even the greatest philosophers can overlook objections.

Rawls proposes another interpretation of Kant's formula. When we apply this formula, Rawls suggests, Kant intends us to imagine that we know nothing about ourselves or our circumstances. We should ask what we could rationally will if we were behind a *veil of ignorance*, not knowing whether we are men or women, rich or poor, fortunate or in need of help. Like Nagel, Rawls supports this interpretation with the claim that it seems needed to defend Kant's formula from objections. [412] But even if Kant ought to have used the idea of a veil of ignorance, that doesn't show that he did. In his discussions of his Formula of Universal Law, Kant never suggests that we ought to imagine that we know nothing about ourselves or our circumstances. [413]

On a third interpretation of Kant's formula, suggested by T. C. Williams, Kant intends us to judge our maxims from the imagined point of view of an impartial observer. Williams similarly defends his interpretation with the claim that it is needed to defend Kant's formula from objections. [414] But when Kant discusses his formula, he never asks us to imagine that we are impartial observers.

Scanlon proposes a fourth interpretation. When we apply Kant's formula, Scanlon suggests, Kant intends us to ask whether *everyone* could rationally will that our maxim be a universal law. [415] But this cannot be what Kant means. Kant writes:

> I ought never to act except in such a way that *I* could also will that my maxim be a universal law. [416]

Kant gives many different statements of his formula, none of which refers to what everyone could will.

These proposals would be better made, not as claims about what Kant means, but as ways of revising Kant's formula so that it can avoid objections of the kind that we have been considering.

Of these proposed revisions, Scanlon's, I believe, is the best. According to the moral belief version of Kant's formula, or

> MB: It is wrong for us to act on some maxim unless *we ourselves* could rationally will it to be true that everyone believes that such acts are morally permitted.

On Scanlon's proposal, this would become

> MB4: It is wrong for us to act on some maxim unless *everyone* could rationally will it to be true that everyone believes that such acts are morally permitted.

This revision is also suggested by some of Kant's claims about two of his other principles, the Formulas of Autonomy and of the Realm of Ends. For example, Kant refers to

> the idea of the will of every rational being as a will giving universal law. [417]

Though Kant never appeals to what everyone could rationally will, that may be only because he assumes that this revision of his formula would make no difference. Kant may assume that what any one person could rationally will must be the same as what everyone else could rationally will. On this assumption, MB and MB4 would always coincide.

This assumption, I have claimed, is false. What could be rationally willed by many of those who are men, rich, or powerful could *not* be rationally willed by many of those who are women, poor, or weak. Since there can be such differences between what different people could rationally will, MB and MB4 sometimes conflict, and we must choose between them. If Kant had seen the need to make this choice, he would have rightly chosen MB4. [418]

Remember next that we ought to revise Kant's formula so that it applies, not to the agent's maxim, but to the morally relevant description of what this person is doing. Our revised formula can therefore become

> MB5: It is wrong to act in some way unless everyone could rationally will it to be true that everyone believes that such acts are morally permitted.

With similar revisions, Kant's Law of Nature Formula would become:

> LN5: It is wrong to act in some way unless everyone could rationally will it to be true that everyone acts in this way, in similar circumstances, whenever they can.

As I explain in a note, it is enough to appeal to MB5. [419]

When people believe that some kind of act is morally permitted, they accept some principle that permits such acts. So MB5 can become

> *the Formula of Universally Willable Principles*: An act is wrong unless such acts are permitted by some principle whose universal acceptance everyone could rationally will.

In Scanlon's words, 'to answer the question of right and wrong what we must ask is.

. . "What general principles of action could we all will?"' [420]

This formula makes our moral reasoning impartial in a way that avoids the Rarity, High Stakes, and Non-Reversibility Objections. Since this formula does not appeal to the agent's maxim, it avoids the Mixed Maxims Objection. Since this formula allows us to appeal to conditional principles, it also avoids the Threshold Objection. We need another revision to avoid the New Ideal World Objection, but that revision would raise some complications that we can here ignore.

After considering some similar objections, as I have said, some people have come to believe that Kant's Formula of Universal Law cannot help us to decide which acts are wrong. When applied to such questions, Wood calls this formula 'radically defective' and 'pretty worthless,' Herman claims that it cannot be made to work, Hill doubts that it can provide 'even a loose and partial action guide', and O'Neill claims that it often gives either unacceptable guidance or no guidance at all. [421] Since these are claims about Kant's actual formula, they are, as I have argued, justified. Whether some act is wrong does not depend on the agent's maxim, and Kant's formula cannot succeed if this formula appeals only to what the agent could rationally will. But we can revise Kant's formula by dropping Kant's appeal to the concept of a maxim in the sense that covers policies, and appealing instead to principles, and to what everyone could rationally will. All these objections then disappear.

If we appeal to the principles that everyone could rationally choose to be the principles that everyone accepts, our view is of the kind that is called *Contractualist*. Several writers, such as Rawls and Scanlon, propose what have been called *Kantian* versions of Contractualism. But the Formula of Universally Willable Principles is, I believe, the version of Contractualism that is closest to Kant's own view. So we can restate this formula, and give it a shorter name. According to

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

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

CHAPTER 15 CONTRACTUALISM

50 The Rational Agreement Formula

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

> *the Rational Agreement Formula*: Everyone ought to follow the principles to whose being universally accepted it would be rational for everyone to agree.

Some Contractualists appeal instead to the principles to whose being universally *followed*---or *successfully* acted upon---it would be rational for everyone to agree. Most of my claims would apply to such versions of Contractualism, to which I shall return. I shall say that we *choose* the principles to whose universal acceptance we agree. We choose rationally, most Contractualists assume, if our choice would be best or expectably-best for ourselves. We can start with that assumption.

Though there are some principles whose universal acceptance would be best for everyone, there are others whose acceptance would be best only for certain people. What would be best for men, for example, would not always be best for women. It may seem that, when people's interests conflict, there would be no principle whose choice would be rational for everyone in self-interested terms. But the Rational Agreement Formula applies only to principles that it would be rational for *everyone* to choose. There would be no point in our choosing principles whose acceptance would be best for ourselves, if some other people could not rationally choose these principles.

What we could rationally choose would also depend on the effects of our failing to reach agreement. Most Contractualists tell us to suppose that, if we failed to agree, no one would accept any moral principles, so that no one would believe that any acts were wrong. Such a world would be likely to be bad for everyone. In this amoral *No-Agreement World*, as Hobbes memorably wrote, our lives would be 'solitary, poor, nasty, brutish, and short'. That would give everyone strong self-interested reasons to try to reach agreement.

We can suppose that, to make this agreement easier to achieve, there would be discussions, and a series of straw votes. But there would have to be some final vote. [422] We must all know that, if we failed to reach agreement in this last round, we would have lost our last chance, since we could not try again. In earlier

rounds, it would be rational for us to vote tactically. We could declare that we intended to choose principles that favoured ourselves, and we would vote for these principles, thereby trying to persuade others to vote for these principles as well. Only in the decisive final vote would it be rational for each of us, given our need to reach agreement, to make our full concessions to others.

Morality, some Contractualists believe, is best regarded as a mutually advantageous bargain. This need not be an actual bargain. When people's interests conflict, it would be rational for everyone to agree on certain principles to resolve these conflicts. By appealing to this fact, these writers argue, we can justify these principles in the actual world, in which there has been no such agreement. We ought to treat each other as we would have rationally agreed to do.

To justify certain principles in this way, however, we must defend the claim that everyone *would* have rationally reached agreement on these principles. And this claim would be hard to defend. When David Gauthier discusses his proposed version of the Rational Agreement Formula, he tells us to 'suppose that after each party advances his initial claim, agreement is reached in a single round of concessions.' [423] But we cannot simply *suppose* that such agreement would be reached. Given our need to reach agreement, it would be rational for each of us to try to predict which principles everyone else would choose, and to choose these principles ourselves. In some cases, each of us might be able to predict what other people would choose. Suppose, for example, that we are trying to reach agreement on how some fixed set of resources would be shared between us. It might be uniquely rational for everyone to choose that everyone should get equal shares, since we could each predict that everyone else would make this choice. But when we are choosing most other moral principles, this coordination problem would have no such obvious solution. In trying to predict what other people would choose, each of us would be groping in the dark. So in the decisive final vote, there would be many conflicting principles that it would be equally rational for everyone to choose. The Rational Agreement Formula would then fail, since there would be no set of principles that everyone ought rationally to choose. [424]

If we ignore this first objection, there is another objection to this version of Contractualism. The No-Agreement World would be less bad for certain people, such as those who have greater abilities, and those who are rich in the non-legal sense that they control more resources. In a world without morality, people with such advantages would be better able to fend for themselves. As everyone would know, these people would have less need to reach this Contractualist agreement. That would give them greater bargaining power. These people could declare that, in the decisive final vote, they will choose certain principles that would allow them to keep their advantages, and would give them further benefits. Such threats might be credible, since these people would be more prepared than others to run the risk of bringing about the No-Agreement World. When certain questions were

being discussed, moreover, it might be better for some people if there was no agreement. One example is the question of how much of their resources the rich ought to give to the poor. If there was no agreement on this question, so that no one accepted any principle about what the rich ought to give, that would be much the same as everyone's believing that the rich were permitted to give nothing. That might be fine with the rich. In these and similar ways, those who had greater bargaining power might be able to use that power to make it rational for others to accept principles that favoured these powerful people.

Some writers accept this implication of the Rational Agreement Formula. That is true of *Hobbesian* Contractualists, like Gauthier, who defend only a minimal version of morality. Gauthier claims that, since morality presupposes mutual benefit, it would not be wrong for us to impose great harms on certain other people, if the existence of these people does not benefit us. On this view, for example, when Europeans founded colonies in North America, they were morally permitted to kill the native inhabitants. [425] Nor can this view directly support requirements to care for people who are congenitally handicapped. [426] Such conclusions, Gauthier concedes, conflict strongly with most people's moral beliefs. But Gauthier rejects appeals to such intuitive beliefs, or to our 'considered moral judgments', which he claims that moral theories ought to ignore. [427]

I have rejected Gauthier's claim that, when we apply the Rational Agreement Formula, it is Gauthier's minimal morality that everyone ought rationally to choose. We ought also, I believe, to reject Gauthier's conclusions. If our considered moral judgments conflict deeply with some moral view, we should reject this view. And, as Locke said of Hobbes, Gauthier's minimal morality does not admit 'a great many plain duties'. [428] Similar claims apply, I believe, to other Hobbesian theories. Hobbesian Contractualists give unsound arguments for unacceptable conclusions.

51 Rawlsian Contractualism

Though Rawls also appeals to the Rational Agreement Formula, he defends more acceptable conclusions. Most of Rawls's claims are about the *justice* of what he calls the *basic structure*, or main institutions, of those societies that are nation-states. These claims are not relevant here. My remarks will only be about Rawls's Contractualist account of morality, which he calls *rightness as fairness*. [429]

When applied to morality, I shall argue, Rawls's version of Contractualism fails. But if we removed the Contractualism from Rawls's great *Theory of Justice*, the result would be a liberal egalitarian view that is both in itself very appealing and well supported by some of Rawls's Non-Contractualist claims and arguments. [430]

In considering Rawlsian Moral Contractualism, we can start with Rawls's assumptions about rationality and reasons. Rawls accepts a desire-based subjective theory, claiming that we ought rationally to try to achieve the aims that, after fully informed and procedurally rational deliberation, we would most want to achieve. Of those who accept this theory, many believe that it coincides with Rational Egoism, which claims that we ought rationally to try to do whatever would be best for ourselves. These people mistakenly assume that, after such deliberation, each of us would always care most about our own well-being in the rest of our lives as a whole.

Rawls does not make that assumption. He considers cases in which justice requires us to act in ways that would be bad for us. Even in such cases, Rawls claims, it might be rational for us to do what justice requires. We would be acting rationally if we would be doing what, all things considered, we most wanted to do. In his words,

> If a person wants with deliberative rationality to act from the standpoint of justice above all else, it is rational for him so to act. [431]

Since Rawls's theory about reasons is desire-based, however, Rawls cannot claim that it would be rational for *everyone* to act justly. When he discusses people whose informed desires would be better fulfilled if they acted unjustly, Rawls claims that these people would not have sufficient reasons to do what justice requires. [432]

On subjective theories, as I have argued, we cannot have reasons to want anything as an end, or for its own sake. If people don't care about something, and they would not care even after fully informed and procedurally rational deliberation, we cannot claim that they have reasons to care. Rawls would accept these claims. In his words:

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently. [433]

Similarly, when Rawls discusses the view that

> something is right. . . when an ideally rational and impartial spectator would approve of it,

he writes:

> Since this definition makes no specific psychological assumptions about the impartial spectator, it yields no principles to account for his approvals. . . [434]

Rawls here assumes that we have no reasons to care about anything. If Rawls believed that we have such reasons, he would not claim that, if we knew only that

someone was *ideally rational*, we could draw no conclusions about what this person would approve. Rawls's claim would instead be that, since this person was ideally rational, he would approve what he had most reason to approve. For example, he would approve of acts that relieved suffering, or saved people's lives.

As a Contractualist, Rawls appeals to the principles that it would be rational for everyone to choose, if we were all trying to reach agreement on the principles that we would all accept. On Rawls's desire-based theory, what it would be rational for people to choose depends on what they would in fact want. Since Rawls cannot predict what people would want, he adds a motivational assumption. He tells us to suppose that, when we were choosing moral principles, everyone's main aim would be to promote their own interests. [435] On this assumption, Rawls's desire-based theory coincides with Rational Egoism. If we cared most about our own interests, it would be rational for us, according to desire-based or aim-based theories, to make the choices that we could expect to best promote these interests. Rawls's motivational assumption therefore allows him to appeal to claims about self-interested rationality. In his words,

> In choosing between principles each tries as best he can to advance his interests. [436]

Rawls revises the Rational Agreement Formula by adding a *veil of ignorance*. According to

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

In explaining why he adds this veil of ignorance, Rawls appeals to the two objections to Hobbesian Contractualism mentioned above.

First, if everyone knew particular facts about themselves and their circumstances---such as their sex, age, abilities, and the resources that they control---we could not hope to work out what it would be rational for everyone to choose. In Rawls's words, 'the bargaining problem. . . would be hopelessly complicated'. [437] There would be no principles to whose universal acceptance it would be rational for everyone to agree. Rawls's veil of ignorance solves this problem. If no one knew any of these facts about how they differed from other people, it would be rational for everyone to choose the same principles, so agreement would be guaranteed. It would be enough to ask what it would be rational for any one person to choose, since the same answer would apply to everyone.

Second, as Rawls points out, if we knew nothing about ourselves or our

circumstances, that would make us impartial. We would not know the facts that might give us greater bargaining power. Nor could anyone choose principles that were biased in their own favour. Though we would be choosing principles for self-interested reasons, our ignorance would ensure that, in choosing principles, we would give equal weight to everyone's well-being. [438]

One of Rawls's main aims, he writes, is to produce a systematic theory which provides an alternative to all forms of Utilitarianism. [439] It is surprising that, in trying to achieve this aim, Rawls proposes his version of Moral Contractualism, which appeals to a combination of self-interested rationality and impartiality. We should expect such a theory to support some view that was, or was close to being, Utilitarian. [440] As Rawls himself points out, Utilitarianism is, roughly, self-interested rationality plus impartiality. [441]

Rawls is aware of this problem. According to one version of Rawls's Formula, when we imagine that we are behind the veil of ignorance, we would assume that we had an equal chance of being in anyone's position. On that assumption, Rawls claims, it would be rational for everyone to choose the principle whose acceptance would make the average level of well-being as high as possible. [442] By choosing this *Utilitarian Average Principle*, each of us would maximize our own expectable level of well-being.

Rawls rejects what we can call this *Equal Chance Formula*. If we were behind the veil of ignorance, Rawls claims, we ought not to assume that we had an equal chance of being in anyone's position. According to Rawls's preferred version of his formula, which we can call the *No Knowledge Formula,* we would have no knowledge of the probabilities. That would make it rational for us, Rawls argues, to choose certain non-Utilitarian principles.

For Rawls's Contractualist theory to achieve his aims, he must defend his rejection of the Equal Chance Formula. When describing his veil of ignorance, Rawls writes

> there seem to be no objective grounds. . . for assuming that one has an equal chance of turning out to be anybody. [443]

This remark treats our imagined state behind the veil of ignorance as if it would be some actual state of affairs, whose nature we would have to accept. But Rawls is proposing a thought-experiment, whose details are up to him. He could tell us to *suppose* that we have an equal chance of being anyone. So Rawls must give some other objection to the Equal Chance Formula. Rawls himself points out that, since there are different Contractualist formulas, which have different implications, he must defend his choice of his particular formula. This formula, he writes, must be

the one that is 'philosophically most favoured', because it 'best expresses the conditions that are widely thought reasonable to impose on the choice of principles'. [444] Could Rawls claim that, compared with the Equal Chance Formula, his No Knowledge Formula better expresses these conditions?

The answer, I believe, is No. Rawls's veil of ignorance is intended to ensure that, in choosing principles, we would be impartial. To achieve this aim, Rawls need not tell us to suppose that we have no knowledge of the probabilities. If we supposed that we had an equal chance of being in anyone's position, that would make us just as impartial. Since there is no other difference between the Equal Chance and No Knowledge Formulas, Rawls's No Knowledge Formula cannot be claimed to be in itself more plausible.

When Rawls discusses what he calls the 'Kantian interpretation' of his theory, he suggests another defence of his No Knowledge Formula. Kantian Contractualism, Rawls writes,

> aims for the thickest possible veil of ignorance. . . The Kantian rationale. . . starts by allowing the parties no information and then adds just enough so that they can make a rational agreement. [445]

By supposing that we know as little as possible, Rawls suggests, we would make our reasoning as similar as possible to the reasoning of our noumenal selves in Kant's timeless noumenal world, and we would thereby best express our freedom and autonomy.

This defence of the No Knowledge Formula does not, I believe, succeed. If we start by supposing that, behind Rawls's veil of ignorance, we would have *no* information, and we ought then to add *just enough* information to make a rational choice possible, we ought to appeal to a more extreme version of the No Knowledge Formula. In making our choices, for example, we need not know that different people have different abilities, or that we live in a world with scarce resources. Even if we did not know such facts, we could know enough to make a rational decision. [446] We would then be closer to achieving Rawls's aim of 'the thickest possible veil of ignorance'. But this version of Contractualism could not be claimed to be the one that, in Rawls's words, 'best expresses the conditions that are widely thought reasonable to impose on the choice of principles'. We cannot reasonably require that those who are choosing moral principles be as ignorant as possible. It is *well*-informed not *ill*-informed choices to which we can more plausibly appeal. [447] Rawls also writes that, on this Kantian version of his view, 'we start from no information at all; for by negative freedom Kant means being able to act independently from the determination of alien causes'. [448] True beliefs are not well regarded as alien causes.

Remember next that, as Rawls claims, the Equal Chance Formula 'leads naturally' to the Utilitarian Average Principle. [449] Since Rawls cannot justify his rejection of Equal Chance version of Rawlsian Contractualism, Rawls's theory does not, as he intends, provide an argument against all forms of Utilitarianism. [450]

Rawls might reply that we can have another kind of reason to reject some formula, or moral theory. We can justifiably reject some formula, however plausible it seems, if this formula's implications conflict too strongly with some of our best considered and firmest moral beliefs. Since Rawls assumes that Utilitarianism conflicts with some of these beliefs, such as the belief that slavery is always wrong, Rawls might claim that we can justifiably reject the Equal Chance Formula on the ground that, in leading to the Utilitarian Average Principle, this formula has unacceptable implications.

If Rawls made this claim, however, his Contractualism would still provide no argument against Utilitarianism. Rawls would be appealing to our non-Utilitarian beliefs to justify our rejecting the Equal Chance Formula and appealing to his No Knowledge Formula. So he could not also claim that, by rejecting the Equal Chance Formula and appealing to his No Knowledge Formula, we could justify our non-Utilitarian beliefs. If we defend some argument only by appealing to certain beliefs, we cannot then defend these beliefs by appealing to this argument. That defence would be circular, by assuming what it was trying to justify.

Rawls might next retreat to the claim that, though the Equal Chance Formula supports Utilitarianism, his No Knowledge Formula supports plausible non-Utilitarian principles. If that were true, Rawls's appeal to his formula would at least show that Veil of Ignorance Contractualists do not have to accept Utilitarian conclusions.

Rawls's Formula does not, however, support plausible non-Utilitarian principles. When he applies his formula, Rawls argues that, if we had no knowledge of the probabilities, we ought rationally to assume the worst, and try to make our worst possible outcome as good as possible. We ought therefore to choose the principles whose acceptance would make the worst off people as well off as possible. Since this argument tells us to *max*imize the *min*imum level of well-being, we can call it the *Maximin Argument*.

This argument has been widely criticised. Even if it were sound, however, it would not support an acceptable non-Utilitarian moral view. Suppose first that we must decide how to use some scarce medical resources, treating various young people who all have some disease. In one of two possible outcomes,

> *Blue* would live to the age of 25, and a thousand other people would all live to 80.

In the other outcome,

Blue would live to 26, and these other people would all live to 30.

People would be relevantly worse off, we can next suppose, if their lives would be shorter. On the Maximin Argument, we ought then to choose the second of these outcomes, giving Blue her extra year of life, since that is what would be best for the person who would be worst off. That is an indefensible conclusion. Though we can plausibly give some priority to benefiting those people who would be worse off, this priority should not be absolute. It would be wrong to give Blue one more year of life, rather than giving fifty more years to each of a thousand other people--- people who, without these extra years, would all die almost as young as Blue. When applied to this and many other cases, the Maximin Argument has implications that are much too extreme.

Rawls accepts what I have just claimed. Though he applies his Maximin Argument to the basic structure of society, Rawls agrees that, when we apply this argument to other questions about distributive justice, this argument's implications are much too extreme. Utilitarian theories, Rawls claims, fail to provide an acceptable general principle of distributive justice. But as Rawls admits, his version of Contractualism also fails to provide such a principle. [451]

We can now turn to other moral questions. On Rawls's Maximin Argument, when we choose between different moral principles, we ought rationally to choose the principles whose acceptance would be best for those who would be worst off. There are many moral questions to which, even if it were sound, the Maximin Argument could not be plausibly applied. Suppose that we are comparing different principles about when we could justifiably fail to keep our promises, or tell lies, or impose risks on other people. It would be hard to decide which are the principles about such questions whose acceptance would be best for the worst off people. Nor could this be the right way to choose between such principles. Suppose that, if we all accepted one of two forms of the practice of promising, or one of two principles about imposing risks, that would give much greater benefits to most people. These facts would not be, as the Maximin Argument implies, morally irrelevant.

Even if Rawls did not appeal to this argument, there is another way in which Rawls's Formula fails to support plausible non-Utilitarian principles. Rawls's version of Contractualism forces us to ignore most non-Utilitarian considerations. According to Utilitarians, when we are choosing between acts or principles, it is enough to know the size and number of the resulting benefits and burdens. Most of us believe that there are several other morally important facts and considerations. We have such beliefs, for example, about how benefits and burdens should be distributed between different people, and about responsibility, desert, deception, coercion,

fairness, gratitude, and autonomy. When we apply Rawls's version of Contractualism, all such considerations are irrelevant, except insofar as they affect our own well-being. Though Rawlsian moral reasoning differs from Utilitarian reasoning, it differs only by subtraction. When Rawls describes how people would choose moral principles from behind his veil of ignorance, he writes that they

> decide solely on the basis of what best seems calculated to further their interests so far as they can ascertain them. [452]

Rawls merely denies these people most of the knowledge that self-interested calculations need. Since Rawls's imagined contractors choose principles for purely self-interested reasons, there is no way in which non-Utilitarian considerations could possibly enter in. [453]

When he first presents his theory, Rawls writes

> It is perfectly possible . . . that some form of the principle of utility would be adopted, and therefore that contract theory leads eventually to a deeper and more roundabout justification of Utilitarianism. [454]

He also writes

> for the contract view, which is the traditional alternative to Utilitarianism, such a conclusion would be a disaster. [455]

Rawls might be able to deny that his version of Contractualism justifies any form of Utilitarianism. But his claim would have to be that, even if his theory led to some Utilitarian conclusion, it is not plausible enough to justify this conclusion. [456]

52 Kantian Contractualism

To reach a more plausible and successful version of Contractualism, we should return to a different formula, and a different view about reasons and rationality. According to

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

Remember next that, according to

> the Rational Agreement Formula: Everyone ought to follow the principles to whose universal acceptance it would be rational for everyone to agree.

These formulas both require unanimity, since they both appeal to the principles whose universal acceptance everyone could rationally choose. But unlike the Rational Agreement Formula, the Kantian Formula does not use the idea of an *agreement*. When we apply the Agreement Formula, we imagine that we are all trying to reach agreement on which principles everyone would accept. Such agreement would be needed, since everyone would accept only the principles that, in this single thought-experiment, *everyone* chose. According to the Kantian Formula, in contrast,

> Everyone ought to follow the principles that everyone could rationally choose, if each person supposed that everyone would accept the principles that *he* or *she herself* chose.

In applying this formula, we carry out *many* thought-experiments, one for each person. In making these separate choices, none of us would need to reach agreement with other people, since each of us would have this power to choose which principles everyone would accept. The Kantian Formula requires unanimity in a quite different way. This formula appeals to the principles that, in these many separate thought-experiments, everyone would have sufficient reasons to choose.

Though Rawls rightly rejects the Rational Agreement Formula, the Kantian Formula is, I believe, more plausible than Rawls's Formula, and better achieves Rawls's aims.

Rawls's veil of ignorance is in part intended to eliminate inequalities in bargaining power. The Kantian Formula achieves this aim in a better way. Since there is no need to reach agreement, there is no scope for bargaining, so no one would have greater bargaining power. When we ask which principles everyone could rationally choose, we can therefore suppose that everyone knows all of the relevant, reason-giving facts, and could therefore respond to these reasons.

Consider next one of Rawls's reasons for rejecting Utilitarianism. Utilitarians believe that it would be right to impose great burdens on a few people, whenever such acts would give a greater sum of benefits to others. In such cases, Rawls claims, justice

> does not allow that the sacrifices imposed on the few are outweighed by the larger sum of advantages enjoyed by the many. [457]

According to several writers, Utilitarians reach such unacceptable conclusions because they merely add together different people's benefits and burdens. In Nagel's phrase, different people's claims are all 'thrown into the hopper', and merged into an impersonal sum. Some of these writers suggest that, to protect people from having such great burdens imposed on them, we should appeal instead to the idea of a unanimous agreement. On this proposal, by requiring such an

agreement, we give everyone a *veto* against being made to bear such burdens, thereby achieving what we can call the *anti-Utilitarian protective aim*.

Vetos, however, can be misused. Precisely *by* requiring such unanimous agreement, the Rational Agreement Formula makes it harder to achieve this protective aim. This formula gives further advantages, not to those who *most* need morality's protection, but to those who *least* need such protection, because their greater abilities, or their control of more resources, gives them greater bargaining power.

Rawls's Formula does little to achieve this protective aim. Though Rawls's veil of ignorance eliminates bargaining power, it also prevents anyone from knowing whether they are one of the few people on whom some Utilitarian principle would require or permit us to impose great burdens. Rawls appeals to the principles whose choice would be rational in self-interested terms. And as I have claimed, Rawls has no relevant objection to the Equal Chance Formula. So he cannot plausibly deny that, from behind the veil of ignorance, we could rationally choose some Utilitarian principle, or some similar but somewhat more cautious principle, running the small risks of bearing some great burden for the sake of much more likely benefits. [458]

The Kantian Formula requires unanimity without appealing either to a veil of ignorance or to a need to reach agreement. Partly for this reason, this formula better achieves the protective aim. If Utilitarians appealed to this formula, they would have to claim that we could rationally choose their principle even if we knew that we were one of the few people on whom these great burdens would be imposed. In at least some cases, we could plausibly reject this claim.

The Kantian Formula has other advantages. Though Rawls's veil of ignorance ensures impartiality, it does that crudely, like frontal lobotomy. The disagreements between different people are not resolved, but suppressed. Since no one knows anything about themselves or their circumstances, unanimity is guaranteed. In the thought-experiments to which the Kantian Formula appeals, there is no veil of ignorance. Everyone would know how their interests conflict with the interests of others. Since unanimity is not guaranteed, it would be morally more significant if unanimity *could* be achieved, because there are some principles that, even with full information, everyone could rationally choose.

Whether there are such principles depends on what we ought to believe about reasons and rationality. If the best theory were either Rational Egoism, or some desire-based or aim-based subjective theory, the Kantian Formula would not succeed. In the thought-experiments to which this formula appeals, there would be no set of principles whose choice would be rational for everyone in self-interested terms. Nor would there be some set of principles whose universal acceptance

would best fulfil everyone's fully informed desires or aims.

We ought, I believe, to reject all subjective theories. And though Rational Egoism is, in being objective and value-based, a theory of the right kind, this theory is too narrow. According to objective theories of the kind that I believe to be the truest or best, we have strong reasons to care about our own well-being, and in a temporally neutral way. But our own well-being is not, as Rational Egoists claim, the one supremely rational ultimate aim. We could rationally care as much about some other things, such as the well-being of others.

Return next to the fact that, since Rawls appeals to the principles that it would be rational to choose for self-interested reasons, there is no way in which, when we apply the Rawlsian Formula, non-Utilitarian considerations can enter in. When we apply the Kantian Formula, we can appeal to every kind of non-deontic reason, so this formula can support non-Utilitarian principles. [459]

For the Kantian Formula to succeed, what we can call its *uniqueness condition* must be sufficiently often met. It must be true that, at least in most cases, there is some relevant principle, and only one such principle, that everyone could rationally choose. If there was no such principle, there would be no principle that the Kantian Formula would require us to follow. This formula might then fail, by failing to disallow acts that are clearly wrong. If everyone could rationally choose two or more seriously conflicting principles, this formula might again fail, in similar though more complicated ways. It would not matter, though, if everyone could rationally choose any of several similar principles. Such principles would be different versions of some more general, higher-level principle, and the choice between these lower-level principles could then be made in some other way. [460] The uniqueness condition would, I believe, be sufficiently often met.

To illustrate the Kantian Formula, we can apply it to an easy question. Suppose that

> some quantity of unowned goods can be shared between different people,
>
> no one has any special claim to these goods, such as a claim based on their having greater needs, or their being worse off than others,

and

> if these goods were equally distributed, that would produce the greatest sum of benefits.

It is clear that, in such cases, everyone should be given equal shares.

Kantians might argue:

> (A) Everyone could rationally choose the principle that, in such cases, gives everyone equal shares.
>
> (B) No one could rationally choose any principle that gave them and the other people in some group less than equal shares.
>
> (C) Only the principle of equal shares gives no one less than equal shares.

Therefore

> (D) This is the only principle that everyone could rationally choose.

If we accept Rational Egoism, we must reject this argument's first premise. On this theory, everyone ought rationally to choose some principle that gave to themselves more than equal shares. We must also reject (A) if we accept a subjective theory about reasons. There are many people whose fully informed desires or aims would not be best fulfilled by their choosing the principle of equal shares. But I believe that, as (A) claims, everyone could rationally choose this principle, since we would all have sufficient reasons to make this choice. We would not be rationally required to choose some principle that gave us *more* than equal shares. As (B) claims, no one could rationally choose any principle that gave them and the other people in some group *less* than equal shares, thereby producing a smaller sum of unequally distributed benefits. As (C) claims, only the principle of equal shares gives no one less than equal shares. So, as this argument shows, this is the only principle that everyone could rationally choose. The Kantian Formula rightly implies that, in such cases, everyone should be given equal shares.

53 Scanlonian Contractualism

We can now introduce another version of Contractualism. According to

> *Scanlon's Formula*: Everyone ought to follow the principles that no one could reasonably reject. [461]

In a fuller statement:

> Some act is wrong just when such acts are disallowed by some principle that no one could reasonably reject, or when any principle permitting such acts could be reasonably rejected by at least one person.

Though 'reasonable' sometimes means the same as 'rational', Scanlon's Formula uses this word in a different, partly moral sense. We are unreasonable in this sense if

we give too little weight to other people's well-being or moral claims. [462]

Some people claim that, because Scanlon appeals to this partly moral sense of 'reasonable', his formula is empty. If we accepted Scanlon's Formula, these people say, that would make no difference to our moral thinking, since everyone could claim that the moral principles which they accept could not be reasonably rejected.

This objection overlooks the fact that, when we apply some Contractualist formula, we cannot appeal to our beliefs about which acts are wrong. [463] Suppose again that, in

> *Means*, Grey and Blue are trapped in slowly collapsing wreckage. Grey is in no danger. I could save Blue's life, but only by using Grey's body as a shield, without her consent, in some way that would destroy Grey's leg.

Many people would believe that it would be wrong for me to save Blue's life in this way. If we accept this view, we might appeal to

> *the Harmful Means Principle*: It is wrong to impose such a serious injury on someone as a means of benefiting other people.

According to another, conflicting view, which we can call

> *the Greater Burden Principle*: We are permitted to impose a burden on someone if that is the only way in which someone else can be saved from some much greater burden.

Scanlon makes various claims about what would be reasonable grounds for rejecting moral principles. According to one such claim,

> it would be unreasonable. . . to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others. [464]

We impose a burden on someone, in Scanlon's intended sense, when we fail to give this person some benefit. Blue could argue that, as Scanlon's claim implies, Grey could not reasonably reject the Greater Burden Principle. Though my acting on this principle would impose a burden on Grey, my acting on the Harmful Means Principle would impose a much greater burden on Blue. Losing a leg is a much smaller burden than failing to have our life saved.

Grey might reply that, in her opinion, Blue could not reasonably reject the Harmful Means Principle. But why would this rejection be unreasonable? Grey might say that she has a right not to be seriously injured without her consent as a means of benefiting someone else. But in claiming that she has this right, Grey would be

implicitly appealing to her belief that it would be wrong for me to injure her in this way. When we apply Scanlon's Formula, we cannot appeal to such *deontic* beliefs. Grey might claim that

> (1) my act would be wrong, because no one could reasonably reject the Harmful Means Principle, which disallows such acts.

But Grey could not defend (1) with the claim that

> (2) no one could reasonably reject this principle because such acts are wrong.

As I have said, if we combined such claims, we would be going round in a circle, getting nowhere. Grey must argue in some other way that no one could reasonably reject the Harmful Means Principle. [465]

As this example shows, Scanlonian Contractualism is far from being empty. When Blue rejects the Harmful Means Principle, Blue can appeal to the fact that, compared with losing a leg, dying is a much greater burden. This is one of the kinds of fact that, on Scanlon's view, can provide reasonable grounds for rejecting some moral principle. When Grey defends the Harmful Means Principle, she cannot appeal to any such fact. Grey's problem is that, unlike the Greater Burden Principle, the Harmful Means Principle is best defended by appealing to our intuitive beliefs about which acts are wrong. Many of us would believe it to be wrong to inflict a serious injury on someone, without this person's consent, even when that is our only way to save someone else's life. But when we apply Contractualist formulas, we cannot appeal to such beliefs.

Like Rawls, Scanlon proposes his Contractualism partly as a way of avoiding Act Utilitarianism, or *AU*. [466] In one way, as we have just seen, Contractualism makes AU easier to defend. Most of us reject AU because this view requires or permits many acts that seem to us to be wrong. As Scanlon writes,

> the implications of Act Utilitarianism are wildly at variance with firmly held moral convictions. [467]

But when we apply some Contractualist formula, and follow the Deontic Beliefs Restriction, we cannot appeal to such convictions.

Even without appealing to such convictions, however, Scanlonian Contractualists can reject Act Utilitarianism. To illustrate Scanlon's Formula, it is worth considering some examples. Suppose that, in

> *Transplant,* I am in hospital, to have some minor operation. You are my

> doctor. You know that, if you secretly killed me, my transplanted organs would be used to save the lives of five other people. [468]

According to

> AU: We ought always to do, or try to do, whatever would benefit people most.

This principle requires you to save these five people by killing me, since that is how you would benefit people most. Most of us would believe this act to be wrong.

We can plausibly defend this belief by appealing to one version of Scanlon's Formula. Suppose we all knew that, whenever we were in hospital, our doctors might secretly kill us so that our organs could be used to save other people's lives. Even if that risk would be very small, this knowledge would make many of us anxious, and would worsen our relation with our doctors. [469] This relation is of great importance, since we often rely on what our doctors decide, or advise us to do, and they may be people whom we expect to help us through the ending of our lives. By appealing to such facts, we could reasonably reject AU. If all doctors followed this principle in such cases, a few more people's lives would be saved. But the saving of these extra lives would be outweighed by these ways in which it would be bad for us and others if, as we all knew, our doctors believed that it could be right to kill us secretly in this way. We can call this the *anxiety and mistrust argument*. [470]

This argument illustrates another way in which, if we appeal to a Contractualist formula, this makes a difference to our moral reasoning. If we consider *Transplant* on its own, we could ignore this argument. Since you could save the five by secretly killing me, your act would produce no anxiety or mistrust. But when we apply some Contractualist formula, such as the Kantian or Scanlonian Formulas, we don't consider particular acts on their own. We ask which are the principles that everyone could rationally choose, or that no one could reasonably reject, if we were choosing the principles that everyone would accept. In answering *this* question, we must take into account the effects of everyone's accepting, and being known to accept, these principles. That makes it irrelevant that, in *Transplant*, your act would be secret, and would therefore produce no anxiety or mistrust.

We can reasonably reject some principle, Scanlon claims, only if we can propose some better alternative. If we reject AU, what alternative should we propose?

It may help to compare *Transplant* with two other cases. Remember that, in

> *Tunnel*, by switching the points on some track, you could redirect a driverless, runaway train, so that it kills me rather than five other people,

and that in

> *Bridge*, you could save the five only by using remote control to make me fall in front of the train, thereby killing me, but also triggering the train's automatic brake.

For one alternative to AU, we might return to

> *the Harmful Means Principle*: It is wrong to impose a great injury on one person as a means of benefiting other people.

What is morally important, on this view, is how your saving of the five would be causally related to the act with which you kill me. It would be wrong for you to save the five in both *Transplant* and *Bridge by* killing me. But it would not be wrong for you to kill me in *Tunnel*, since you would here be killing me, not as a means of saving the five, but only as the foreseen side-effect of redirecting the train. Many of us would accept these claims, believing my act to be wrong in *Bridge* but permissible in *Tunnel*. When we apply Scanlon's Formula, can we plausibly defend this distinction?

The answer, I suggest, is No. When we consider cases like *Tunnel* and *Bridge*, we have strong reasons to care whether we would live or die, but no strong reasons to care how our death might be causally related to the saving of other people's lives. In making this claim, I am not assuming that only outcomes matter. We can have reasons to care about how some outcomes are produced. But when someone else could act in some way that would both kill us but also save several other people's lives, we would have no strong reason to prefer to be killed as a side-effect of the saving of these people's lives rather than as a means. Given these facts, Scanlon's Formula seems to count against the view that there is an important moral difference between your acts in *Tunnel* and *Bridge*. If I could *not* reasonably reject some principle that would permit you to kill me in *Tunnel*, it seems doubtful that I could reasonably reject every principle that would permit you to kill me in *Bridge*. Scanlon's Formula seems to imply that these acts are either both wrong, or both morally permitted.

Consider next another alternative to AU, which is suggested by the anxiety and mistrust argument. According to what we can call

> *the Emergency Principle*: Doctors must never kill their patients as a means of saving more lives. In certain *non-medical emergencies*, however, everyone is permitted to do whatever would save the most lives.

These non-medical emergencies are cases that involve unintended and immediate threats to people's lives, such as some fire, flood, avalanche, or driverless run-away train. [471] The Emergency Principle condemns your saving the five by killing me in *Transplant*, since you are here my doctor. But this principle permits you to save the

five in a way that kills me, in both *Tunnel* and *Bridge*, because these are non-medical emergencies, and in these cases I would be a stranger to you.

Compared with the Harmful Means Principle, Scanlon's Formula seems more strongly to support the Emergency Principle. What is morally important, this principle assumes, is not the *causal* relation between your saving of the five and your killing of me, but the *personal* relation between you and me in *Transplant*, and the other differences between medical and non-medical emergencies. These are the kinds of fact to which, when applying Scanlon's Formula, we can more plausibly appeal. We have reasons to want our doctors to believe that they must never kill their patients as a means of saving other people's lives---or, we can add, even as a side-effect. While our relation to our doctors is of great importance, we have no such personal relation to those who might kill us or save our lives in these rare non-medical emergencies. And we have reasons to want such people to believe that, in such cases, they ought to save as many lives as possible. We would know that, if our lives were threatened in such an emergency, we would be more likely to be one of the people whose lives would be saved.

54 The Deontic Beliefs Restriction

Suppose that, after thinking hard about these imagined cases, we believe that you would be morally permitted to kill me, in *Tunnel*, as a foreseen side-effect of saving the five, but that it would be wrong for you, in *Bridge*, to kill me as a means. We may then accept the Harmful Means Principle, which draws this distinction. Suppose next that, for the reasons I have just given, we cannot successfully defend this principle by appealing to Scanlon's Formula. This and other similar principles are best defended by appealing to our intuitive beliefs about which acts are wrong. But when we apply Contractualist formulas, we cannot appeal to these beliefs. Nor can we appeal to these beliefs when we apply Kant's Formula of Universal Law.

We might now challenge this Deontic Beliefs Restriction. When we try to answer moral questions by applying these Kantian or Contractualist formulas, why should we ignore our beliefs about which acts are wrong?

Kantians and Contractualists might reply that, if we appealed to such deontic beliefs, their formulas would be circular, in a way that made them useless. As I have said, there is no point in claiming both that

> acts are wrong when any principle permitting them would fail some Kantian or Contractualist test,

and that

> principles would fail this test when and because the acts they permit are wrong.

This reply is not, however, enough. Even if these formulas would be useless unless we follow the Deontic Beliefs Restriction, that does not show that we ought to think about morality by applying these formulas.

Another reply appeals to a distinction that is *meta-ethical*, in the sense that it makes claims about the nature and justifiability of moral beliefs and claims. According to *Intuitionists*, Rawls writes, there are certain independent truths about which acts are wrong, and about which facts give us reasons.[472] Two examples are the truths that slavery is wrong, and that we have reasons to prevent or relieve suffering. These truths are *independent* in the sense that they are not created or constructed by us. According to a different view, which Rawls calls *Constructivism*, there are no such truths. [473] On this view, what is right or wrong depends entirely on which principles it would be rational for us to choose in some Kantian or Contractualist thought-experiment. In Rawls's phrase, it's for us to decide what the moral facts are to be. [474] If we are Constructivist Contractualists, and we believe that it would be rational to choose principles that permit slavery, we ought to conclude that slavery is not wrong. Though slavery may seem to us to be wrong, Constructivists reject appeals to our moral intuitions, which some of them claim to involve prejudice, or cultural conditioning, or to be mere illusions.

I shall here assume that we ought to reject these *sceptical*, anti-intuitionist views. Rawls does not commit himself to Constructivism, and he often assumes that there are some independent moral truths, such as the truth that slavery is wrong. When we try to achieve what Rawls calls reflective equilibrium, we should appeal to all of our beliefs, including our intuitive beliefs about the wrongness of some kinds of act. As Scanlon writes:

> this method, properly understood, is. . . the best way of making up one's mind about moral matters. . . Indeed, it is the only defensible method: apparent alternatives to it are illusory. [475]

If Kantians and Contractualists accept that our moral reasoning should appeal to such intuitive beliefs, they must defend the Deontic Beliefs Restriction in some other way.

There is one straightforward and wholly satisfactory defence. In describing this defence, we can first distinguish between two senses in which some property of an act, or some fact about this act, might make this act wrong. When some property of an act makes this act wrong, it does not *cause* it to be wrong. In one trivial sense,

wrongness is the property that *non-causally* makes acts wrong. That is like the sense in which blueness is the property that makes things blue, and illegality is the property that makes acts illegal. It is in a different and highly important sense that when acts have certain other properties---such as that of causing pointless suffering, or being a lying promise---these facts may non-causally make these acts wrong. Causing pointless suffering isn't the same as being wrong. But if some act causes pointless suffering, this fact may make this act wrong by making it have the different property of being wrong. Moral theories should try to describe the properties or facts that, in this sense, can make acts wrong. [476]

Scanlon once claimed that his Contractualism gives an account, not of what *makes* acts wrong, but of wrongness itself, or of *what it is* for acts to be wrong. This claim was, I believe, a mistake. To see why, we can first restate the Kantian Contractualist Formula. According to

> KF2: An act is wrong just when such acts are disallowed by one of the principles whose universal acceptance everyone could rationally will.

Suppose next that, in

> the *Kantian* sense, 'wrong' means 'disallowed by the principles whose universal acceptance everyone could rationally will'.

If Kantian Contractualists used 'wrong' in this sense, they could claim to be giving an account of one kind of wrongness. On this view, when acts are disallowed by such a principle, that's *what it is* for these acts to be wrong in this Kantian sense. But KF2 would then be a concealed tautology, one of whose open forms would be

> KF3: An act is disallowed by such a principle just when such acts are disallowed by such a principle.

And this claim is not worth making. Kantian Contractualists ought instead to use 'wrong' in one or more non-Kantian senses. KF2 would not then be trivial, since this claim would mean that, when some act is disallowed by such a principle, that makes this act wrong in such other senses. For example, Kantian Contractualists might claim

> KF4: When some act is disallowed by one of the principles whose universal acceptance everyone could rationally will, that makes this act wrong in the senses of being unjustifiable to others, blameworthy, and an act that gives its agent reasons to feel remorse and gives others reasons for indignation.

If we are Kantian Contractualists, we should not claim that our formula describes the *only* property or fact that makes acts wrong in these other senses. There are other wrong-making properties or facts that would often have more importance. Our

claim should instead be that this formula describes a *higher-level* wrong-making property or fact, under which all other such properties or facts can be subsumed, or gathered. When some act is a lying promise, for example, this fact may make this an act that is disallowed by one of the principles whose universal acceptance everyone could rationally will. According to this version of Kantian Contractualism, both of these facts could then be truly claimed to make this act wrong.

Scanlon's theory should, I believe, take the same form. According to

> Scanlon's Formula: An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject.

If Scanlon was here using 'wrong' in another Contractualist sense, to mean 'disallowed by such an unrejectable principle', he could claim that his formula gives an account of this Contractualist kind of wrongness, or of *what it is* for acts to be wrong in this sense. But his formula would then be another concealed tautology, one of whose open forms would be the claim that acts are disallowed by such unrejectable principles just when these acts are disallowed by such principles. We could all accept that claim, whatever our moral beliefs. Scanlon's claim should instead be that, if some act is disallowed by some principle that could not be reasonably rejected, that makes this act wrong in one or more non-Contractualist senses.

Scanlon now accepts that his view should take this form. [477] We can therefore say that, on Scanlon's theory, when acts have certain other properties, that makes these acts disallowed by some unrejectable principle, and these facts can all be truly claimed to make these acts in other senses wrong.

If Contractualists make such claims, they can defend the Deontic Beliefs Restriction without rejecting our moral intuitions as worthless. On these versions of Contractualism, it is only *while* we are asking what Contractualist formulas imply that we should not appeal to our beliefs about the wrongness of the acts that we are considering. We can appeal to these beliefs at a later stage, when we are deciding whether we ought to accept these formulas. As when considering any other claim about which acts are wrong, we could justifiably reject any Contractualist formula if this formula's implications conflict too often and too strongly with our intuitive moral beliefs. [478]

On this version of Scanlon's view, he does not reject appeals to our intuitive beliefs. Scanlon shows that, as well as having such beliefs about which acts are wrong, we have and can usefully appeal to intuitive beliefs about what are reasonable grounds for rejecting moral principles. That is Scanlon's greatest contribution to our moral thinking.

CHAPTER 16 CONSEQUENTIALISM

55 Consequentialist Theories

Before we ask what is implied by Kantian Contractualism, it may help to return to the relation between what is good and what is right.

Pain is bad, some of us truly believe, in the sense of being something that we have reasons to want to avoid. But some great philosophers did not have such beliefs. Hume, for example, does not use 'good' or 'bad' in reason-implying senses. This may be why Hume claims that it cannot be unreasonable, or contrary to reason, to prefer our own acknowledged lesser good to our greater good. If Hume had used 'lesser good' to mean 'what we have less reason to prefer', he could not have believed that no such preference could be unreasonable. Hume often uses 'good' and 'evil' merely to mean 'pleasure' and 'pain'. [479]

While Hume would have thought it trivial to claim that pain is evil, Kant sometimes rejects this claim. For example, Kant writes:

> good or evil is, strictly speaking, applied to actions, not to the person's state of feeling. . . Thus one may laugh at the Stoic who in the most intense pains of gout cried out, 'Pain, however you torment me, I will still never admit that you are something evil (*kakon*, *malum*)', nevertheless, he was right. [480]

When Kant claims that pain cannot be evil, he means that pain cannot be morally bad. Like Hume, Kant seems sometimes to be unaware of, or to forget, the reason-implying sense in which it is bad to be in pain. [481]

So does Ross. If some event would be bad, Ross assumes, we have a prima facie duty to prevent this event, if we can. Because we have no such duty to prevent ourselves from being in pain, Ross concludes that our own pain is not bad. More exactly, Ross, suggests, our pain *is* bad, but only from other people's point of view. [482] Ross reaches this strange conclusion because he ignores the reason-implying senses in which things can be non-morally good or bad.

As well as being bad *for* the person who is in pain, pain is also *impersonally* bad. In Nagel's words, 'suffering is a bad thing, period, and not just for the sufferer.' [483] Many people believe that, though outcomes can be good or bad for particular people, there is no sense in which outcomes could be impersonally good or bad. [484] But, as I have said, we can explain such a sense. When we are comparing different

possible outcomes, and we claim that some outcome would be

> *impersonally best* in the *impartial-reason-implying* sense, we mean that this is the outcome that, from an impartial point of view, everyone would have most reason to want, or to hope will come about.

When we consider possible events that would involve and affect only strangers, our actual point of view is impartial. But we also have impartial reasons when our point of view is not impartial, as is true, for example, when we could relieve either our own or someone else's pain. All pain is bad in the sense that we all have reasons to regret anyone's being in pain, whatever that person's relation to us. And we all have reasons to want everyone's life to go well.

If we accept some subjective theory about reasons, or Rational Egoism, we must deny that outcomes could be in this sense good or bad. On these theories, there are no outcomes that everyone has some reason to want, or to regret. It could not be in this sense bad if some plague or earthquake killed many people, since this outcome would not be bad for everyone, nor would everyone have desire-based or aim-based reasons to want such people not to be killed. But we ought, I have argued, to reject these theories.

In what follows, I shall use 'best' in the impartial-reason-implying sense. There are often two or more possible outcomes that might be called 'equal-best'. Since that phrase misleadingly suggests precision, it would be better to call such outcomes not worse than any of the others. [485] To save words, however, I shall use 'best' to refer to all such outcomes.

Though any plausible moral theory could appeal to facts about the goodness of outcomes, certain theories take such facts to be fundamental. According to what I am now calling

> *Consequentialism*: Whether our acts are right or wrong depends only on facts about how it would be best for things to go.

Consequentialist theories can differ in several ways, since they can make conflicting claims both about what is good and bad, and about how the rightness of our acts depends on facts about what would be best.

Some Consequentialists are *Utilitarians*, who believe that

> (A) things go best when they go in the way that would, on the whole, benefit people most, by giving them the greatest total sum of benefits minus burdens.

Other Consequentialists believe that the goodness of outcomes depends in part on other facts. Some people, for example, believe that

> (B) how well things go depends in part on how benefits and burdens are distributed between different people.

On two such views, one of two outcomes might be better, though it would involve a smaller sum of benefits minus burdens, because these benefits and burdens would be more equally distributed, or because more of the benefits or fewer of the burdens would go to people who were worse off.

The word 'Consequentialist' is in one way misleading, as is talk of the goodness of *outcomes* and of the acts that *make* things go best. These words suggest that, on these theories, all that matters is the future, and the effects of our acts. Consequentialists can reject those claims. The goodness of some outcomes might depend in part on facts about the past. It might be better, for example, if benefits went to people who had earlier been worse off, or if we kept our promises to those who are dead, or if people are punished only if they earlier committed some crime. And some acts, intentions, and motives may be in themselves good or bad, whatever their effects. Kind acts may be good, for example, even when they fail, and it may often be in itself bad when people are deceived or coerced. When we ask whether it would be best if something happened, or if someone acted in some way, we are asking what, from an impartial point of view, everyone would have most reason to want, or to hope. This sense of 'best' leaves it entirely open which are the ways in which we would have most reason to want things to go.

There is, however, one kind of value to which Consequentialist theories cannot appeal. Some Consequentialists believe that

> (C) when people act rightly for the right reasons, these acts are in themselves good, and wrong acts are in themselves bad.

As I explain in a note, the rightness or wrongness of our acts cannot depend on whether these acts are in these ways good or bad.[486]

All Consequentialists appeal to claims about what would make things go best. We can call this the *Consequentialist Criterion*. *Direct* Consequentialists apply this criterion directly to everything: not just to acts, but also to rules, laws, customs, desires, emotions, beliefs, the distribution of wealth, the state of the Earth's atmosphere, and anything else that might make things go better or worse. When these people apply this criterion to acts, they are *Act Consequentialists*. Some of these people claim that

> (D) everyone ought always to do whatever would in fact make things go best.

Others claim that

> (E) everyone ought always to do, or try to do, whatever would be most likely to make things go best, or more precisely what would make things go *expectably-best*. [487]

If (D) uses 'ought' in the fact-relative sense and (E) uses 'ought' in the evidence-relative or belief-relative senses, these claims do not conflict. In most of what follows we can ignore the difference between these claims. And I shall often use 'best' to mean 'best or expectably-best'.

Indirect Consequentialists apply the Consequentialist Criterion directly to some things but only *indirectly* to others. *Rule Consequentialists* apply this criterion directly to rules or principles, but only indirectly to acts. Some of these people believe that

> (F) everyone ought to follow the principles whose universal acceptance would make things go best.

On this view, though the best principles are the ones whose universal acceptance would make things go best, the best or right acts are not the acts that would make things go best, but the acts that are required or permitted by the best principles. It would be wrong to do what would make things go best when such acts are disallowed by one of the best principles. *Motive Consequentialists* similarly claim that, though the best motives are the ones whose being had by everyone would make things go best, the best or right acts are not the acts that would make things go best, but the acts that would be done by people with the best motives. These theories overlap with those systematic forms of *virtue ethics* which appeal to the character-traits and other dispositions that best promote human flourishing or well-being. There could be many other forms of Indirect Consequentialism. [488]

56 Consequentialist Maxims

Some Consequentialists might apply their criterion directly to maxims, and only indirectly to acts. Of the possible maxims on which everyone might act, some would be

> *optimific* in the sense that, if everyone acted on these maxims, things would go in the ways that would be impartially best.

According to what we can call

> *Maxim Consequentialism:* Everyone ought to act only on these optimific maxims.

It is worth returning briefly to one of Kant's formulas. Some Kantians might argue:

> (G) Each of us is permitted to act on some maxim if we could rationally will it to be true that everyone acts on this maxim.
>
> (H) Some people could rationally will it to be true that everyone acts on the optimific maxims.

Therefore

> These people are permitted to act on these maxims.

(G) is Kant's Law of Nature Formula. If (H) is true, Kant's formula permits some people to be Maxim Consequentialists, who act on these optimific maxims.

In assessing this argument, we must appeal to some view about reasons and rationality. According to wide value-based objective views of the kind that I believe we should accept, (H) is true. If everyone acted on the optimific maxims, things would go in ways that would both be impartially best and be best *for* some people. These *fortunate* people would have both impartial and personal reasons to will it to be true that everyone acts on these maxims, and at least some of these people would not have any stronger conflicting reasons.

When we apply Kant's formula, some writers claim, we ought to appeal only to a rational requirement to avoid inconsistency, or contradictions in our will. On this assumption, (H) is true. There would be some people who could rationally will it to be true that everyone acts on the optimific maxims, since that would involve no inconsistencies or contradictions in these people's wills. Other writers claim that we are rationally required to will what would best fulfil our true needs as rational agents. [489] On this assumption, there would again be some fortunate people who could rationally will it to be true that everyone acts on the optimific maxims. Things would go best in such a world in part because many people's true needs as agents would be best fulfilled.

(H) is also true on subjective theories about reasons. Of the fortunate people, some would care strongly about the well-being of others, and would want things to go in the ways that would be best. [490] Some of these people would have desires that would be best fulfilled if everyone acted on the optimific maxims.

Rational Egoists might reject (H). We are rationally required, these people believe, to choose whatever would be best for ourselves. It would be best for each person, Rational Egoists might claim, if everyone acted on certain maxims that were not optimific, because some of these acts would give this person extra benefits, in ways that imposed greater burdens on others. But this claim, I believe, is false. As before, some of the fortunate people would care strongly about the well-being of

others, and if things went in the ways that would be impartially best, that would be best for some of these people.

Similar claims apply to any other plausible or widely accepted view about reasons and rationality. On all such views, there would be some people who could rationally will it to be true that everyone acts on the optimific maxims. Kant's original Law of Nature Formula, we can therefore claim, permits some people to be Maxim Consequentialists.

It is an objection to Kant's formula that it permits only *some* people to be Maxim Consequentialists, since such moral claims ought to apply to everyone. We can call this *the Relativism Objection*. To answer this objection, we can revise Kant's formula so that it appeals, not to what the agent could rationally will, but to what everyone could rationally will. This revised formula has implications that apply to everyone.

We have other strong reasons, I have argued, to revise Kant's formulas in this and certain other ways. These revisions lead us to the Kantian Contractualist Formula. So we can now ask what this formula implies.

57 The Kantian Argument

Of the principles that everyone might accept, some might be

> *UA-optimific* in the sense that these are the principles whose universal acceptance would make things go best.

According to the *universal acceptance* version of Rule Consequentialism, or

> *UARC:* Everyone ought to follow these optimific principles.

When we consider some kinds of case, there might be two or more optimific principles that were significantly different. Rule Consequentialists would then have to choose between these principles in some other way. This question is best considered later. So we can here suppose that there is only one set of UA-optimific principles.

Kantians could argue:

> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.
>
> (B) Everyone could rationally choose whatever they would have sufficient reasons to choose.

(C) There are some UA-optimific principles.

(D) These are the principles that everyone would have the strongest impartial reasons to choose.

(E) No one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons.

Therefore

(F) Everyone would have sufficient reasons to choose these optimific principles.

(G) There are no other significantly non-optimific principles that everyone would have sufficient reasons to choose.

Therefore

(H) It is only these optimific principles that everyone would have sufficient reasons to choose, and could therefore rationally choose.

Therefore

Everyone ought to follow these principles.

This argument is valid. (A) is the Kantian Contractualist Formula. So if this argument's other premises are true, this formula requires everyone to follow these optimific principles. We can call this *the Kantian Argument for Rule Consequentialism*.

When we apply the Kantian Formula, we ask which principles each person could rationally choose, if this person supposed that he or she had the power to choose which principles would be accepted by everyone, both now and throughout the future. This formula appeals to the principles that, in these many imagined cases, everyone could rationally choose. We should assume that, in making these choices, everyone would know all of the relevant facts. On that assumption, as premise (B) claims, everyone could rationally choose what they would have sufficient reasons to choose.

We are supposing that, as (C) claims, there is some set of principles that are UA-optimific. Of all the principles that everyone might accept, these are the principles whose universal acceptance would make things go best in the impartial-reason-implying sense. If everyone accepted these principles, things would go in the ways in which everyone would have the strongest impartial reasons to want things to go. That is true by definition. So, as premise (D) claims, these are the principles whose

universal acceptance everyone would have the strongest impartial reasons to choose.[491]

According to premise (E), no one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons. This premise needs to be defended. If we were choosing principles from an impartial point of view, it is the optimific principles that everyone would have most reason to choose. But in the thought-experiments to which this Kantian Formula appeals, we would *not* be choosing principles from an impartial point of view. Our choices would affect our own lives, and the lives of those other people to whom we have close ties, such as our close relatives and those we love. So we might have strong personal and partial reasons *not* to choose the optimific principles.

To decide whether everyone could rationally choose these principles, we must know what the alternatives would be. It will be enough here to consider other principles that would be *significantly* non-optimific, in the sense that their universal acceptance would make the future history of the world go, in certain ways, *much* worse. We need not compare the optimific principles with any principles that are only *slightly* non-optimific, since their acceptance would make things go in ways that would be only slightly worse. As before, we should first try to get the main outlines right. Details can wait.

58 Self-interested Reasons

In asking whether premise (E) is true, we should consider the strongest reasons that anyone might have not to choose that everyone accepts the optimific principles. Of our reasons not to choose these principles, some might be provided by facts about our own well-being. If everyone accepted the optimific principles, that would be very bad for certain people. These people would have strong self-interested reasons not to choose these principles.

I might be such a person. Suppose again that, in

> *Lifeboat*, I am stranded on one rock, and five people are stranded on another. Before the rising tide covers both rocks, you could use a lifeboat to save either me or the five. I and the five are all strangers to you and to each other, and we are in other ways relevantly similar. We are all young, and we would all lose, in dying, many years of happy life.

Any optimific principle would require you to save the five, since it would be worse if more people died. According to one such principle, which we can call

> *the Numbers Principle*: When we could save either of two groups of people,

> who are all strangers to us and are in other ways relevantly similar, we ought to save the group that contains more people.

Suppose next that my rock is nearer to you. According to

> *the Nearness Principle*: In such cases, we ought to save the group that is nearer to us. [492]

If everyone accepted the Numbers Principle rather than the Nearness Principle, there would be many other cases in which some people would act on this principle, so many more people's lives would be saved. This fact would give me strong impartial reasons to choose that everyone accepts the Numbers Principle. But I would know that, if I made this choice, you would act on this principle by saving the five, and I would die, thereby losing many years of happy life. [493] This fact would give me strong self-interested reasons to choose the Nearness Principle, since you would then save my life. According to premise (E), these self-interested reasons would not be decisively stronger than, or outweigh, my impartial reasons to choose the Numbers Principle. Is that true?

According to subjective theories about reasons, the answer depends on my desires or aims. If I cared enough about the well-being of other people, I could rationally choose that everyone accepts the Numbers Principle. But if we are Subjectivists, we must reject the Kantian Formula. In most cases, there would be no principles that everyone would have sufficient desire-based or aim-based reasons to choose. As I have argued, however, we ought to reject Subjectivism, and accept some Objectivist view, which appeals to value-based object-given reasons.

According to one such view,

> *Rational Egoism*: We always have most reason to do whatever would be best for ourselves.

On this view, premise (E) is false. I could not rationally choose that everyone accepts the Numbers Principle, since that choice would be worse for me. But we ought, I believe, to reject this view.

According to a view at the opposite extreme,

> *Rational Impartialism*: We always have most reason to do whatever would be impartially best.

On this view, we would be rationally required to sacrifice our life if we could thereby save several strangers. If that were true, cases like *Lifeboat* would provide no objection to premise (E). I would be rationally required to choose that everyone accepts some optimific principle, such as the Numbers Principle. [494] But we ought

also, I believe, to reject this view.

According to

> *wide value-based objective views*: When one of two possible acts would make things go in some way that would be impartially better, but the other act would make things go better either for ourselves or for other people to whom we have close ties, we often have sufficient reasons to act in either way.

On such views, we are often rationally permitted but not rationally required to give significantly greater weight, or strong priority, both to our own well-being and to the well-being of those to whom we have close ties, such as our close relatives and those we love. We ought, I believe, to accept some view of this kind.

On the views that seem to me most plausible, if we could save either our own life or the lives of several strangers, we would have sufficient reasons to act in either way. In *Lifeboat*, I could rationally choose that you save me; but I could also rationally choose instead that you save the five. So I could rationally choose that everyone accepts the Numbers Principle.

According to some more egoistic objective views, we are rationally required to give strong priority to our own well-being. I would not have sufficient reasons to give up my life unless I would thereby save as many as a hundred or a thousand other people. But in the thought-experiment to which the Kantian Formula appeals, I would have the power to choose which principles everyone would accept, both now and in all future centuries. The principles I chose would be accepted by many billions of people. If I chose that everyone accepts the Numbers Principle rather than the Nearness Principle, my choice would affect how people would later act in very many other cases of this kind. Though I would die, my choice would indirectly save at least a million other people. Millions of people now die each year whose lives could have been easily saved. So even on these more egoistic views, I would have sufficient reasons to give up my life to save these very many other people.

This case is only one example. But if, as I believe, I could rationally choose this optimific principle even at the cost of my own life, similar claims apply to all of the many cases in which, because the stakes are lower, no one's choice of an optimific principle would involve so great a sacrifice of their own well-being. [495]

Suppose next that my belief is mistaken. We ought, I have claimed, to reject Rational Egoism. But there is another, more plausible view that is relevant here. On this view, we could often rationally choose to bear some significant burden when we could thereby save many other people from similar burdens. That is not true, however, when this burden would be as great as dying young, and thereby losing

many years of happy life. I could not rationally choose the Numbers Principle, because I could not rationally choose to give up my life, however many other people's lives my choice would save. We can call this view *High Stakes Egoism*.

If this view were true, *Lifeboat* would provide an objection, not only to premise (E) of the Kantian Argument for Rule Consequentialism, but also to the Kantian Contractualist Formula. Just as I could not rationally choose any principle that required you to save the five rather than me, the five could not rationally choose any principle that required you to save me rather than them. In this and other such cases, there would be no principle that everyone could rationally choose, so there would be no principle that the Kantian Formula would require us to follow. If we could save either one stranger or a million others, this formula would permit us to act in either way. That is an unacceptable conclusion.

High Stakes Egoism is, I believe, false. But it is worth describing how, if this view were true, we could respond to this objection to the Kantian Formula.

Contractualists appeal to the principles that it would be rational for everyone to choose, if we were choosing in some way that would make our choices sufficiently impartial. Rawls suggests that, to achieve such impartiality, we should appeal to the principles that it would be rational for everyone to choose from behind some *veil of ignorance*, which prevented us from knowing particular facts about ourselves or our situation. I have claimed that, when we apply the Kantian Contractualist Formula, we have no need for such a veil of ignorance. There would always be some relevant principle that, even with full knowledge, everyone could rationally choose.

We are now supposing that, in one kind of case, my claim is mistaken. In these cases, we could save the lives of either of two groups of strangers, one of which contains more people. According to High Stakes Egoism, when the people in these groups were choosing principles that apply to such cases, they would be rationally required to give absolute priority to the saving of their own lives. The Kantian Formula would here fail because these people's choices would be wholly self-interested. To avoid this objection, we could revise this formula. When we apply the Kantian Formula to such cases, we might appeal to the principles that these people could rationally choose from an impartial point of view. Or we might partly follow Rawls by adding a *local* veil of ignorance, so that these people did not know whether they were in the smaller or the larger group. On both these versions of the Kantian Formula, these people could all rationally choose some optimific principle that would require us to save the group that contained more people.

The Kantian Formula might be more sweepingly revised, by appealing to principles that would *all* be chosen either from an impartial point of view, or from behind a *global* veil of ignorance. But that would make this formula less appealing in ways

that I describe in Section 52. And there would be no need for such a revision. High Stakes Egoism applies only to cases in which, if we chose some optimific principle, this choice would impose on us some very great burden, such as dying young or having to endure prolonged agony. We could rationally choose to accept some lesser injury, such as becoming deaf, or losing a leg, when our choice would indirectly save many other people from such injuries. So we could still claim that, in nearly all cases in which people's interests conflict, there would be some principle that, even with full knowledge and from their actual partial point of view, all of these people could rationally choose.

If we ought to reject High Stakes Egoism, as I believe, the Kantian Formula does not need to be even partly revised in such a way.

59 Altruistic and Deontic Reasons

Of our reasons not to choose the optimific principles, others might be provided by facts about certain other people's well-being. Suppose that, in

> *Second Lifeboat*, you could save either your child or five strangers.

We may believe that, even if you could rationally give up your *own* life to save five strangers, you could not rationally give up your *child's* life to save these strangers, nor could you rationally choose that we all accept some optimific principle that would require this act. This case may then seem to provide an objection to premise (E).

The optimific principles would *not*, however, require you to save these five strangers rather than your child. Suppose that we all accepted and acted on some principle that required us to give no priority to saving our own children from death or lesser harms. In such a world, things would go in some ways better, since more children's lives would be saved and fewer children would be harmed. But these good effects would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would give no such priority to saving our own children from harm, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and other similar facts, the optimific principles would in many cases permit us, and in many others require us, to give strong priority to our own children's well-being.

This objection could be transferred, however, to a different kind of case. Suppose that, in

> *Third Lifeboat*, it is I who could save either your child or five other children.

> These six children are all strangers to me.

Any optimific principle would require *me* to save the other five children. And we might claim that

> (I) you could not rationally choose that everyone accepts such an optimific principle, since you would have decisive reasons to choose instead that I accept some principle that would require me to save your child.

You would have such decisive reasons, we might claim, because you would have a duty to make the choice that would save your child's life.

There are other ways in which, by appealing to our moral beliefs, we might argue that we could not rationally choose that everyone accepts certain optimific principles. We may believe that, if everyone accepted these principles, that would sometimes lead us or others to act wrongly. The wrongness of such acts, we might claim, would give us decisive reasons not to choose that everyone accepts these principles.

As I have often said, however, when we apply the Kantian Formula or any other Contractualist formula, we cannot appeal to our beliefs about which acts are wrong. If we claim that

> some act is wrong because we could not all rationally choose any principle that permits such acts,

it would be pointless also to claim that

> we could not all rationally choose any such principle because such acts are wrong.

It would be similarly pointless to claim both that

> everyone ought to follow certain principles because these are the only principles that everyone could rationally choose,

and that

> these are the only principles that everyone could rationally choose because these are the principles that everyone ought to follow.

If we combined these claims, the Kantian Formula would achieve nothing. So when we apply this formula, we must ignore our beliefs about which acts are wrong. We can appeal to these beliefs only at a later stage, after we have worked out what this formula implies, and we are asking whether, given these implications, we ought to accept this formula.

Since we cannot appeal to our beliefs about your duties to your child, could we defend (I) in some other way? We could most plausibly appeal, I believe, to your love for your child. Rather than trying to ignore your duties to your child, it will be simpler to change our example. Suppose that, in

> *Fourth Lifeboat*, I could save either someone whom you love or five other people. These six people are all strangers to me.

Any optimific principle would require me to save the other five people. It might now be claimed that

> (J) you could not rationally choose that everyone accepts some optimific principle, since you would have decisive reasons to choose that I accept some other principle which required me to save the person whom you love.

Though this claim is plausible, it is not, I believe, true.

It may seem absurd to deny that you would have decisive reasons to choose this other principle. Could Romeo or Isolde have rationally chosen to let Juliet or Tristan die? While discussing a similar example, Williams writes:

> deep attachments to other persons. . . cannot embody the impartial view, and. . . also run the risk of offending against it. . . yet unless such things exist, there will not be enough substance or convictions in a man's life to compel his allegiance to life itself. Life has to have substance if anything is to have sense, including adherence to the impartial system; but if it has substance, then it cannot grant supreme importance to the impartial system. . . [496]

I am not appealing, however, to the kind of impartial system that Williams here movingly rejects. As I have just said, the optimific principles would often either permit or require us to give strong priority to the well-being of those to whom we have close ties. And in claiming that we could rationally choose that everyone accepts these principles, I am not assuming that we are rationally required to give equal weight to everyone's well-being. I assume only that, though we are rationally permitted to give strong priority to the well-being of ourselves and certain other people, we are also rationally permitted to give great weight to the well-being of strangers.

As my claims about *Lifeboat* imply, the person whom you love could rationally choose that everyone accepts some optimific principle. Though this person would then die, this choice would indirectly save very many other people's lives. This fact would give this person sufficient reasons to make this choice.

When someone whom we love could rationally choose to bear some burden for the sake of benefits to others, this fact does not imply that *we* could rationally choose that

this person bears this burden. We might be rationally required to give to the well-being of those we love much more weight than we are rationally required to give to our own well-being. We might not have sufficient reasons to save five, or fifty, or even five hundred strangers rather than saving someone whom we love. But in *Fourth Lifeboat* you would know that, if you chose that everyone accepts some optimific principle, your choice would indirectly save the lives of a much greater number of other people. You would have sufficient reasons, I believe, to make the choice that would save these many other people. It is, I agree, absurd to imagine Romeo or Isolde choosing to let Juliet or Tristan die. If you were Romeo or Isolde, you would not *in fact* make the choice that would save these many other people. But we often know that people won't in fact do what they have sufficient reasons to do. Since you would have sufficient reasons to choose some optimific principle, *Fourth Lifeboat* does not, I believe, provide an objection to premise (E), or to the Kantian Formula.

Suppose next that my belief is mistaken. It might be claimed that, when the stakes are as high as this, we are rationally required to give absolute priority to the well-being of those we love. If that were true, there would be no principle applying to such cases that everyone could rationally choose, so there would be no principle that, according to the Kantian Formula, everyone ought to follow. This formula would not require me to save even a million strangers rather than the person whom you love. That is another unacceptable conclusion. This objection is like the one that appeals to High Stakes Egoism. As before, the Kantian Formula could be revised by adding some local veil of ignorance. But this revision is not, I believe needed.

60 The Wrong-Making Features Objection

On some value-based objective theories, there are some things that are worth doing, and some other aims that are worth achieving, in ways that do not depend, or depend only, on their contributions to anyone's well-being. Scanlon's examples are 'friendship, other valuable personal relations, and the achievement of various forms of excellence, such as in art or science.' [497] These we can call *perfectionist* aims.

On such views, it would be in itself good in the impartial-reason-implying sense if we and others had these valuable personal relations, and achieved these other forms of excellence. The optimific principles might require us to try to achieve some perfectionist aims, and to help other people to do the same. Since these are views about how it would be best for things to go, these claims could not give us reasons to reject the optimific principles.

On some views, however, we might also have some *personal* and *partial* perfectionist reasons. These are not self-interested reasons, since to achieve some perfectionist

aim we may have to sacrifice much of our well-being. But these reasons might conflict with our reasons to make things go impartially better in such perfectionist ways. Suppose that I could save either the only copy of my great nearly finished novel or the only copies of five similarly great novels by other writers. I might have personal perfectionist reasons not to choose any optimific principle that would require me to save these other people's novels rather than saving mine. But these reasons would not, I believe, outweigh my impartial reasons to choose this principle. I could rationally give up my novel to save these five other similarly great novels. If my belief were mistaken, we could again revise the Kantian Formula. But that would make little difference, since such cases would be rare.

There is another, more important possibility. Suppose that some optimific principle requires certain acts that we believe to be wrong. When we apply the Kantian Formula, we cannot appeal either to our belief that certain acts are wrong, or to the *deontic* reasons that the wrongness of these acts might provide. But we can appeal to the features of these acts that, in our opinion, *make* them wrong. And we might claim that

> (K) these wrong-making features give us decisive *non-deontic* reasons not to act in these ways, and not to choose that everyone accepts the optimific principle that requires such acts.

If there were certain acts of which (K) was true, that would provide an objection to premise (E) of the Kantian Argument for Rule Consequentialism, since there would be an optimific principle that we would not have sufficient reasons to choose. We can call this *the Wrong-Making Features Objection*.

This objection rightly assumes that, of the features that can make acts wrong, some would also give us decisive non-deontic reasons. If certain acts would cause pointless suffering, for example, this fact would give us decisive reasons not to act in these ways. These reasons would not be deontic, since they would not be provided by the fact that these acts would be wrong. The wrongness of these acts would at most give us further reasons not to act in these ways. But (K) could not be truly applied to these acts, since the optimific principles would not require us to cause pointless suffering.

(K) seems most likely to be true when applied to acts that would have good effects but would also violate some principle about the wrongness of treating people in some way. Return to

> *Bridge*, in which you cannot save the five except by causing me to fall in front of the runaway train, thereby killing me.

Suppose we believe that this act would be wrong, and that its wrong-making feature is the fact that

> (L) you would be killing me as a means of saving these other people.

To state one version of objection (K), we might claim both that

> (M) the optimific principles would require us, in cases like *Bridge*, to kill one person as a means of saving several others, since we would thereby make things go better,

and that

> (N) the wrong-making feature of such acts would give us a decisive non-deontic reason not to act in this way, and not to choose any optimific principle that would require such acts.

(M) is not obviously true. For various reasons that I mention above and below, the optimific principles would often permit or even require us *not* to do what would make things go best. But we can here suppose that (M) is true. It will be enough to ask whether claims like (M) and (N) could *both* be true.

For the optimific principles to require certain acts, it must be true that

> (O) when we consider these acts from an impartial point of view, we would have most reason to want everyone to act in these ways.

If we did not have such impartial reasons, it would not be better in the impartial-reason-implying sense if everyone acted in these ways, so the optimific principles would not require such acts. Our point of view is impartial when we are considering cases that involve people who are all strangers to us. That is true of nearly all actual cases, since nearly everyone is a stranger to us. So we can also claim that if

> (P) the optimific principles require certain acts,

it must be true that

> (Q) we would have most reason to want nearly everyone to act in these ways.

On the objection we are now considering,

> (R) some of these acts have certain features that would give everyone decisive non-deontic reasons not to act in these ways.

At least in most cases, I believe, (P), (Q), and (R) could not all be true. When

applied to *Bridge*, for example, these claims would imply that

> (S) you would have a decisive non-deontic reason not to save the five by killing me,

but that

> (T) you would also have most reason to want or hope that some stranger would arrive and act instead of you, saving the five by killing me.

On this view, though everyone would have decisive non-deontic reasons not to kill someone as a means of saving more lives, what everyone would have most reason to want, from an impartial point of view, is that everyone who can act in this way *does* kill someone as a means of saving more lives. These two kinds of reason could not, I believe, be so directly opposed. We could not have such impartial reasons to want everyone to do what everyone had such decisive non-deontic reasons *not* to do. So (S) and (T) could not both be true.

Similar claims apply to other cases. Of the features that make certain acts wrong, most give us non-deontic reasons not to act in these ways. At least in most cases, these features also give us reasons to want other people not to act in these ways. That is most obviously true of those wrong acts that harm other people, since we all have impartial reasons to want other people not to be harmed. But similar claims would apply to acts that had other wrong-making features. Suppose, for example, that it would be wrong to deceive or coerce other people as a means of producing certain benefits. The wrong-making features of these acts might give everyone decisive non-deontic reasons not to act in these ways. If that were true, could it also be true that, from an impartial point of view, we would have most reason to want everyone to act in these ways? Our answer should, I believe, be No. If the nature of deception and coercion gave everyone decisive non-deontic reasons, in such cases, *not* to deceive and coerce others, we could not also have such impartial reasons to *want* everyone, in such cases, to deceive or coerce others. That would be a strangely schizophrenic or internally conflicting view. And if we did not have such impartial reasons, the optimific principles would not require such acts.

There may, however, be one kind of exception. Suppose that, in

> *Lesser Evil*, you know that, unless you save the five by killing me, *Grey* and *Green* will save the five by each killing two other people.

Of those who believe it to be wrong to kill someone as a means of saving other people, most would believe that such an act would be wrong even if, as in *Lesser Evil*, this act is the only way to prevent more acts of the same kind. Even if this act would be wrong, however, we would have impartial reasons to want you to act in this way. Though it would be bad if you killed me as a means, it would clearly be

even worse if Grey and Green both acted wrongly in this way, by each killing two people as a means. [498] So if we learnt that you had acted wrongly in this way, thereby preventing the wrong acts of both Grey and Green, we ought to regard this fact as, in a sober way, good news. Similar claims apply if we set aside our beliefs about which acts are wrong, as we must do when applying the Kantian Formula. If everyone had such decisive non-deontic reasons *not* to act in some way, we could not, I have claimed, have impartial reasons to *want* everyone to act in this way. That would be a schizophrenic view. But we might have impartial reasons to want *no one* to act in this way *except* when such an act is the only way to prevent more such acts. That would not be a schizophrenic view.

According to the objection that we are now discussing

> (U) The optimific principles require us to act in certain ways, though these acts have wrong-making features that give everyone decisive non-deontic reasons not to act in these ways, and not to choose that everyone accepts these principles.

As I have argued, we can reply that

> (V) if these acts had such features, the optimific principles would *not* require us to act in these ways, except perhaps when such an act would be the only way to prevent more such acts.

If (V) is true, as I believe, this objection would at most apply to only a few cases, such as *Lesser Evil*. I shall now argue that, even in these cases, this objection would fail. If you are inclined to agree, you might skip the next section.

61 Decisive Non-Deontic Reasons

If you saved the five, in *Bridge* or *Lesser Evil*, you would be doing that by killing me. We can next ask whether, as this objection claims, this feature of your act *would* give you a decisive non-deontic reason not to act in this way. We can first reconsider *Tunnel*: the case in which, if you redirected a runaway train,

> (W) you would save the five, but in a way that also killed me.

This fact, we can plausibly believe, would give you a strong non-deontic reason not to act in this way. It would be awful to do what you knew would kill an innocent person. This may be why many people believe that you would merely be morally permitted, rather than morally required, to save the five by redirecting this train. But, as these people all believe, the awfulness of killing someone would not give you a *decisive* non-deontic reason not to act in this way. If you would be morally permitted to redirect this train, though you would thereby kill me, the fact that you

would be saving several people's lives would give you a sufficient reason to act in this way.

Similar claims apply to *Bridge*, in which if you caused me to fall onto the track

> (L) you would be killing me as a means of saving the five.

It would again be awful to save the five by killing an innocent person. This feature of this act might give you a strong non-deontic reason not to act in this way. As in *Tunnel*, however, this non-deontic reason could not decisively outweigh your reason to do what would save several people's lives. If *Bridge* is significantly different from *Tunnel*, as many people would believe, this difference could not, I believe, be that, since you would be killing me as a means, you would have a decisive *non*-deontic reason not to act in this way. This feature of this act might give you a decisive reason not to act in this way. But it could do that, I believe, only *by making this act wrong*. This decisive reason would have to be *deontic*. If that is true, the objection we are now considering fails. You would not have a decisive *non*-deontic reason not to act in this way.

Similar remarks apply to other kinds of case. I suggest that

> (X) if the optimific principles require certain acts that we believe to be wrong, the features or facts that, in our opinion, make these acts wrong would not give us decisive *non*-deontic reasons not to act in these ways. What might be true is only that, by making these acts wrong, these facts would give us decisive deontic reasons not to act in these ways.

The optimific principles would require several kinds of act that many people believe to be wrong. These principles might, for example, require some of us to use artificial contraceptives, or to perform or have an abortion, or to help someone to die in a swifter, better way, or to steal from certain rich people and give what we steal to the poor. If we had decisive reasons not to act in these ways, these reasons, I suggest, would have to be provided by the wrongness of these acts.

We should expect (X) to be true. If the optimific principles require some kind of act, we must have strong impartial reasons to want everyone to act in this way. If we did not have such reasons, it would not be better if everyone acted in these ways, so the optimific principles would not require such acts. Since we would have strong impartial reasons to want everyone to act in this way, we should expect that these reasons could not be decisively outweighed except by the fact that such acts would be wrong. I defend (X) further in Appendix D.

Though I am strongly inclined to believe that (X) is true, it is again worth supposing that I am mistaken. Suppose that the optimific principles require certain acts that we believe to be wrong, and that the features that, in our opinion, make these acts wrong *would* give us decisive *non*-deontic reasons not to act in these ways. These beliefs would not, by themselves, provide an objection to premise (E). This objection must claim that

> (Y) these wrong-making features would also give us decisive non-deontic reasons not to choose that everyone accepts the optimific principle that requires such acts.

Only (Y) would count against (E), by implying that there is some optimific principle that we would not have sufficient reasons to choose.

(Y) is a claim, not about our reasons for acting in certain ways, but about our reasons for choosing that everyone accepts some principle. These are quite different questions. Consider, for example, some kind of act that would be bad for us, but would give some greater benefit to others. Even if we had strong reasons not to act in this way, we might have decisive reasons both to want everyone to act in this way, and to choose that everyone accepts some principle that requires such acts. If everyone acted in this way, for example, that might be better for everyone.

(Y) seems most likely to be true when applied to acts that violate some deontological constraint. Our main example is *Bridge*. We are supposing both that, in this case, the optimific principles would require you to save the five, and that this act would be made to be wrong by the fact that you would be killing me as a means. According to (Y), this fact would give you a decisive non-deontic reason not to choose that everyone accepts any such optimific principle. We should ask what this reason might be.

Since this reason must be non-deontic, it could not be provided by the wrongness of such acts. We might appeal again to the awfulness of saving several people's lives by killing an innocent person. The awfulness of such an act, we can plausibly believe, would give you a strong non-deontic reason to want not to be morally required to act in this way. But in a case like *Tunnel*, as we have seen, this reason would not be decisive, since you would have sufficient reasons to save the five in a way that would also kill me. And if the optimific principles required you, in *Bridge*, to save the five by killing me, this would have to be because the relevant facts gave you impartial reasons to want everyone, in such cases, to act in such ways. These facts would also give you reasons to want everyone to accept some principle that requires them to act in this way. These impartial reasons could not, I believe, be *decisively* outweighed by your personal *non*-deontic reason to want *yourself not* to be required to act in this way.

In defending this belief, I shall make some wider claims, which apply to all cases. If the optimific principles require us to act in some way, the relevant facts must give us impartial reasons to want everyone, in relevantly similar cases, to act in this way. Only then would it be better if everyone acted in this way. Since we would be considering nearly all actual cases from an impartial point of view, we would have most reason to want nearly everyone to act in this way. If we choose that everyone accepts the principle that requires such acts, our choice would indirectly bring it about that most people *would* do what we had most reason to want nearly everyone to do. These facts would give us strong impartial reasons to choose that everyone accepts this principle. According to premise (E), these reasons would not be decisively outweighed by any relevant conflicting reason. We are now asking whether, as (Y) claims, there are some cases in which (E) is false.

It will help to remember here the other kinds of case that raise the strongest objections to (E). If we choose that everyone accepts some optimific principle, this choice might be very bad either for ourselves or for certain people to whom we have close ties, such as those we love. In *Lifeboat*, for example, if I chose that everyone accepts the Numbers Principle, you would save the five rather than me, and I would lose many years of happy life. This fact would give me a very strong personal reason not to choose the Numbers Principle. But this reason would not, I believe, be decisive. By choosing that everyone accepts this optimific principle, I would indirectly save many other people's lives, and this fact would give me sufficient reasons to make this choice.

We are now considering a different kind of reason. In the cases to which (Y) might apply, the relevant facts would give us strong impartial reasons both to want everyone to act in some way, and to choose that everyone accepts some optimific principle that requires such acts. But these impartial reasons, (Y) claims, would be decisively outweighed by some conflicting non-deontic reason. Any such reason would have to be much stronger than the personal reasons I have just mentioned, such as our reasons to want not to die young. Only if this reason was much stronger could it decisively outweigh these conflicting impartial reasons. There is, I believe, only one third kind of reason that might be clearly stronger than, and decisively outweigh, both such strong personal reasons and such strong impartial reasons. If we would have some decisive reason not to make some choice, despite the fact that this choice would either (1) be much better for ourselves or those we love, or (2) would make things go impartially much better, this reason would have to be provided by the fact that this choice would be morally wrong. We could not have decisive *non*-deontic reasons not to make this choice. If that is so, as I believe, (Y) could not be true, so this objection to (E) fails.

62 What Everyone Could Rationally Will

According to premise (E), no one's impartial reasons to choose the optimific principle would be decisively outweighed by any relevant conflicting reasons. In defending (E), I have appealed to several claims that I believe to be true, and then argued that, even if I am mistaken, (E) would still be true, or could be made true by some revision of the Kantian Formula. Premise (E) is in this way *robust*.

It is worth supposing that I have made yet another mistake. Suppose that, in some cases, (Y) *is* true, because we would have a decisive non-deontic reason not to choose that everyone accepts some optimific principle. Suppose also that this objection could not be met by any similar revision of the Kantian Formula. In such cases, (E) would be false. The Kantian Argument could not show that the Kantian Formula always requires us to follow the optimific principles. We would have to revise this argument's conclusion.

This argument would then be in a different way robust, since this revision would be slight. For the reasons given above, if there were cases in which (Y) was true, such cases would be rare. (Y) might be true only in cases like *Lesser Evil*, in which some optimific principle required some act as the only way to prevent more such acts. Since such cases would be rare, the Kantian Argument might still show that, in nearly all actual cases, the Kantian Formula requires us to follow the optimific principles. Kantian Contractualism would then be, in its implications, close to Rule Consequentialism. There might be less disagreement between these theories than there is between some different versions of Rule Consequentialism.

Remember next that, in *supposing* that (Y) is sometimes true, I am supposing that several of my earlier claims are mistaken. (Y), I believe, is never true. If that is so, this argument's conclusion does not in fact need to be revised.

There is, I believe, no other strong objection to (E). If that is so, we ought to accept premises (B) to (E). Everyone would have strong impartial reasons to choose the optimific principles, and these reasons would not be decisively outweighed by any relevant conflicting reasons.

Since we ought to accept these claims, we ought to accept this argument's first conclusion. As (F) claims, everyone would have sufficient reasons to choose that everyone accepts the optimific principles.

According to this argument's remaining premise:

> (G) There are no other, significantly non-optimific principles whose universal acceptance everyone would have sufficient reasons to choose.

Compared with (E), this premise is much easier to defend. If everyone accepted any such other principle, things would go in ways that would be impartially much

worse. That is what is meant by the claim that these other principles are significantly non-optimific. These facts would give everyone strong impartial reasons not to choose that everyone accepts any such principle. Since most people would have no conflicting personal reasons, most people could not rationally make this choice. And in nearly all these cases, if everyone accepted any such non-optimific principle, things would also go much worse for some unfortunate people. It is even clearer that *these* people could not rationally choose that everyone accepts this principle, since these people would have both strong impartial reasons and strong personal reasons not to make this choice. In *Earthquake*, for example, White could not rationally choose that we all accept some non-optimific principle that required me to save Grey's leg rather than White's life. And in *Lifeboat*, none of the five could rationally choose that we all accept some non-optimific principle that required you to save me rather than saving all of the five. So, as (G) claims, there are no significantly non-optimific principles that everyone would have sufficient reasons to choose. [499]

(B), (F), and (G) together imply

> (H) It is only the optimific principles whose universal acceptance everyone would have sufficient reasons to choose, and could therefore rationally choose.

When combined with (H), the Kantian Formula implies that everyone ought to follow these principles.

We can now restate this argument more briefly. Kantians could claim:

> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally choose, or will.
>
> (C) There are some principles whose universal acceptance would make things go best.
>
> (F) Everyone could rationally will that everyone accepts these principles.
>
> (H) These are the only principles whose universal acceptance everyone could rationally will.

Therefore

> UARC: These are the principles that everyone ought to follow.

(A) is the Kantian Contractualist Formula, and UARC is one version of Rule

Consequentialism. We are assuming (C). I have, I believe, successfully defended (F) and (H). So this Kantian Formula requires everyone to follow these Rule Consequentialist principles.

This argument, we may suspect, must have at least one Consequentialist premise. If that were true, this argument would be uninteresting. We would expect Consequentialist premises to imply Consequentialist conclusions. And such an argument would not give Non-Consequentialists any reason to change their view.

This argument's premises are not, however, Consequentialist. The argument assumes that outcomes can be better or worse in the impartial-reason-implying sense. But Non-Consequentialists can accept that assumption. Many Non-Consequentialists believe, for example, that it would be worse if more people suffer, or die young. These people reject Consequentialism, not because they deny that outcomes can be in this sense better or worse, but because they believe that the rightness of acts does not depend only on facts about how it would be best for things to go. This argument also assumes that there are some principles whose universal acceptance would make things go best. But this assumption is not Consequentialist. We may believe that there are such optimific principles, but also believe that we ought to reject some of these principles, because they require or permit some acts that are wrong.

Since this argument does not have any premise that assumes the truth of Consequentialism, it is worth explaining how this argument validly implies its Consequentialist conclusion.

Consequentialists appeal to claims about what would be best in the impartial-reason-implying sense. These are claims about what, from an impartial point of view, everyone would have most reason to want, or choose. The strongest objections to Consequentialism are provided by some of our intuitive beliefs about which acts are wrong.

Contractualists appeal to the principles that it would be rational for everyone to choose, if we were all choosing in some way that would make our choices sufficiently impartial. Some Contractualists claim that, to achieve such impartiality, it is enough to appeal to the principles that it would be rational for everyone to choose, if everyone needed to reach agreement on these principles. Other Contractualists, such as Rawls, add a veil of ignorance. Kantian Contractualists achieve impartiality by appealing to what everyone could rationally choose, if each person supposed that he or she had the power to choose which principles everyone would accept. Impartiality is here achieved, without any need to reach agreement or any veil of ignorance, by the requirement of unanimity. In arguing that there are principles that everyone could rationally choose, I have appealed to another feature of Contractualism. When we apply any Contractualist formula, we cannot appeal to

our intuitive beliefs about which acts are wrong.

We can now explain how, without having any Consequentialist premise, this argument validly implies its Consequentialist conclusion. As I have just said:

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

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

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

CHAPTER 17 CONCLUSIONS

63 Kantian Consequentialism

Return next to Act Consequentialism, or

> AC: Everyone ought always to do, or try to do, whatever would make things go best.

Is this principle UA-optimific, by being the principle whose universal acceptance would make things go best?

As Sidgwick argued, the answer is No. [500] If everyone always tried to do whatever would make things go best, these attempts would often fail. When predicting the effects of possible acts, people would often make mistakes, or deceive themselves in self-benefiting ways. It would be easy, for example, to believe that we were justified in stealing or lying, because we falsely believed that the benefits to us would outweigh the burdens that our acts would impose on others. If we were all Act Consequentialists, that would also undermine or weaken some valuable practices or institutions, such as the practice of trust-requiring promises. If everyone had the motives of an Act Consequentialist, that would be bad in other ways. For it to be true that everyone nearly always tried to make things go best, most of us would have to lose too many of the strong loves, loyalties, personal aims, and other motives in which much of our happiness consists, and that also make our lives in other ways worth living. For these and other such reasons, we can claim that

> (A) if everyone accepted AC, things would go worse than they would go if everyone accepted certain other principles.

These other, *UA-optimific* principles would partly overlap with the principles of common sense morality. These principles would often require us, for example, not to steal, lie, or break our promises, even when such acts would predictably make things go best. These principles would permit us to give some kinds of strong priority to our own well-being. And they would often permit us, and often require us, to give some kinds of strong priority to the well-being of certain other people, such as our close relatives and friends, and those to whom we may be related in various other ways, such as our pupils, patients, clients, colleagues, customers, neighbours, and those whom we represent. Since AC is not the principle whose universal acceptance would make things go best, the Kantian Formula does not require us to be Act Consequentialists.

We have been discussing the *universal acceptance* version of Rule Consequentialism, or UARC. According to a different version of this theory, which we can call

> *UFRC*: Everyone ought to follow the principles of which it is true that, if they were *universally followed*, things would go best.

Such principles we can call *UF-optimific*. We *follow* some principle when we succeed in doing what this principle requires. For example, we would be following AC if we always did whatever would make things go best.

We have also been discussing what we can now call the *acceptance version* of Kantian Contractualism, or *AKC*. According to a different version of the Kantian Formula, which we can call

> *FKC*: Everyone ought to follow the principles whose being universally followed everyone could rationally will, or choose.

The Kantian Argument discussed above could be revised to show that

> (B) it is only the UF-optimific principles whose being universally followed everyone could rationally will.

This other version of the Kantian Formula therefore requires us to follow these principles.

According to some writers, the Act Consequentialist principle is UF-optimific. For example, Shelly Kagan claims that

> (C) if everyone always followed AC, by doing whatever would make things go best, things would go best.

This claim may seem undeniable. And if this claim were true, this version of the Kantian Formula would require us to be Act Consequentialists. [501]

(C) is not, I believe, true. When we ask whether things would go best if everyone followed AC, we should consider all of the ways in which such a world would differ from the other possible worlds in which everyone followed various other principles. We should take into account, not only the effects of people's acts, but also the effects of people's intending to act in these ways, and having the motives that would lead them to act in these ways. [502] For some of the reasons that Sidgwick gave, we can claim that

> (D) if everyone always followed AC, things would go worse than they would go if everyone always followed certain other principles.

If everyone always did whatever would make things go best, everyone's *acts* would,

in most cases, have the best possible effects. [503] Things would go better than they would go if everyone always tried to do whatever would make things go best, but such attempts often failed. But the good effects of everyone's acts would again be outweighed, I believe, by the ways in which it would be worse if we all had the motives that would lead us to follow AC. As before, in losing many of our strong loves, loyalties, and personal aims, many of us would lose too much of what makes our lives worth living. So this version of the Kantian Formula does not require us to be Act Consequentialists.

This formula does, however, require us to follow the principles that are UF-optimific. And compared with the UA-optimific principles, these principles are more similar to AC. [504] So this version of the Kantian Formula supports a moral view that is significantly closer to Act Consequentialism.

To cover both versions of the Kantian Formula, we can restate Kantian Contractualism as

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.

Principles could be *universal laws* by being either universally accepted, or universally followed.

Since these different versions of KC and RC have different implications, we might have to choose between them. In making this choice, we would have to consider several questions that I shall not consider here. But I shall mention one possibility. We ought, I have claimed, to distinguish different senses of 'ought' and 'wrong', which we can use in different parts of our moral theory, to answer different questions. It is worth drawing other such distinctions. For example, it is one question what *we ought all ideally to do* if we suppose that we would all succeed. Our answers to this question will be our *ideal act theory*, or what some call our *full compliance theory*. It is another question what we ought to do when we know that some other people will act wrongly. Some call this our *partial compliance theory*. We can also ask what we ought to try to do when we take into account various other facts, such as facts about the mistakes that people would be likely to make, and facts about people's motives, desires, and dispositions. Another question is which motives we ought to have, and what we ought to be disposed to do. Our answers to this question would be our *motive theory*, which would itself have ideal and non-ideal parts. If we are Kantian Contractualists and Rule Consequentialists, we may not need to choose between at least some of these different versions of KC and RC, since we might appeal to these different versions, and use these different senses of 'ought' and 'wrong', in such different parts of our moral theory. [505]

There may be another complication. I have supposed that there is one set of principles that are UA-optimific, and another set that are UF-optimific. If there were two or more such sets, which were significantly different, we would have to choose between these sets of principles in some other way. There are several possibilities, which I shall not consider here.

We can now return to another part of Kant's view. According to what I have called Kant's

> *Formula of the Greatest Good*: Everyone ought to strive to promote a world of universal virtue and deserved happiness.

We can best promote this world, Kant claims, by following the moral law, as described by Kant's other formulas. Some of these formulas, I have argued, are best revised and combined in Kantian Contractualism. So Kant might have claimed:

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.
>
> (E) What everyone could rationally will to be such laws are the principles whose being universal laws would make things go best, by bringing the world closest to its ideal state.
>
> (F) This ideal state would be a world of universal virtue and deserved happiness.

Therefore

> Everyone ought to follow the principles whose being universal laws would best promote this ideal world.

This argument would give Kant's moral theory its most unified and harmonious form. [506] Kant's Formula of the Greatest Good would describe a single ultimate end or aim which everyone ought to try to achieve, and Kantian Contractualism would describe the moral law whose being universally accepted or followed would best achieve this aim.

Of this argument's premises, KC is Kantian Contractualism. The Kantian Argument in Chapter 16 could be turned, with some revisions, into a defence of (E). (F) is Kant's description of the ideal world that he calls the Greatest Good.

We ought, I have argued, to revise (F). It would be bad, Kant claims, if people had

more happiness, or less suffering, than they deserve. But Kant also claims

> (G) If all of our decisions were merely events in time, no one could deserve to suffer.

We ought, I have argued, to accept this claim. As I have said, we can add:

> (H) All of our decisions *are* merely such events.

Therefore

> (I) No one could deserve to suffer.

Nor could anyone deserve to be less happy. If we subtract Kant's claims about desert, Kant's ideal world would be a world of universal virtue and happiness. In considering worlds that are not ideal, we would again have to decide which worlds would be closer to the ideal. It would always be better, I believe, not only if there was less suffering and more happiness, but also if more of this happiness came to people who were less happy, or who suffered more. We might add that our well-being does not consist merely in happiness and avoiding suffering, and that how well things go in part depends on other facts that are not about anyone's well-being.

Kant's claims about his ideal world raise another question. In asking how we could get closest to Kant's ideal, we must compare the goodness of virtue and happiness. [507] On one view, the goodness of virtue is infinitely greater, so that if anyone became slightly more virtuous, or slightly less vicious, this change would be better than the achievement of any amount of happiness, however great, or the prevention of any amount of suffering. For this view to seem plausible, I believe, we must assume that we have some kind of freedom that could make us responsible for our acts in some desert-implying way. If there could be no such freedom, as I have claimed, we ought to accept a very different view. If someone is morally bad, by being a cruel murderer for example, this would be bad for the murderer, his victim, and others, and this would also be a bad state of affairs, which we would all have reasons to regret, and try to prevent. But the badness of someone's being a cruel murderer is, I believe, relevantly similar to the badness of someone's being insane. Such badness can be easily outweighed by the badness of great suffering.

This rejection of desert may seem to take us far from Kant's view. But Kant sometimes makes such claims, as when he refers to

> the supreme end, the happiness of all mankind. [508]

And Kant also writes:

> If we conduct ourselves in such a way that, if everyone else so conducted

> themselves, the greatest happiness would arise, then we have so conducted ourselves as to be worthy of happiness. [509]

Kant here asserts a hedonistic version of Rule Consequentialism.

I shall now sum up these conclusions. Moral principles could be universal laws by being either universally accepted or universally followed. Kantians, I have claimed, can argue:

> KC: Everyone ought to follow the principles that everyone could rationally will to be universal laws.
>
> (J) There are certain principles whose being universal laws would make things go best.
>
> (K) These are the only principles that everyone could rationally will to be universal laws.

Therefore

> RC: Everyone ought to follow these optimific principles.

KC and RC are the most general statements of Kantian Contractualism and Rule Consequentialism. We are supposing that (J) is true. I have, I believe, successfully defended (K). So Kantian Contractualism implies Rule Consequentialism.

Since that is true, these theories can be combined. According to what we can call

> *Kantian Rule Consequentialism*: Everyone ought to follow the optimific principles, because these are the only principles that everyone could rationally will to be universal laws.

64 Climbing the Mountain

Remember next that, according to

> Scanlon's Formula: Everyone ought to follow the principles that no one could reasonably reject.

Kantians might argue:

> (A) If someone could not rationally will that some principle be a universal law, there must be facts which give this person a strong objection to this

principle.

(B) If there is some conflicting principle that everyone *could* rationally will to be a universal law, no one's objection to this alternative could be as strong.

Therefore

(C) When there is only one relevant principle that everyone could rationally will to be a universal law, no one's objection to this principle could be as strong as the strongest objections to every alternative.

(D) No one could reasonably reject some principle if there are stronger objections to every alternative.

Therefore

(E) When there is only one relevant principle that everyone could rationally will to be a universal law, no one could reasonably reject this principle.

(F) Since there are stronger objections to every alternative, these alternatives could all be reasonably rejected.

Therefore

(G) When there is only one relevant principle that everyone could rationally will to be a universal law, this is the only relevant principle that no one could reasonably reject.

(H) There is only one set of principles that everyone could rationally will to be universal laws.

Therefore

(I) These are the only principles that no one could reasonably reject.

We can call this *the Convergence Argument.* If this argument is sound, Kantian and Scanlonian Contractualism can be combined. The principles that no one could reasonably reject are the same as the principles that everyone could rationally will to be universal laws.

This argument applies, not to Scanlon's present theory, but to what I believe to be the best version of Scanlonian Contractualism. I defend this belief, and discuss this argument further, in Chapters 21 to 23.

This combined theory, as I have argued, can also include Rule Consequentialism. According to what we can call this

> *Triple Theory*: An act is wrong if and only if, or *just when*, such acts are disallowed by some principle that is
>
> > (1) one of the principles whose being universal laws would make things go best,
> >
> > (2) one of the only principles whose being universal laws everyone could rationally will,
>
> and
>
> > (3) a principle that no one could reasonably reject.

More briefly,

> TT: An act is wrong just when such acts are disallowed by some principle that is optimific, uniquely universally willable, and not reasonably rejectable.

We can call these the *triply supported* principles. If some principle could have any of these three properties without having the others, we would have to ask which of these properties had most moral importance. But these three properties, I have argued, are had by all and only the same principles. If that is true, we could claim

> (J) Moral principles are not reasonably rejectable just when they are uniquely universally willable, and they are uniquely so willable just when they are optimific.

We could also claim

> (K) When some principle is optimific, that makes it one of the only principles that are universally willable,

and

> (L) When some principle is one of the only principles that are universally willable, that makes it one of the principles that no one could reasonably reject. [510]

We might add:

> (M) When acts are disallowed by some principle that is optimific, universally willable, and not reasonably rejectable, that makes these acts unjustifiable to others.

(N) Such acts would be blameworthy, and would give their agents reasons to feel remorse, and give others reasons for indignation.

(O) Everyone has reasons never to act in these ways. These reasons are always sufficient, and often decisive.

For the reasons that I earlier gave, this Triple Theory should claim to describe, not wrongness itself, but one of the properties or facts that make acts wrong. There are several other, more particular wrong-making properties or facts, such as the properties of causing pointless suffering or coercing others for our own convenience. The Triple Theory should claim to describe a single *higher-level* wrong-making property, under which all other such properties can be subsumed, or gathered. This higher-level property is the complex property of being disallowed by some principle of which (1), (2), and (3) are true. When acts have certain other properties, that makes them acts that would be disallowed by such a triply supported principle, and all these facts could be claimed to make these acts wrong. Each of these facts, we might add, would give everyone further reasons not to act in these ways.

If we accept this Triple Theory, we should admit that, in explaining why many kinds of act are wrong, we would not need to claim that such acts are disallowed by some triply supported principle. In some cases such a claim would be, not merely unnecessary, but also puzzling or offensive. This is like the fact that, after some rape or murder, we ought not to say 'What if everyone did that?' or 'What if everyone believed such acts to be permitted?' Some acts are open to objections that are both clearer and stronger than the objections to these acts that are provided by Kant's formulas, or by any version of Contractualism or Rule Consequentialism.

In many other cases, however, it may help to ask whether some act is permitted or disallowed by some triply supported principle. It may be unclear, for example, whether it would be wrong to break some law, or tell some lie to achieve some good end, or coerce someone in some way for this or someone else's good, or steal some object that its owner never uses, or fail to help some people who are in great need, or fail to vote, or have, in an overpopulated world, more than two children. If any of these kinds of act would be disallowed by one of the principles whose acceptance would make things go best, and by one of the only principles whose being universal laws everyone could rationally will, and by a principle that no one could reasonably reject, these facts would provide some of the strongest objections to these acts.

Remember next that, on the Triple Theory, an act is wrong *just when* such acts are disallowed by the triply supported principles. There are several lower level wrong-making properties, and several principles that disallow acts with these properties. The Triple Theory makes claims about what all these properties and principles have in common. If this theory's claims are true, that would give us

deeper explanations of why these principles are justified, and why these acts are wrong. One aim of such a theory, as Scanlon writes, is to provide 'a general criterion of wrongness that explains and links these more specific wrong-making properties'. [511]

For some moral theory to succeed, it must have plausible implications. The Triple Theory has many such implications. But after we have worked out what this theory implies, and we have carefully considered all of the relevant facts and arguments, this theory might conflict with our intuitive beliefs about the wrongness of certain acts. If there are many such conflicts, or these intuitive beliefs are very strong, we could then justifiably reject this theory. If instead these conflicts are significantly less deep, or less common, we could justifiably follow this theory in revising some of our intuitive moral beliefs. [512]

We have such intuitive beliefs, not only about which acts are wrong, but also about which principles or theories might be true. So as well as having plausible implications, any successful principle or theory must be in itself plausible. Only such a principle or theory could *support* our more particular moral beliefs.

Kantian Contractualism passes this test. If some act is disallowed by one of the only principles whose being a universal law everyone could rationally will, this fact can be plausibly claimed to be one of the facts that make this act wrong.

Scanlonian Contractualism may seem to be, not merely plausible, but undeniable. Suppose I claimed:

> Though my act is disallowed by some principle that no one could reasonably reject, I deny that such acts are wrong.

This claim may seem close to a contradiction. Though I am rejecting this principle, I am also conceding, it seems, that this rejection is unreasonable. And if my rejection of this principle is unreasonable, this rejection could not be justified, so I could not defensibly deny that such acts are wrong. If Scanlon's Formula seems undeniable, however, that is because this formula does not explicitly include the Deontic Beliefs Restriction. In a fuller statement, this formula might claim:

> An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject, on grounds *other than* their belief that this principle is mistaken, because it disallows some acts that are not wrong.

It would not be self-contradictory to claim that, even though some kind of act is disallowed by such a principle, this principle *is* mistaken, because such acts are not wrong.

Kantian Contractualism can be combined, I believe, with the best version of Scanlonian Contractualism. But my arguments for this belief may fail. We would then have to choose between these theories.

Kantian Contractualism could still be combined, however, with Rule Consequentialism. I have argued that

> (K) when some principle is optimific, that makes it one of the principles whose being universal laws everyone could rationally will,

and that

> (P) there are no other principles whose being universal laws everyone could rationally will.

If these claims are true, Kantian Contractualism and Rule Consequentialism fit together like two pieces in a jig-saw puzzle. [513]

Of the Triple Theory's components, Rule Consequentialism is, in one way, the hardest to defend. Some Rule Consequentialists appeal to the claim that

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

This claim is in itself very plausible, and is not challenged by any of the arguments that I have given. If we reject (Q), that is because this claim supports Act Consequentialism, which conflicts too often, or too strongly, with our intuitive beliefs about which acts are wrong. Rule Consequentialism conflicts much less with these intuitive beliefs. But if Rule Consequentialists appeal to (Q), their view faces a strong objection. On this view, though the best principles are the principles that are optimific, the right acts are *not* the acts that are optimific, but the acts that are required or permitted by the best principles. It would be wrong to act in ways that these principles disallow, even if we knew that these acts would make things go best. We can plausibly object that, if all that ultimately matters is how well things go, it could not be wrong to do what we knew would make things go best.

Rule Consequentialism may instead be founded on Kantian Contractualism. What is fundamental here is not a belief about what ultimately matters. It is the belief that we ought to follow the principles whose being universally accepted, or followed, everyone could rationally will. Because Kantian Rule Consequentialists do not assume that all that ultimately matters is how well things go, their view avoids the objection that I have just described. When acts are wrong, these people believe, that is not merely or mainly because such acts are disallowed by one of the optimific principles. These acts are also wrong because they are disallowed by one of the only set of principles whose being universal laws everyone could rationally will. [514]

If Kantian Contractualism implies Rule Consequentialism, as I have claimed, that does not make the resulting view wholly Consequentialist. Though this view is Consequentialist in its claims about which *principles* we ought to follow, it is not Consequentialist either in its claims about *why* we ought to follow these principles, or in its claims about which *acts* are wrong. This view, we might say, is only *one-third* Consequentialist.

In these chapters I have argued that, with some revisions and additions, Kant's most important claims are these:

> (R) Everyone ought to treat everyone only in ways to which they could rationally consent.
>
> (S) Everyone ought to regard everyone with respect, and never merely as a means. Even the morally worst people have as much dignity or worth as anyone else.
>
> (T) If all of our decisions are merely events in time, we cannot be responsible for our acts in any way that could make us deserve to suffer, or to be less happy.
>
> (U) Everyone ought to follow the principles whose being universal laws would make things go best, because these are the only principles whose being universal laws everyone could rationally will.

We ought, I believe, to accept (S) and (T), and we have strong reasons to accept (R) and (U).

It may be worth explaining why I have spent so long defending (U). Of our reasons for doubting that there are moral truths, one of the strongest is provided by some kinds of moral disagreement. Most moral disagreements do not count strongly against the belief that there are moral truths, since these disagreements depend on different people's having conflicting empirical or religious beliefs, or on their having conflicting interests, or on their using different concepts, or these disagreements are about borderline cases, or they depend on the false assumption that all questions must have answers, or precise answers. But some disagreements are not of these kinds. These disagreements are deepest when we are considering, not the wrongness of particular acts, but the nature of morality and moral reasoning, and what is implied by different views about these questions. If we and others hold conflicting views, and we have no reason to believe that *we* are the people who are more likely to be right, that should at least make us doubt our view. It may also give us reasons to doubt that any of us could be right.

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

I have also argued that some things matter in the reason-implying sense. There are some aims, such as avoiding and preventing suffering, that we all have reasons to want to achieve, and try to achieve. Reasons for acting all derive their force from the facts that give us reasons to have such aims.

What now matters most is that we rich people give up some of our luxuries, ceasing to overheat the Earth's atmosphere, and taking care of this planet in other ways, so that it continues to support intelligent life. If we are the only rational animals in the Universe, it matters even more whether we shall have descendants during the billions of years in which that would be possible. Some of our descendants might live lives and create worlds that, though failing to justify past suffering, would give us all, including those who suffered, reasons to be glad that the Universe exists.

PART FOUR COMMENTARIES

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 book 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.

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.

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 that 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 his concluding remarks, which give the book as a whole its name. As he notes, the formula 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" (000).

The suggestion, if I am interpreting it correctly, is that there is a single true morality, crystallized in a single supreme principle that 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 his concluding remarks 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 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* (p. 000) and *Means* (p. 000), 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 Contractualsim 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." (p. 000) 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] 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*

[1] Parfit objects, more specifically, to Korsgaard and O'Neill's interpretation of Kant's claims, according to which "(B) It is wrong to treat people in any way to which they cannot possibly consent because we have not given them the possibility of giving or refusing consent." (p. 000)

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." (p. 000) 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]

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,

---

[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 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.

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 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 (p. 000).

---

[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.

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 (p. 000). 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" (p. 000). 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" (p. 000).

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 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

---

[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.

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 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,

---

[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, p.102)

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" (p. 000).

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" (000). 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 (p. 000) 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 a single person over the bridge to prevent the runaway trolley from killing the five others who are in the trolley's path.

Parfit realizes that insofar as one imagines oneself in the positions of Grey or the single person on the bridge, one may rationally want such principles not to be followed. One may rationally want principles 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 other 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 single person off the bridge is not just the idea that this person, 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 single person for the greater good. Imagining oneself in the position of this person, one might want it to be the case that insofar as it is anyone's decision whether this person should give up his life to save the five, it should be *this person's* decision. And this thought seems to me one that can be entertained and supported even if one is not in this person's position.

---

[7] Strictly speaking, the agent in Parfit's *Bridge* case is not in a position literally to *push* the single person off the bridge, but rather to use a remote control device to cause this person 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. (See p. 000).

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 single person on the bridge alone to the principle that favors pushing him in front of the trolley.[9] [10]

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 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.

---

[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.

[10] 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 great injury on one person as a means of benefiting others." (p.212) According to Parfit, "the Harmful Means Principle is best defended by appealing to our intuitive beliefs about which acts are wrong." (p. 213-214) My remarks do not appeal to such intuitions, however.

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 the final sentences of Parfit's essay, in which he 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" (p. 000).

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 nonconsequentialist (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 conseqeuntialist 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 essay 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 forego 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.[11] 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 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.

---

[11] 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.

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 the single person on the bridge, and see a runaway trolley speeding below on its way to kill five people if nothing is done to interfere. If I push this person 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 (p. 000) 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 or strongly suspect that we 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's last sentences seem to suggest, 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 7, 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 this book he "shall use 'ought morally' and 'wrong' vaguely, in some combination of these senses" (p. 000). To his credit, he recognizes that "there are deep and difficult questions about how these concepts are related", and admits that he "shall say little about these meta-ethical questions" in this book, on the grounds that "these questions will be easier to answer when we have made more progress with our moral thinking" (p. 000). This is fair enough, and it is both striking and impressive how well Parfit characterizes this range (geographical pun unintended) of metaethical questions that he does not delve into but which, as he sees, ultimately bears on the claims he has defended in this book. *On What Matters* can fairly be said to constitute progress in our moral thinking. To assess its significance, we need to return to some of the metaethical questions which Parfit has left for another day, another hike, another book.

HUMANITY AS END IN ITSELF ALLEN WOOD

[Note to the copy-editor. This Commentary's endnotes should become footnotes, as with the other three Commentaries. I don't know how to make this change. Nor do I know how to eliminate the two black lines in the last few pages of Susan Wolf's Commentary directly above.]

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" (G 4:429-430),[515] 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, 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 systematize and correct Common Sense,[516]

which continues to exercise authority within moral theory independently of first principles, and might even help to shape the formulation of moral principles.[517]

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 principle.[518] 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".[519] 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.

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' (G 4:388) 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 relate 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.[520]

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 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.[521]

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, 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."[522] 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.

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' is 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.[523] 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 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 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 discredit the Sidgwickian program broadly conceived. Like many ambitious philosophical projects, 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.[524] 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.[525]

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.

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 ourselves 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 life boats. They should to find out why this happened, and take steps to minimize the chances of its happening ever again. We recently saw this point illustrated dramatically 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,[526] 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.[527]

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 life boat 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 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.

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 insuring 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 o 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.

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.

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.[528]

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 on that unargued assumption.[529] 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 a majority ("the taxpayers") are often used to rationalize the oppression of underprivileged minorities ("the underclass"). Appeals to rights and appeals to welfare are equally open to abuse. Hence from the standpoint of moral theory, surely the best course is to keep an open mind about what rights people have and what considerations might ground them. If it is an unargued assumption of trolley problems, and hence of the moral intuitions they evoke, that all such rights must be based solely on the considerations on which these problems focus, then that is a reason for doubting that these intuitions provide reliable data for moral theorizing.

Extreme situations. To others, trolley problems may appeal because it seems to them that the only honest way to confront many social policy decisions is to see them as frank trade-offs between the deepest interests of different people. It is simply a fact about many social policy decisions that if they are made one way, then *these* people will be hurt and if they are made the other way, then *those other people* will be hurt. But it does not follow from this fact that the correct way to view all such situations is to see them simply, or even primarily, in this light. One important reason why people are regarded as having rights or entitlements -- and why most people are forbidden or even prevented from directly choosing between the competing interests of others -- is that it is in general *evil* to decide between competing interests merely on such a basis. That is the real reason why, for instance, doctors are not permitted to carve up a healthy person in order to distribute their vital organs among five people needing organ transplants. It is also why railways workers and people walking across bridges have rights to be protected that interlopers on tracks do not have, and why bystanders are not permitted to switch the points on trains or operate trapdoors in bridges in order save five by killing one.

There are some extreme and desperate situations in human life – such as war or anarchy, or sometimes pestilence, famine or natural disaster – in which it can look as if the only way to think rationally about them is simply to consider coldly and grimly the numbers of people, the amounts of benefit and harm, and the kind of actions available to you that will produce the benefit and harm. But it is significant that we should think of such decisions as being made coldly and grimly, calculating consequences with a kind of economist's tunnel-vision while totally denying all our normal human thoughts and feelings. For those are situations in which human beings have been deprived of humanizing social institutions (like those that should provide enough lifeboats, prevent runaway trains and trolleys, keep interlopers off tracks and bystanders away from switches, and so forth) that make it rationally possible *not* to look at matters in that way. I grant you that trolley problems might help you to think in a rational (if dehumanized) fashion about situations in which that is the only way left to think about them because the situations themselves have already been dehumanized. That is a powerful argument *against* using trolley problems in moral philosophy.

We think of war as a morally unacceptable condition, in large part because in war it can indeed seem rational for people to think about their lives and the lives of others in truly monstrous ways. One of our primary tasks as human beings is to view things in better ways, and if necessary to make changes in the world (regulating the behavior of doctors and trolley systems) so as to bring it about that there are other ways of viewing things rationally. If you take some part of human life (such as health care delivery) which is not inherently as barbarous as war, and come to regard this as the only rational way to think about it – or especially if you come to

regard this as the only rational way to think about the fundamental principles of morality generally -- then that amounts to a voluntary decision on your part to turn health care, or even human life as a whole, into something horrible and inhuman, something like war, that ought never to exist.

The realm of ends. FH, the principle that humanity in every person has dignity as an end in itself, may give us reasons refusing to look at the world in the way trolley problems tend to induce us to look at it. But perhaps the Kantian ideal of a realm of ends provides an even more direct route to the same conclusions. It implies that we should not think about moral problems in terms of trade offs between competing human ends, but should try to understand the answer to every problem as one that treats all people as ends, and leaves out no human ends except those that exclude themselves from the harmonious system (or realm) of all rational ends. For in a realm of ends, no one would have to choose between one life and five simply on the basis of numbers – since every life, considered simply as such, would have equal dignity as part of the realm of ends. Thus no one's life would have to be sacrificed unless their actions excluded its preservation from the harmonious system of ends.

No doubt human vulnerability to nature, and even more human wickedness, will forever prevent there actually being such a realm of ends. That is why there will probably always be such things as hurricanes, shipwrecks, unjust economic systems and wars. That is why there are problems about the distribution of such things as healthcare that (especially in a fundamentally unjust and inhuman society like ours) seem to come down to stark tradeoffs between the deepest interests of different people and groups. Consequently, there will always be a place for the kinds of issues trolley problems are meant to address. That is my one concession to those philosophers who like to think about trolley problems. It is a significant concession, but a much more limited one than it might at first seem. For because people can, to some extent, create a realm of ends in their relations with each other and in their ways of thinking about these relations, it also means that these problems are not as universal in their moral significance as many philosophers think. Because the actual operation of trains and trolleys, for example, is subject to a considerable degree of responsible human control and regulation, runaway trolleys are not in fact very good examples of situations in which there arise the kinds of problems the trolley problem philosophers want to address.

More importantly, trolley problem cases do not represent the fundamental issues with which moral principles must deal. On the contrary, these kinds of problems mark the limits of the power of moral thought to deal with problems of human life. The kind of thinking they force on us rather constitutes the way we have to think about things precisely where our moral aspirations have essentially failed. If it ought to be our chief moral concern to make human life, as far as we can, into a realm of ends, then from the standpoint of morality preventing people from having

to think about competing human interests in ways trolley problems encourage you to do always takes precedence in principle over any rule or policy about what an agent should actually *do* in a situation such as *Lifeboat*, *Tunnel* or *Bridge*. If that is true, then the use of trolley problems by moral philosophers to test fundamental moral principles involves a deep misconception about the ways of thinking that should be fundamental in moral philosophy.

Fans of trolley problems have suggested to me that these problems are intended to be philosophically useful because they enable us to abstract in quite precise ways from everyday situations, eliciting our intuitions about what is morally essential apart from the irrelevant complexities and "noise" of real world situations that get in the way of our seeing clearly what these intuitions are. But I have already suggested why I cannot accept that. Trolley problems seem to me to abstract not from what is irrelevant, but from what is morally vital about all the situations that most resemble them in real life. At the very least, trolley problems presuppose (rather than establish) that certain things are morally fundamental, and my own view is that these presuppositions are at least highly doubtful, probably perniciously false, and that trolley problems (or people's responses to them) do nothing at all to support or confirm these presuppositions. Instead, they only provide a kind of illegitimate pseudo-support for them, as well as the opportunity to do moral philosophy in a manner that encourages us not to question them.[530]

Notes

A MISMATCH OF METHODS BARBARA HERMAN

1. Derek Parfit's *On What Matters* offers an avowedly hybrid theory of morality, or at least of the part of morality that tells us which acts are wrong. The theory is elaborated by way of an extended and inventive critical reconstruction of Kant's ethics as a kind of contractualism. What makes it hybrid is the conjunction of the contractualist framework with an account of value that is for the most part concerned with outcome effects on well-being, taken in a very wide-ranging way.[12] Despite the embrace of a Kantian contractualist framework – the fundamental aim of morality is not to make things go best, but to find principles of action that everyone could rationally will (p. 000) – since the values that inform rational willing are (for the most part) about what is nonmorally best, the hybrid theory winds up having a strongly consequentialist cast.

That a normative theory is hybrid is not in itself grounds for criticism. What is puzzling is a hybrid methodological approach to understanding the ambitions of Kant's moral theory, since it is anything but hybrid. The defining feature of Kant's theory is that goodness is a function of, and not independent of, moral principle.[13] While I think Parfit is often correct in rejecting some of the versions of Kantian claims and arguments that he finds in the literature, I don't think his revisionary interpretive project, which aims to elicit the best in Kant's ethics by evaluating and revising its claims in terms of nonmoral good outcomes, can capture what is most distinctive about Kant's theory. The mismatch of methods is too profound.

For the mismatch of methods to be a source of serious worry, we would want to know two things. One is that it really does have far-reaching and distorting effects on moral judgment and thought; the other that there is a version of Kant's ethics as a unified (non-hybrid) theory that is plausible. These are larger projects than can be attempted here. What I aim to do instead is work through some examples that show the depth and extent of the mismatch problem, and then offer some interpretive resistance to the hybrid arguments that provides a better fit with

---

[12] For Parfit, reasons that bear on judgment and action are value-responsive, though, here following Sidgwick, Parfit holds that personal and impersonal reasons enter moral judgment with separate and independent weight.

[13] This is the point of the Paradox of Method (*Critique of Practical Reason* 5:63): well-being considerations are facts that support preferences, but not values (at least not directly).

what Kant says, and hews to the spirit of the unified project. There will not be space to fully cite or defend each and every claim I make about Kant; the claims will perforce be provisional, their value in the plausibility and distinctiveness of the interpretations they suggest. I will argue at greater length that two regions of normative worry that prompt the demand for hybrid repair – making moral space for our personal concerns and the power of good ends to justify prima facie wrong actions as means – are not problems for Kant's theory when it is not interpreted in an unnecessarily narrow way.

2. To elicit some of the elements of the mismatch, let's begin by considering the way the hybrid revisionist approaches Kant on lying.[14] First, he looks at things Kant says: that telling lies is among the morally worst things a person can do, that lying is wrong because lies fail to respect the value of rational agency (especially the liar's own agency), and, famously, that one may not lie, regardless of the consequences. Most of this is deemed obviously incorrect. But what is not thought incorrect is the Kantian idea of morality restricting actions on the grounds that a principle permitting them cannot be rationally willed (though Kant's own view of how to understand the condition of universal rational willing is regarded as mistaken or confused). Then the revisionist offers a better, although hybrid, argument. It will go something like this. Depending on circumstances, lies can be either beneficial or harmful. Most often they are attempts to secure some advantage for the liar by controlling the information available to victims (though controlling information can also be beneficial and so possibly rational). When advantage-lying is widespread, it undermines the trust conditions necessary for cooperative activity, itself a great good. Therefore, a principle of general permissiveness about lying would not be rational to will: since lying is so often a useful means, permissiveness would likely lead to more lying than trust, and so cooperation, could survive. But a principle that permitted lying when necessary to save wrongfully threatened lives would not be interfering with interests we have reason to protect and would have little or no undermining effect on trust. So advantage-lying is shown to be wrong; not all lying is wrong; and the rationale for the wrongness points not to the value of rational agency, but to the benefits of cooperation. In this way, the revisionist retains the Kantian (contractualist) spirit and get a much more plausible moral view. The consequentialism figures in the revisionary account twice – in the values appealed to and in the treatment of the universality condition as setting up a comparison between how we would fare were advantage-lying, as opposed to life-saving lying, permissible.

---

[14] This is not an exact report of Parfit's discussion of lying, but a compressed variant that captures its main elements.

> Now, whatever the correct view of Kant on lying is, a best version of *it* is not going to be found in the terms of costs and benefits, and not through an argument that appeals to the comparative cost/benefit value of a selectively permissive principle's general acceptance. That way takes lying (and truth-telling) out of the center of the moral theory, and regards its moral significance merely instrumentally. But if the ambition of Kant's moral philosophy is a unified theory of value and principle *within* an account of practical reason, if it's supposed not to be possible for a maxim of lying to be principle of rational willing, we ought to be looking at the relation lying creates between rational agents as one that in some way violates a principle of (or implied by) their common rational nature.[15] However such a view is laid out, well-being outcomes won't be given an independent role in the argument.[16] Granted it's not easy to say what it could mean to take rationality or rational nature to be of value,[17] and I agree that the idea of respect for persons that is supposed to follow from it risks being either empty or a container for one's preferred account of human status. Nonetheless, if there is a deep insight that Kant offers, whatever the difficulty of working it out, it is nowhere else than in the account of value and the principles of action-evaluation derived from the constitutive principles of a rational will. It may be that Kant's theory cannot realize its ambitions, but as I hope to show later on in this paper, I don't think the best interpretation of Kant has yet reached that stage of the dialectic.[18]

---

[15] I will return to Kant's account of lying in section 8.

[16] As I read the *Groundwork* tests, they do not ask which of two hypothetical worlds would be *better* for us, but rather which principles of action are consistent with constitutive norms of rational willing. Compossibility is not the kind of outcome the hybrid theorist has in mind. Kant thinks that were we all to act morally we would realize a kingdom of ends, and our actions and maxims are to be consistent with that effect, the kingdom of ends does not represent an *outcome* value in the sense of providing an aim or reason for action. The same is true, I believe, of Kant's notion of the Highest Good: it is not an object we can aim for except in the sense that we seek our own and others' happiness in morally directed ways.

[17] Parfit's own struggles with this set of ideas are exemplary and informative (cf. *OWM* 134-140) – he has an unerring feel for the wince feature of appealing but bad arguments.

[18] To be clear about this, I do not mean to suggest that there is a way to make Kant's Formula of Universal Law work after all. I don't think there is. But since I also doubt that the Formula was ever intended to do the work of establishing permissions and requirements (it can explain the wrongness in wrong action, but cannot by itself tell

3. If the example of advantage-lying displays one aspect of the mismatch in methods, a different register of the divide between Kant's theory and Parfit's methodology can be seen in their treatment of the role of motives in assessments of wrongdoing. For Parfit, it is almost never the case that wrongness of action is determined or even affected by an agent's motive. If, as he argues, the values that justify moral principles look to outcome-events (outcomes that would come about were a principle generally accepted), then (most) wrongful actions will either generate bad outcomes directly, or they are of a kind which if (believed to be) permitted would summatively generate bad outcomes (or significantly worse outcomes when compared with the consequences of agent's acting in conformity to some competitor principle). What makes an action wrong is then directly a function of what does or would happen, not about why an action was done (the motive here regarded as a cause of action[19]). Motive may matter to other questions – about character, reliability, the kinds of relations a person acting from this or that motive can reasonably sustain – but it does not figure in the explanation of the wrongness of wrongful action.

So a selfish motive won't make a rescue wrong, and even a morally bad motive won't transfer its negative quality to a morality-conforming action that it brings about. In a related example, Parfit has us imagine a coffee-ordering gangster, motivated to do whatever it takes to make the world conform to his desires. He is ready to cause all kinds of mayhem if anyone crosses him, and regards the barista as he would a potentially recalcitrant soda machine that he will lash out at if it balks at dispensing his drink. But no one does cross him; the coffee is ordered and paid for. Since the act is one that satisfies moral principle (paying for purchases, or somesuch), nothing bad has happened. He is a nasty guy you wouldn't want to have around, but for all that, unless and until he does something impermissible, the moral problem is all a matter of potential and probabilities – of bad motives, not bad action.[20]

---

us which actions are wrong), other elements of Kant's theory must be brought in to do that.

[19] Sometimes when Parfit talks about motives he means the attitude an actor has in acting: whether I regard you as a rational person, a moral subject, or as a mere means. Since quite nasty attitudes can coexist with permissible actions – the attitude's negative potential remains unrealized – the issue of relevance to wrongness is the same.

[20] The reward-motivated life-saver may seem to be a purer case since there seems to be no question that he aims to do something good; but suppose that as he is swimming to

One might wonder whether it is true that nothing bad happened. The barista was surely put at risk in ways he ought not to have been. Were we to assume motivational transparency, it would seem odd to say that nothing wrong has happened if you escape harm only by avoiding eye-contact or placating or doing whatever is needed to avoid setting off those around you who are motivationally primed for easy violence. Making it through a minefield is not a walk in the park. But let's leave this worry aside.

One aspect of our interest in moral wrongness would seem to support the irrelevance of motive conclusion. If we are attending to wrongful action with an eye to (possibly or even in principle) interfering with it, so long as the gangster does what is in the circumstance required – he pays for his coffee – there is nothing happening that we should prevent (and in the case of the reward-seeking life-saver, we might even have reason to help him). The more general thought would be that to judge an action wrong we must also hold that it would be better (morally better?) if it didn't happen – that its happening is an occasion, at the least, for regret. And of course it is not better that the coffee not be paid for or the life not saved. Regrets about the action seem irrational here.[21]

However, the conditions for regrets about others' doings are often different from those that apply to the agent acting. A reformed gangster might reasonably look back at the coffee scene with a kind of horror: there I was, he thinks, ready to take the guy out if he said one off word to me! It's easy enough to imagine him concluding that what he had done was wrong: it was a matter of sheer luck that there was a benign outcome. It would not be inapt for him to wish it had not happened: not the paying for the coffee, of course, but the entire episode. If a sign of wrongdoing is guilt, or a sense that apology might be in order, motive or attitude can suffice to trigger it, and a change in attitude is often integral to the work of moral repair for what was done. (That the subject of one's action be aware of the wrong done to him is not necessary for apology to be in order.) These are reasons for thinking that the moral bearing of an agent's attitude or motive touches more than the quality of his character or the associated likelihood that he will do as morality directs. They show an agent acting in a way he should not have. They are not reasons hybrid theory wants to register in its account of wrongness, because its consequentialist account of value propels it to implicitly model wrongness on a

---

the rescue, a greater reward is announced for saving a victim downstream: what does he now have reason to do?

[21] And there are lots of things, away from the action, we get to criticize or regret about the agent.

legalistic notion of impermissibility.[22] And while impermissibility may fairly mark out the class of wrongful actions that are wrong no matter what the agent's motive, it need not, and for Kant, as far as I can tell, it does not, exhaust the category of moral wrongness in acting.[23]

So what is it in Kant's view that could make motive relevant to determinations of wrongness? Why, in the moral assessment of an action, should we care about its underlying cause? I would put it this way: for Kant, wrongness marks incorrect *ways of acting* and not merely actions that fail to conform to principles applying to action-(intended) outcome pairs. An agent who ignores or fails to respond appropriately to the morally relevant features of his circumstances acts in a way that is wrong. And this is so whether or not his external action and (intended) end are what they would be if he had acted correctly. Returning to the gangster: in a narrow (legalistic) focus, he orders and then pays for coffee – nothing is wrong. But when we widen the focus, more is going on. For one thing, he doesn't see the ordering as calling for payment; he'll pay if nothing provokes him. Nor does he see his *ordering and paying for* coffee as a required way of getting it: he would steal from the coffee seller if that was worth the trouble (*OWM* 91). So if one thought with Kant that wrongness arises from the principles of the deliberating agent and is about whether, through them, she has a sound route of reasoning to her action, the gangster would be in the wrong twice over.[24] Since it is the agent's motive that is responsible for the correct elements playing the correct role in the production of an action, motive matters to the wrongfulness of what is done. On Kant's view, as I understand it, avoiding impermissibility and avoiding wrongness are not the same thing; actions can be "not impermissible" and yet wrong.[25]

---

[22] I suspect the legalism is quite deep. After all, if one thinks that motives matter, in asking whether it would be rational to will the universal acceptance of acting "this way" – acting to save *sub specie* getting a greater reward, paying the tab *sub specie* its being the path of least annoyance – wrongness would be motive-sensitive. If, on the other hand, one thinks about wrongness by analogy with what cannot be lawfully brought about, motives are not relevant. But an analogy does not provide an argument for regarding moral wrongness in this way.

[23] Just as acting "according to duty" is not the same as acting as one ought.

[24] With others, I read the universality condition on the Kantian side as about form: a requirement that materially conditioned practical inference satisfy a matter-independent standard of correctness.

[25] Interestingly, we can speak of degrees of wrongness, though not degrees of impermissibility. When I know something untoward will happen to you as a result of what I do, but I do not value my action because of it, that is less bad (in the dimensions of

4. At this point, I can imagine someone asking whether the version of Kant I'm putting forward doesn't ignore or elide his famous distinction between morally worthy and duty-conforming actions, the former requiring that the action be done from a moral motive, the latter motive-indifferent. A first thing to note is that the question already suggests a position: that moral worth is something added-on, post permissibility, as it were. *Given* an action according to duty, *that same action* would have moral worth if done with a special attitude or from the motive of duty. Such a description misses the point of the distinction *moral worth* names. In looking to the moral content of the maxim on which the agent acts, Kant points to a condition of the *action's* value (not, as the question suggests, the *agent's* value). An externally conforming action that lacks moral worth is a behavior whose connection to moral correctness is conditional or accidental. It is in that sense *not* a (morally) correct action. There may be epistemic barriers to determining whether an action is correct or not (though one shouldn't exaggerate opacity: we often can tell when an agent is not acting correctly by seeing how she responds to failures), and there are independent reasons why we might not want to interfere with actions that are in external conformity with moral principle (there are also often good reasons not to want to interrogate agents about their conforming actions unless we are in an instructional or advisory relation with them). If I am the person acted upon, if I am not intimate with the agent, or not relying on her as a moral reasoner (we engage in one-time transactions, not long-term or complex projects), then her getting things *as if* right (according to duty) may be enough. But from the point of view of the deliberating agent it is not the same: how she regards what getting it right amounts to partly determines what she is doing.[26]

I think that the tendency to think that moral worth is about something else accepts the idea that there is a clear notion of "doing the right thing" that survives coming to do it the right way or the wrong way. We are uneasy about this sort of idea in other areas – addition, belief- formation – where we judge accidental correctness as it tracks the genuine article, and so correct, but once removed. An unjustified true belief is of course true, but it is also *qua* belief (that is, strictly) incorrect or wrong or defective. I think Kant has a similar view about moral worth and wrongness. An action that has moral worth, one done from the motive of duty, is an action

---

wrongness) than my directly intending it, or seeing it as a positive effect or even a second-order motivating benefit (that is, I would not act as I do for the extra benefit, but it might add to the value of my action weighed against some other option).

[26] The formal requirement is that one act only on maxims through which one *can* at the same time will...and not, act in conformity with that principle through which one *could* at the same time will....

arrived at under the non-accidental regulation of moral principle (that's what it is to act from a motive of duty).[27] The primary notion is not one of avoiding *getting something wrong* (acting contrary to duty), but of *getting something right*.

The doctrine of moral worth is not the only place where Kant is taken to be offering a motive-independent notion of wrongness; also noted are his views of perfect duties and duties of justice. Neither view supports the general thesis of motive-independent wrongness. In both cases, the error in thinking that they do is instructive.

Perfect duties are described in the *Groundwork* as duties "that admit no exception in favor of inclination" (4:421n), and so seem to be motive-independent.[28] But since inclination is only one kind of motive, or source of motives, the description leaves it open whether perfect duties might admit of exceptions in favor of motives of a different sort. And this makes sense, given Kant's theory of action, where motive is an agent's source of interest in an end, and so in action as a means (mere efficacy of means doesn't justify acting). Motives range widely, from such inclination-based concerns for self, family and friends to the rational interest we have in moral ends. Since the same action-end pair can hang together quite differently for agents with different motives, it is possible that some kinds of action (deceitful promises, say) could be wrong when employed as a means for any end of self-interest, but not wrong if the end is supported by a moral interest in saving a life. (Its not the intended end qua good state of affairs that justifies; the motive condition implies that justification depends on an agent's having a morally correct conception of her end. I will have more to say about this condition later.)

Duties of justice are indeed about external actions only; motives are not relevant to their correct performance. However, duties of justice are not one of the classes of moral duties, on all fours, as it were, with duties of truth-telling or aid or respect or friendship. For Kant, they are institution-based duties whose point is to secure conditions of equal external freedom (a "like liberty for all" condition).[29] They only

---

[27] That is, moral worth is not just about an attitude one has towards one's action, at least no more so than weighing evidence is just an attitude one might have about belief-formation.

[28] Imperfect duties impose requirements directly on ends, only indirectly on actions.

[29] For a very clear presentation of the point and nature of duties of justice, see Arthur Ripstein, "Authority and Coercion," *Philosophy and Public Affairs*, 32:1 (Winter 2004). 2-35. There is a moral duty to enter the state so that there might be duties of justice and so morally sanctioned regulation of property and contract.

come into existence through the legislative activity of a state (or civic union with the authority to compel compliance).[30] Theft, to take Kant's central example, is contrary to a duty of justice not because it's an instance of free-riding on a convention (Parfit is quite right in requiring that we be able to distinguish decent from both trivial and abhorrent conventions). Rather, for reasons having to do with the conditions of human rational agency – we must be able to rightfully exclude others from the use of some things – we are under moral compulsion to live in a state where the boundaries of property are settled by law and have a duty to abide by its laws. (This is one of the ways we show that the rules of property have a different status, a different kind authority, than the rules of a chess club.) *Given* property laws, appropriating what belongs to someone else is wrong. It is a violation of a duty of justice even if the appropriation is for a morally good end. Such a purpose may be a reason for a court to be lenient, or, perhaps, for the law to be written with something like a moral eminent domain clause to cover such cases (one might then be acting as an agent of the state in taking what's needed to save a life). The *moral* wrong in stealing involves the invasion of a region under the rightful authority of another; but its wrongness depends on the region being defined by and under the protection of a state (or other system of enforcement).[31]

In sum, neither the doctrine of moral worth, nor Kant's account of perfect duties, nor his introduction of duties of justice support the view that the fundamental category of moral wrongness for Kant is motive-independent. While this is not enough to make the case for the relevance of motives to moral wrongness, it should be enough to give us reason to think more about what's at stake here. After all, Kant's treatment of moral action need not square with a contemporary agenda that focuses on standards of impermissibility.

Where does this leave us in thinking about the challenges Parfit directs at Kant's moral theory? If the separation of the two methodologies is so wide that there is not ground for agreement even about the kind of thing we look to when we assess wrongness in action, then, apart from interest in specific topics, there may not be

---

[30] They are an essential part of a complete moral theory because they provide necessary background conditions for many moral obligations. Because we have a *moral* duty to obey the law, we can fulfill duties of justice "from duty." But in that sense, we can also obey traffic laws from duty; that doesn't make "no right on red" a moral duty.

[31] For this reason, a violation of a duty of justice is to be regarded as an act against the state. Duties of justice (or *recht*) include requirements to pay debts and keep promises, but only as they occur in the context of contracts. The *moral* duty of promise-keeping will have a different source, and its violations may or may not be motive-independent.

much to be gained from a point-by-point comparison of interpretations of Kant's arguments and Parfit's hybrid reconstruction. They are simply too far apart.[32] With that in mind, I propose to use the rest of this paper to do some work on the Kant side of things: since part of the appeal of the hybrid theory comes from its avoiding or transcending perceived limitations of Kant's views, if there is a better interpretation of Kant that is not limited in those ways, we may yet make some progress.

5. Although Kant's theory has a great deal to say about one aspect of morality we care about – that our pursuit of ends be constrained by rational principle – it has seemed distressingly insensitive to another – that its principles not direct us to act in ways we would find, to use Parfit's word, awful. Because the hybrid theory is directly responsive to our natural concerns and nonmoral ends, it presents itself as more reasonable – its prohibitions and permissions, if honored, would make things go better for us. Kant's theory, by contrast, can seem indifferent to what we care about, implacable in its demands, even when the outcomes it blocks are self-evidently good. Now were it true, as I suggested in the previous section it might be, that Kant's framework allows that in freighted circumstances we *can* sometimes be justified in acting in ways that we normally must not, we might conclude that there is reason to rethink the terms of Kantian moral requirement as well as its fit with ends we care about. In that spirit, I will offer a sketch of the elements of a deliberation-centered reading of Kant's moral theory, with some focus on its treatment of nonmoral ends, and then return to the case of the necessary lie to see whether the theory, so-interpreted, can do better with the moral problem of ends and means.[33]

First, the sketch. Morality, for Kant, belongs to the domain of practical reason – its principle is practical reason's first principle. To speak of reason, whether practical or theoretical, is to indicate a subject-matter that is about warranted transitions from thought to thought, thought to belief, thought to intention or choice (or between the propositions or sentences that represent them). So if the categorical imperative is or expresses a principle of practical reason, then it is a principle of inference,

---

[32] This takes no position on the best version of contractualist theory inspired by Kant.

[33] The first parts of this account are drawn from my "Reasoning to Obligation," *Inquiry* 49:1 (February 2006), 44-61.

directing (correct) reasoning from one place to another in just the sense that *modus ponens* does – though by a different rule, of course.[34]

Take one of Kant's examples.[35] Someone gives me something of value to hold for her; no one else knows I have it; she dies before it is to be returned. *Correct* practical reasoning takes me from some premise about ownership to some conclusion about what is to be done by way of a principle or rule of inference that in its most abstract form says: "act only on that principle (maxim) that can at the same time be willed a universal law." If my principle is instead "to increase my property by every safe means," then it directs my reasoning to an intention to keep the object in a way that involves a contradiction in just the sense that it would if I used a principle that warranted reasoning to not q from p, and if p then q. That is, *my* principle is not a possible instance of the correct principle of reasoning-to-action. There remain, to be sure, the familiar difficulties – how to formulate maxims, the proper understanding of universalization, etc. – but, setting them aside, it seems to me most plausible that this is the right way to approach Kant's account of moral reasoning and his account therefore of what makes actions wrong.

Now if practical reasoning emulates the form of reasoning in general, it needs access to true premises, and those would have to be premises of or about ends. That many have thought Kant clearly asserts otherwise comes, I believe, from a confusion about the argument structure of the *Groundwork*, one that mistakes a claim about the condition of application of the categorical imperative (which *is* end-independent[36]) for a claim about the irrelevance of ends, or premises of ends, for moral reasoning.[37] If, as Kant thinks, reason can determine the will to action, its

---

[34] Part of Kant's purpose in insisting on the possibility of synthetic *a priori* judgment is to extend the domain of necessary connection between cognitions.

[35] *Critique of Practical Reason* 5:27-28.

[36] The possibility of a categorical imperative depends, Kant says, on "a practical proposition that does not derive the volition of an action analytically from another volition already presupposed (for we have no such perfect will), but connects it immediately with the concept of the will of a rational being as something that is not contained in it" (*Groundwork* 4:420n). It follows that the practical proposition, as a rule of inference, applies a rational standard to all willed action – means taken for some end – without regard to the content of the agent's willed end.

[37] See *Groundwork* 4:414-415 where we are clearly set up to expect an account of objective ends or goods. It is *further* ends that are excluded. Humanity as an end-in-itself is an objective end, but it is a formal end, and, uninterpreted, cannot anchor deliberation.

principle ought to tell us that there are ends we may not have as well as ends we must have.[38] Where there are ends we may not have, *no* reasoning from them can be sound. And with morally required ends, any intention correctly derived from them will have a moral content – that is, the agent's conception of her action will draw down from the end a moral point or purpose in so acting, as well as a sense of the action's material efficacy.[39] Since, as we shall see, necessary or obligatory ends offer moral housing for our nonmoral interests, they are the right kind of thing to look to if we want to see whether and how what matters to us personally also matters morally. And last, obligatory ends will, if anything can, offer resources of justification beyond the familiar universalization rule. In exploring this possibility, I will examine two cases, one where Kant clearly does think a morally necessary end justifies normally forbidden means, and one where he should. In both I will argue that the justification the end provides is not, in any ordinary sense, instrumental. Given the place of such ends within practical reasoning, this is as it should be.

6. Kant argues that there are two, and only two, obligatory ends: of our own perfection and of the happiness of others.[40] Slightly filled out they amount to this: towards ourselves we have the end of developing and maintaining our moral and rational abilities; towards others we are to attend to the agency-related effects of our actions on their pursuit of happiness. Kant's argument for these two ends is brief and obscure; for our purposes, an intuitive gloss should suffice.

We begin by asking what, from the point of view of practical reason, demands attention? With respect to *actions*, we are not to act on any principles inconsistent with universal law-giving (that we cannot also will to become universal law). A different kind of problem arises in the normal course of adopting *ends*, whether or not they prompt us to wrongful actions, as we develop and pursue our idea of

---

[38] So Kant plainly argues in *Metaphysics of Morals* 6:385.

[39] That there need not be two equi-fundamental principles, one for actions and one for ends, is the point of Kant's "Paradox of Method": the moral law is a positive synthetic *a priori* principle for the correct use of the faculty of free willing – in objectively determining good willing, it must be or determine the will's principle *and* its object.

[40] *Metaphysics of Morals* 6:386. There is in fact another kind of required end that comes from the side of *Recht* that obligates us to create and support the state. But, as we shall see in section 7, these are not, strictly, possible ends for each separate individual: no one can act for them unless others do so as well.

happiness. For what we pursue under this idea may not be, from the point of view of practical reason, acceptable. Sloth, greed, sloppiness about what we believe and how we reason, neglect of core abilities, can be the effect of the pursuit of happiness when it is not under the regulatory control of any higher idea than some ordered satisfaction of our inclinations.[41] These effects are not themselves ends, nor likely intended; it is rather that, given our psychology, they are examples of dangers to practical rationality that won't be averted unless we have special reason to attend to the possible effects of some of our ends on our rational functioning. If we do, some of our ends will have to be abandoned; others will need to be pursued in more reason-friendly ways; ends we may not have wanted to have we may have to take on, given rational needs or rational damage that must be repaired. From the point of view of practical reason, it cannot be a matter of indifference that our psychology, which is affected by what we do, is vulnerable to disabilities that can render us less able to respond to rational requirements. The problem is not that we will then be moved to do wrong; even if, by a fluke, we never do, we would not be reasoning-to-action well, and so not willing as we ought. In this way we get the obligatory end of one's own perfection, which gives rise to various duties-to-self.

There is a parallel story for the obligatory end of the happiness of others. Although each person necessarily pursues her own idea of happiness, others have a large effect on the pursuit at almost every step of the way, from the array of ideas we are given about how to live to the provision (or not) of all kinds of help. But suppose we ask, as we did above, what can be at issue here from the point of view of practical reason? That is, why should someone else's pursuit of happiness be made, by practical reason, of concern to me? Presumably for the very same reasons we just canvassed in the obligation to self: an additional but equally fundamental fact of our psychology is that we are not monads, not autarchic systems of desire. A person's rational abilities are (partly) formed by others, (partly) sustained by them; rational abilities are vulnerable to the effects of poverty, humiliation, and sustained misdirection. At the extreme, making someone's life too hard *or* too easy can affect their ability to sustain or value rational activity. Ignoring the awkward personification, the question then is: How could impartial practical reason be indifferent to activity that undermined our own *or others'* ability to engage in reasoning?

---

[41] One might think that rational prudence would do the work here: the thought that we are a perduring being and that our future selves have a claim on our present attention. But many of these vices give us no reason to want the future to be different from the present. They affect the horizon of our practical imagination, leaving us with no reason to expect projects and needs that may come that should constrain how we treat ourselves now.

What do obligatory ends require of us? They are to shape both our pursuit of and our idea of happiness. They do not require that when deciding between going to a concert and spending an evening with friends I should deliberate about self- (or other-) improvement; that would be absurd. But they do imply that if I work so much that I have no time for friends or pleasures, I may be neglecting myself in ways I ought not, or may have failed to understand the material conditions of healthy human agency.[42] There is something of general concern that I should not have ignored. Likewise, as my actions affect others, I may not be indifferent to costs I impose, and never casual about respect. When the norms or standards that obligatory ends provide are not met, our willing is morally faulty. That is, unless there is a course of reasoning from the obligatory end(s) to the action, it is not fully justified (regardless of whether the action is externally permissible).[43] And while the obligatory end is usually not the only premise in reasoning, and often not the active one in determining choice, it should always be one of the agent's practical premises.

Suppose, oblivious or indifferent to the effects of my plans on my (or anybody's) rational agency, I decide to spend the weekend at the beach as a happiness-promoting kind of thing. Do we really want to say that there is moral fault in doing this – in going to the beach to have some fun? There's an analogue notion of "acting from" and "acting according to" for ends. What made the prudent shopkeeper's action seem unobjectionable is that it was the action that would be performed by someone who willed well – that is why it is according to duty. Likewise, what makes the end of going to the beach seem all right is that it is the sort of end (seeking enjoyment) that could be an end for someone acting under the authority of the obligatory end of self-perfection. But in just the way that the prudent shopkeeper's action is morally unstable and lacks moral content (the action is not tracking anything moral), so too the simple end of going to the beach adopted without regard to obligatory ends is morally empty and therefore morally risky. It is of course not *very* risky when compared with the end of seeking enjoyment from crack cocaine; but if, for the agent acting, that difference makes no difference, then

---

[42] Since this is one of the regions where truths about the individual trump general claims about persons, failure to pursue characteristically healthy human goals is a serious warning sign, not necessarily a failure.

[43] Again, the extension of judgments of wrongness from action to volition is Kant's central point: if moral wrongness is about faulty reasoning, an action-centered notion of wrongness (or impermissibility) may play a pragmatic role, but it does not fully capture the nature of moral error.

from the moral point of view, that indifference *is* very risky indeed. It then might not seem so off to say that I am morally wrong in acting on my plan.[44]

Obligatory ends thus bring a wide range of ordinary human concerns inside morality. Although our ideas of happiness may have to undergo some revision and development in order to relocate, the familiar elements of self- and other-concern remain, and remain central to our purposes. In securing norms of regard for the well-being of self and other, obligatory ends make these considerations anchors for sound reasoning to action. It remains to be seen whether obligatory ends can justify actions that ordinary ends cannot. Can they show it is all right to lie or coerce or harm for their sake?

I don't mean to suggest that obligatory ends might be crucibles of moral alchemy, able to turn immoral actions into moral ones; if they provide broader justification, it is as premises that affect the moral content of the volitions that follow from them. This is true of other moral ends as well. When acting under the end of friendship, an otherwise permissible action that causes my friend concern may, for that reason, be wrong; or, given special facts of need and intimacy, some morally difficult avenues of action are opened (think about the space of jokes and teasing). The question about obligatory ends is not whether they affect morally available means (they do), but how we are to determine their justificatory scope. The obligatory end of others' happiness may justify some paternalism, but we don't expect it to justify killing one to benefit many (or any) – it can't transform the content of willing *that* action as a means. But perhaps it can reach to a lie for the sake of a life.

7. For guidance in thinking about how moral ends might justify suspect means, I am going to draw on a different region of argument where Kant clearly does appeal to a certain kind of moral end to show how something normally forbidden is permitted – indeed is morally necessary. The argument is about coercion into political union.[45] The formation and preservation of a state that meets rule-of-law

---

[44] If actions whose maxims have moral content exhibit good willing, then it is present when ordinary actions are done for the sake of an obligatory end. Needing a break from work, I decide to go to a movie. I could do it simply for pleasure; I could also be aware that such pleasures are part of a healthy life and act for that reason as well. Since in the latter case the reasoning is, in moral terms, both valid and sound, it does seem to be an instance of good willing (though not of moral worth, since the movie-going is not itself a dutiful action).

[45] What follows is drawn from the *Rechtslehre*, part one of the *Metaphysics of Morals* 6:252-261.

standards has a special role in Kant's moral theory since membership in such a state is a necessary condition of external freedom of action. Through its coercive and adjudicative institutions, the state secures the integrity of body from assault and makes possible sustained possession and exchange of property. Where these conditions are not met, the plurality of individuals' rational actions in and on the world, and so also their happiness, cannot be coherently pursued. Persons are therefore strongly obligated to form a state if there is none, and to sustain the one they have. The end is not one anyone can aim at alone; it is one, Kant argues, we can and must compel others to pursue with us. This sets the problem. If the argument is a moral one, it might seem that the compulsion should be forbidden. But then morality would appear to block its own real possibility. If, however, the argument is *not* moral, we would have a political and trumping contra-moral obligation. Together, these positions form what one might call an antinomy of obligation. As with any Kantian antinomy, it is best resolved by rethinking the assumptions that generate it. In this case, the problem derives from the characterization of the entitlement to compel co-citizenship: whether it can be shown consistent with the autonomy of the rational will.

It is a core feature of Kant's ethics that something's being good for you to do does not entitle me to coerce your doing it. Yet here, the fact that "we" (including you) must (for our good) live in a state apparently entitles us to compel entry and prevent exit (from some civil state or other). Now coercion is a matter of using force or threat of force to induce another to will against her will. But if what is at stake is putting persons and their actions under the authority of the state, it is not clear that the will is forced, or forced in a way that makes coercion morally objectionable. We might think of it this way. When the police set up a road-block, they have the authority to compel me to stop; regardless of what I prefer to be doing, their act is coercive, but not in a morally objectionable way. And that suggests the antinomy might also be about authority. It would be resolved if the authority to compel (into the state) is entailed by what each and every agent necessarily wills.

Kant makes just such an argument. In summary form, it goes like this. In taking possession of any object for our use, we necessarily will that others refrain from taking it (if I take the apple for eating, or plant a crop, I will that it be mine, not yours). Since we cannot live without taking possession of objects, and the condition of our effective willing that others refrain from taking what we have is, Kant argues, the state ("Only in the civil condition can something external be mine or yours"), then in taking possession of anything, we in effect will that condition, and so the state's authority, as a necessary means. In this sense the authority of the state over the will of each is willed by each, and willed by each on condition that it is reciprocally willed by others, which it necessarily is. Thus civil union, under law, arises in and through the reciprocal rational conditions of possession

(property). Since it is not an authority we can rationally avoid, in being materially compelled to act in accordance with the authority of our own will, we are not wrongfully coerced.

Resolving the antinomy in this way keeps morality from blocking its expression in the world, and it does so in a way that explains why one may not do just anything to compel civil entry (or forbid exit). Since the condition of being under law is a moral status, the terms of being brought under the state's authority have to be compatible with one's standing as an equal citizen. This then explains why territorial expansion through war and colonization is impermissible. As Kant remarks, the conditions of civil union arise from the conditions of living together; a state has no authority to create the conditions artificially.

So rather than being a disturbing embrace of contra-moral action, compelled civil union provides an example of how something that has the look of justified wrongdoing turns out to be uncompromisingly moral. Moreover, although compelled citizenship is justified with respect to a moral purpose – securing the conditions of free action – it is not justified simply as a means-to-an-end, even a moral end, but as an action on a principle with moral content (as a kind of moral self-actualization). This justification then frames and shapes subsequent moral reasoning.

The sort of thing I have in mind is this. If the state is morally justified, a variety of roles that are necessary to its function will also be (police, legislator, judge, soldier, but also doctor, educator, welfare-provider). When inhabiting these roles, individuals are allowed to act in ways they would not be permitted to act in the service of their own ends (e.g., police use of coercive force; a hospital policy of triage). We can say: the roles constitute ends of reasoning, so that actions that flow from social roles are, morally speaking, not the same actions they would be if derived from private ends. Of course the justificatory reach of institutions is limited: some actions that might instrumentally promote the function of a social role are not consistent with or would undermine the moral rationale for the institution's sphere of permission. Public officials are permitted, even required, to use force to gain compliance with the law, but they may not use bribes as a means to the same purpose. The impartial use of force is a condition of free action and so is consistent with the moral purpose of the state; bribery by public officials undermines the rule of law conditions of cooperation that a state's existence is to make possible.

There is much more to be drawn from the argument for compelled civil union, but with respect to understanding the work of obligatory ends, two things are most useful: one is the way in which the value a morally necessary end represents enters reasoning about means, and the other is the idea of a common end.

The first draws on basic facts about ends and means. We act by taking means for ends; we reason from ends to means. If the ends we reasoned from were desired states of affairs, then the reasoning would be familiarly instrumental. Further checks on such reasoning tend to be lateral, about costs to other ends one is or will be seeking. The moral check on purely instrumental reasoning is on means *simpliciter*: were one to know nothing more about the end than that it is desired, we should ask: can it permissibly be brought about *this* way? The obligatory end has additional effects on downstream reasoning in at least two ways. First, because an obligatory end, or an instance of such an end, has moral content, in acting under its auspices we are to conceive of what we would do as both morally and causally sufficient for the end. One can't act for the end of "helping persons in need" and take as one's means impoverishing Peter to aid Paul.[46] Likewise, in taking on the moral project of making oneself more focused and attentive to detail, a regimen that caused near-obsessive behavior in this regard, while effective in one sense, would undermine the value of an end which was about the enhancement of abilities of discernment and judgment. In general, the effect of the moral content of an obligatory end *narrows* the class of otherwise permissible instrumental means by requiring that they be (and be seen to be) consistent with the value the end represents.

Sometimes, however, the effect of obligatory ends on moral reasoning is a potential *widening* of the range of means, allowing us, maybe even directing us, to do things we otherwise could not. Following the lesson of compelled civil union, we will not see the moral end as having a kind of weight that private ends lack, its value simply overriding whatever consideration opposes the questionable action. Nor will these actions be justified in spare instrumental terms. Rather, the value in the morally necessary end supports reasoning to an action-type that is only externally congruent with forbidden action. In the case of compelled civil union, what looked like a brute exercise of force turned out to be an action that all are rationally required to will. It is coercion, but not wrongful coercion.

The shift in moral valence that comes with the detail of ends can be seen in more ordinary examples. Compare the situation in which your child is drowning in your pool and I can save him only by, without asking, using your life-preserver, and the situation where it's my child in my pool and I must take your life-preserver. Let's assume I am justified in using what is not my own in both cases. But the actions are not the same kind. In the first case I would say I act for you, using what is yours as an extension of your agency, so that my taking is justified by what you are obligated to will. In the second case there is the balance of harms, the

---

[46] Assuming the impoverishing is not by way of an impermissible act, this not only could be but arguably is a variant of a possible law of nature.

reasonable imposition of burdens, an occasion for replacement and apology, none of which makes sense in the first case. What sense could there be in *apologizing* for using your stuff to save your child?

I don't insist on this way of describing these cases, only that it is a possible way to think about them, and a natural one, once we allow the idea that there is more going on than causal fit when reasoning from end to means from a morally required end. We are not asking, "May I take this means to my end?" but, "Does this end-means pair satisfy the full moral conditions on willing?" In the terms of our earlier discussion (in section 4), it is an instance of motive, reflected in an end, affecting the (moral) identity of an action.

The second lesson to be taken from the argument for political union concerns its being a *common end*: there is something necessary for each of us to do that none can effect without others having and acting for the end as well. Obligatory ends are also common ends, though not for the same reason. Because they are ends of practical reason, each of us has a duty to adopt them. But the fact that they are ends I am under obligation to have does not make what they require my project in more than a locating sense.[47] We are all rationally required to acknowledge and adopt the obligatory end of helping others or promoting their rational well-being. That here and now it's me who must help is only indirectly of moral significance. I respond to (what we would call) an impersonal reason, based in what I rightly regard as a non-optional end. (By contrast, where ends are private ends, that they are mine is not just a matter of location: they belong to me.) So there is a sense in which, like the case of compelled civil union, obligatory ends give us a common project; but unlike the case of compelled civil union, we can, indeed we must be prepared to take on parts of the project separately.[48] For an end to be a common end it need not also be a cooperative one.

8. Armed with these features of obligatory ends – their effect on the moral content of means, the widening of the range of options, and the idea of a common end – we are in a position to make some progress with the kind of case that Kant is thought to manage so badly, where morality seems to require us to act without regard to consequences that we have compelling reason to avoid or prevent. I will focus on

---

[47] This is a point Thomas Nagel made in *The Possibility of Altruism* and John Rawls took up into his reading of Kant.

[48] Even when we act in concert, say through charitable organizations, it is our individual obligations that are being met, though more efficiently, through shared efforts.

the "murderer at the door" scenario, largely because of its unfortunate fame, but also because in working through it we gain some insight about why truth-telling is so important to Kant's deliberation-centered ethics.[49]

Let's set the stage in the usual way. Confronted by a murderer demanding information concerning the whereabouts your friend, his intended victim, you think you should lie to prevent the murderer's succeeding. It is natural to regard your lie as a means to misdirect the murderer and save your friend's life. This, of course, is what Kant objects to: that your purpose in lying is to provide a benefit (or avert a harm) does not make it not wrong to do. Though the principle seems true enough in the abstract, its application to this case strikes almost everyone as absurd: if the lie here is wrong at all, that wrong is surely outweighed by the greater wrong it prevents. Kant seems unable to accept this because of the great disvalue he accords lying to promote one's ends. Whereas we are not sure that the lie in these circumstances is wrong at all.

We can't finesse the issue by arguing directly from the end or even the duty of saving a life. Saving a life is not in general a morally trumping aim (we can't maim or torture in order to save); whether it is ever a trumping aim is the question. We do no better arguing from preventing wrongdoing or a wrongful harm when there is no set calculus for balancing wrongs. Indeed, the issue won't even be raised properly unless we come to terms with Kant's views about the moral significance of lying. The best place to begin, then , is with the specific objection to lying for the sake of one's own ends. We will later consider whether and how lying (and truth-telling) might be affected by obligatory ends: that is, whether and how the kind of end in question makes a moral difference on what may be done. The route we'll follow will take us through less familiar territory about speech and reasoning, the normative import of ends, and the moral significance of different ways of preventing wrongdoing.

So why might Kant have such intense concern with speaking the truth? We start with the fact that normal communicative speech carries a truth presumption: absent good reason to believe otherwise, we have warrant to accept what is said as true (or believed to be true by the speaker), and within limits, are right to depend on it. Whatever the source of the truth presumption – be it in reason, the logic or grammar of assertion, or the conditions of trust – it is clearly in the extension of both

---

[49] Allen Wood gives good reasons for thinking we have grossly misread Kant's "Supposed Right to Lie" (in chapter 14 of his *Kantian Ethics* (Cambridge University Press, 2008)). Here I start out with the old assumptions, though my conclusion fits better with Wood's, and indeed, with other of Kant's discussions about lying. For the record, Kant does not hold that lying is always wrong.

obligatory ends. Since, for Kant, correct reasoning in general ultimately depends on our being able to reason together, the obligatory ends' requirement that we attend to the conditions of rational agency in ourselves and others makes the truth presumption a central concern of a common end. In these terms we should say that the wrong in instrumental lying arises from a deceptive employment of the invitation to believe carried by ordinary speech, reliance on which is exploited to make the victim's reasoning conform to a purpose that is not her own (or, more precisely, not her own in the right way).

If the truth presumption is essential to the well-functioning of rational agents, by-passing it, even for a good end, would seem to involve the kind of insult to persons' status as rational agents that morality prohibits. That suggests that the natural question to ask about the forced speech situation created by the murderer is whether it somehow voids the presumption. Kant made a debater's objection to the claim that the murderer had no *right* to the truth, but it's not obvious that he had to reject the idea that in some speech conditions the truth presumption might be canceled.

There are, after all, all sorts of occasions in which we indicate that our false speech should not be taken as a lie: when we tell tall-tales, or make jokes, bluff in games, write fiction, perform political satire, and so on. Social conventions mark out arenas for white lies and tactful omissions. In the would-be murderer's case, we might argue that because the speech is compelled, or would abet wrongdoing, the context of action itself signals that the truth-presumption is suspended.[50] However, unlike jokes and tall tales where we know the speech is not intended to be truthful (or where truthfulness is not its point), or conventions of tact with which all or most are familiar, in this case one of the parties, the aggressor, depends in his reasoning on the fact that the truth presumption is in play with its usual force.

In ordinary circumstances, whether or not we like the way someone would act, whether it is for or against our interests, autonomy demands respect for a person's agency and for its expression in reasoning to action. We may not undermine another's reasoning for the sake of our own ends by introducing false beliefs or misleading truths, or even by making so much noise that she cannot think. Out of respect, we may decide not to correct errors, or limit our interventions to advice. Sometimes this is because we are not certain what the agent intends, but often even when we are, we accept the authority each has to put the elements of a life together her own way. Though one person may know more than another (in general or in one case) or deliberate with greater facility, no one has privileged access to

---

[50] Kant says we cannot impute the harmful consequences of rightful action to the agent; might this change when a rightful action abets wrongdoing?

correctness in moral reasoning (moral error is not typically a result of difference in skill or epistemic position). In that sense, we have equal status as reasoners.

On the other hand, not every course of reasoning warrants respect. The aggressor's reasoning is not just faulty; it issues in a demand on *our* speech that contravenes the core value of truthfulness, and betrays the common end of reasoning well. It is part of normal speech conditions that we use one another's truthful speech for our own purposes, regardless of whether the speaker knows what our purposes are or agrees with them. Here, however, the aggressor seeks our speech in the spirit of commandeering a weapon. He would impress our speech into the service of a contra-moral purpose – one to which there is no sound deliberative route. For that reason his demand cancels, or has no claim on, the truth presumption. Our being released from a requirement of truthful speech does not, however, get us all the way to the lie. It is because, in addition to the betrayal of the truth presumption, the aim of the aggressor's unimpeded faulty reasoning is harm to another, that we have reason not to let the situation take its course. That is what makes the defensive lie a real option.

Note that if it turns out that we may lie to resist the impressment of our speech, the *first* purpose of our intervention would not be protection of the victim, but something like preventive policing of our shared moral space in response to the aggressor's betrayal of the common end. Consider a case where our forced speech will abet faulty reasoning that would, by lucky accident, produce a beneficial outcome; we would have the same basis of action against the forced speech, but good reason to let the situation take its course.

Of course, because the aggressor is no less a rational agent despite his wrongful action, he remains within the scope of morality. If he has a heart attack on our doorstep we have whatever obligation we ever have to call an ambulance (and to tell him the truth about the help that's available). Nor are we free to do just anything in the service of moral policing; its tools are subject to the same prohibitions as the actions it targets. That is why it matters that the defensive false speech not be like the altruistic lie, an attempt to redirect the aggressor (by exploiting the truth presumption as a means of taking control of his reasoning and action) for the good end of saving our friend's life.[51] But if the aggressor's own reasoning deforms the speech situation, suspending the truth presumption, we are not in the condition of the ordinary wrongful lie. Our false speech would impede him in reasoning through to his violent purpose, but it need not aim at hijacking his will, and therefore does not share the wrong of the ordinary lie. It's a lie, but perhaps not a wrongful lie.

---

[51] It is this assumption of authority over the course of the aggressor's reasoning that makes the altruistic liar partly responsible for any new risks.

Although it is the aggressor's creation of the forced speech situation that signals the change in presumption, if, without increasing the risk to the victim, we can manage without the lie, we should.[52] There is a point to being silent, or to speaking uninformatively, if one can. Such often-mocked casuistical maneuvers show respect for the truth presumption, and have the additional moral advantage of shifting the burden to the hearer, who bears responsibility for the morally compromised circumstances. Still, because the circumstances of action are not truth-demanding, and practical exigencies may leave little room for moral finesse, the straight lie may be without fault. Reasoned to from the common end, it honors rather than betrays what Kant calls "the supreme rightful condition in statements."[53]

We thus approach the conclusion that appropriately conceived false speech can be morally permitted, perhaps even required. It would be morally on a par with other kinds of prevention that impede the completion of bad reasoning in wrongful action. Harking back to the example of compelled civil union, we might draw on an analogy with the policing acts of the state whose justification is that they are "a hindering of a hindrance to freedom."

There is, however, an apparent disanalogy between the necessary end in the argument for compelled civil union and the status of the common end and so of our entitlement to address the malfeasor as its agent. In the argument for civil union, something that the agent necessarily wills (property) has civil union as its necessary (and so omnilateral) condition. But what would count here as such prior willing? While the truth presumption belongs to both obligatory ends (neither towards our own nor others' rational well-being can we be indifferent to the conditions of correct reasoning), an obligatory end is, *qua* end, an agent's end only if she adopts it. The truth presumption *is* necessary to communicative discourse, and so to (human) rational willing in general; it is not a necessary condition of speech as such and so not necessarily willed by all. What I think we should say is that since each of us has necessary and sufficient reason to adopt obligatory ends – that's part of what it

---

[52] I am indebted here to Collin O'Neil for his insightful work on the moral differences between the lie direct and the constrained misdirection in other forms of misleading speech (see his *The Ethics of Communication*, UCLA PhD Dissertation, 2007).

[53] "A Supposed Right to Lie" 8:429. This is as close as I can get to making sense of Kant's claim that we may not forego truthfulness in speech for the sake of some contingent purpose. I think it is in fact quite close.

is for an end to be obligatory – we are entitled to regard everyone, and so the malfeasor, *as if* he had the end: we impute it to him.[54]

Ends have three distinct normative roles. First, as our purposes, ends mark out targets of action; they are what we deliberate *from*. Unless an agent adopts an end, he cannot reason from it to action. Second, ends represent standards, or regulative rules for action: reason itself gives a regulative end, imposing norms of consistency, order, and justification. And third, ends indicate the kinds of reasons agents can offer that shape acceptable interactions. Normally, when someone says 'no' to an end, he has reasons that warrant our respecting his decision. But someone who refuses to adopt the end of helping others isn't thereby free from moral criticism. And the murderer-at-the-door has no reasons for refusing the common end that should concern us. In imputing the common end, we engage the second and third normative elements, and take them to warrant our acting *as if* the first were true as well.

Imputing an end is not such a strange thing to do. Seeing a geyser of water erupting from the front of your house, I enter your property to shut off the main valve to prevent flooding, though I don't actually know you care to protect your house (you might be flooding it yourself to collect insurance, or turning your house into a performance piece). I act because it is reasonable to assume you do have the end it is ordinary to have in such circumstances. I regard myself as acting on your behalf, completing the reasoning to action you would make for your ends were you here. If there is a gap, it is an epistemic one. But when, as in the earlier case, I use what is yours to save your child from drowning in your pool, while again I don't know what you want or intend, the assumptions I make about your end are not bridging an epistemic gap; there is no gap. It is not just reasonable to assume you have the end, it is an end you (morally) *must* have. We are warranted in imputing it to you. If, in turning off your water in the first case, I've made a mistake about your ends, I should apologize. It's a reasonable error. Perhaps you ought to have warned me that you were doing something so unusual.[55] There's no such mistake (or warning) possible in the drowning case: your wanting

---

[54] The doctrine of imputation in Kant's *Metaphysics of Morals* (6:227-228) is about actions and their consequences, not ends. I extend the use of this term because with the idea of imputing ends I want to argue that certain moral conditions explain when it is right to say that a standard applies to a will. Imputing other content derived from obligatory ends is possible, but isn't relevant here to the justification of the defensive lie.

[55] The more we see morality, or parts of it, as a common project, the more responsibility we have for recognizing and flagging special contexts.

the insurance from your child's accidental death does not introduce any reasons that need to be overcome when I act.

We sometimes impute an end as a way of making sense of someone's practical reasoning (as the best account of what is affecting or shaping her reasoning). We also impute hidden motives and unacknowledged ambitions; we impute meaning to speech that is not entirely from the agent, but belongs to context or a dominant ideology (in some circumstances, we impute insensitivity at the telling of a tasteless joke).[56] We hold people in various jobs and offices to standards, criticizing them for failure, without regard to their volitional commitment: that is, given a role, we may impute ends. Imputed ends are one way of explaining what entitles us to integrity in a banker, or to reasonable care in a technician handling our x-rays – regardless of what they in fact will, we are right to complain about the person when the integrity or the care are absent.[57] This is not to say that imputed ends are just as good as the real thing. Where an end is merely imputed, an agent can fail, or reason badly, but unless it actually is her end, she cannot reason well.

So I think we may properly impute the common end to the murderer. He has no reason, in this case, or in general, that would defeat the imputation: he has sufficient reason to adopt the end, and no good reason not to. We therefore do him no disrespect as a reasoner in acting towards him as if he shared the end.

Now, one thing I have not discussed about obligatory ends is the fact that the duties they give rise to are imperfect. Although the obligatory end that directs us to the (rational) well-being of others implies that no one's well-being can in principle be a matter of indifference to us, because the duty is imperfect, we each act for the end in different ways, on different occasions. Imperfect duties introduce a kind of division of labor – each of us has a role, set by our location, our relationships, and our resources, in the service of the end.[58] It might then seem that little can be said in advance about how we are each to act for the sake of the common end, and that suggests that the *imputed* end is idle: it could never be the basis of (even

---

[56] Judgments of negligence often involve imputation, but what is imputed is knowledge of a morally relevant action-guiding fact; the end is not in dispute.

[57] Although an agent may not embrace the standard for action, we can say it belongs to her in the imputed sense, and her reasoning to action is subject to criticism if it is not consistent with what follows from the imputed end. With respect to an end that is merely imputed, an agent can fail, or reason badly, but unless it is her end, she cannot reason well.

[58] In this way obligatory ends shape all of our lives, but don't give all lives the same shape.

counterfactual) reasoning to action for the agent to whom it is only imputed. Indeed, it might seem hard even to mount criticism in its name.

But the common end is not idle. The same grounds that we have for imputing the end at all are sufficient to support a *general* duty of truthfulness in communication: the normal truth presumption that is a condition of human rational well-being generally. (That is, any sound reason we may have to lie will not be when the conditions of the truth presumption apply.) That is why, whether or not it causes harm, the advantage lie is wrong: it misuses the presumption as a private source of power over others. In the case at hand, the misuse occurs in the creation of a context of forced speech, when nothing the speaker can say in response is consistent with moral ends. So the standard of the common end applies. As agents of the common end we are then warranted in intervening for its sake; our targeted deception is a reassertion of its authority.

This gives us a rather distinctive account of what the justified lie accomplishes. The malfeasor is prevented from acting contrary to the conditions of an imputed end – not an end he has, but an end that we are entitled to use as a standard of judgment for his reasoning. More specifically, the intervention targets an inappropriate chain of reasoning that gives rise to an illegitimate demand on shared conditions of speech – illegitimate from both parties' points of view (one actual, one imputed). The false speech does not force the faulty reasoner into conformity with good reasoning (again, *he* has to reason correctly for that); nor can it bring him to act in light of the imputed obligatory end he has failed to adopt. The aim of the speech is to create an impediment to the completion of his reasoning. The impeding shows no disrespect, either for the reasoning or for the reasoner, because the malfeasor has no reason for what he would do that can be respected. If he were sliding on ice towards danger to himself or harm to another, I could respectfully impede his progress. In this case, we would impede an attempt to cross a boundary of protection for truthful speech.

Clearly, this is a narrow result. It does not show that we can lie to prevent harms. It does not show that there is an exception to a truth-telling principle for the sake of protecting a life. It does not justify an exception to a rule against lying. What we learn is that an end or value that normally calls for truth-telling (making it our default position) in this context does not: the factual premises in the case involve a misuse of the truth presumption that then alters the deliberative outcome. The value content of an obligatory end works down the chain of reasoning to permit or require resistance to the misuse of the truth conditions of speech. It thereby tells us how we are to understand and so justify this lie.

If the forced speech feature is absent, the reasoning to an intervention would perforce be different. Suppose one is not compelled to speak; may one volunteer a lie with the aim of sending murderer elsewhere? Since in such a case one makes

use of the truth presumption as a means to exercise power over others, then no. (Thus the claim that once one uses a lie to orchestrate events, one assumes some responsibility for bad outcomes to which the lie contributes. No such shift occurs in the forced speech situation.) An altruistic lie is not morally different than altruistic acts that involve physical detention, constraint or injury. Nothing in the content of the obligatory end yields permission to exercise intrusive power over another. With the justified lie to the murderer, by contrast, the agent acts, as he always should, as an agent of the common end, his targeted false speech a reassertion of its authority.

Reasoning from obligatory ends we *can* have moral cause to make someone's deliberation and so his action more difficult. We tell him that his action will impair our friendship in the hope that this fact will affect his deliberations, not just as a disincentive but as cause to rethink. We can stagger information in the hope that having to wait will create an occasion for clearer-headed deliberation. We prevaricate. The aim is to keep things open and avert danger; as a private agent, we are not entitled to seize another and author his future.

But suppose it all goes wrong. To stop the murderer we would have to disable, confine, or hurt him. What I would say, though can't argue for it here, is this. If an agent of the state – the police, for example – could intervene with force, we may also. Not, however, as private agents pursuing good ends, but as surrogates for public authority when it is not available, and for public ends (we would act to disable the aggressor for the sake of public order). The model is the citizen's arrest, where force is used, but not immoral means. A private agent who uses force does something wrong because there is no valid route from the moral content of his good end to the use of force; the public action, however, has its source in the work of the state which allows for the use of force. Though externally the same, the public and the private actions have different moral content. The rejoinder that the private use of force cannot be impermissible because no one could have good reason to prevent the intervention mistakes other public reasons (a prosecutor's discretion, for example) for moral justification. But these are difficult matters, and for another time.

The purpose of engaging in this lengthy casuistical exercise was to illustrate what can happen when we have obligatory ends at the head of a chain of reasoning to action: a wider range of means is morally allowed, even some we would have thought were ruled out, and consequences are shown to count without ceding ground to moral instrumentalism. Until the casuistry is more fully elaborated, we won't know whether the route through obligatory ends offers enough to accommodate the moral intuitions that Kantian theory has seemed to ignore. But even this fragment of an account is rich enough in resources to encourage the project of a unified (non-hybrid) interpretation of Kant's ethics.

HOW I AM NOT A KANTIAN [59] T. M. SCANLON

On What Matters begins with a vigorous defense of a cognitivist and value-based account of reasons. It ends with a striking claim of a convergence between Kantian, Consequentialist and Contractualist moral theories. In these comments I will concentrate on the relation between these two parts of Parfit's rich and provocative book.

Questions about reasons are fundamental to Parfit's conclusion because the theories whose convergence is in question all characterize right and wrong in terms of what people have reason to want, or could rationally do. The three theories Parfit is considering are:

> *The Kantian Contractualist Formula*: Everyone ought to follow the principles whose universal acceptance everyone could rationally will.
>
> *Scanlon's Formula*: An act is wrong if it would be disallowed by any principle that no one could reasonably reject.
>
> *Kantian Rule Consequentialism*: Everyone ought to follow the principles that are optimific, because these are the only principles that everyone could rationally will to be universal laws.

Parfit acknowledges that the two theories he labels "Kantian" diverge from what Kant himself said. But he regards this as no objection to what he is doing. "We are asking," he writes, "whether Kant's ideas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise Kant's formulas in a way that improves them, we are developing a Kantian moral theory" (p. 000).

I agree that it can be a valuable project to develop a moral theory that is similar to Kant's in some ways but departs from it in others. But I believe that one of the ways in which the theories Parfit lays out diverge from Kant's own view deserves attention. The degree to which Parfit's conclusion should seem surprising depends to a certain extent on how close the theories he is discussing are to Kant's. More important, an examination of one way in which these theories differ from Kant's will bring out some of the difficulties faced by an account of reasons of the

---

[59] I am grateful to Derek Parfit for many discussions of these issues as well as for helpful comments on an earlier version of this paper.

kind that Parfit and I favor, and hence also by a moral theory based on such an account.

I will not engage in detailed exegesis of Kant's texts, but will base my discussion of these issues on a few broad claims about Kant's view of rationality and morality which I hope are relatively uncontroversial. For simplicity, I will concentrate on Kant's Formula of Universal Law, and on Kant's discussion of this formula in his *Groundwork of the Metaphysics of Morals*. A full discussion would need to take into account other formulations of the Categorical Imperative as well as what Kant says in other works. But this will suffice for the mainly comparative points that I want to make.

I begin with an observation about the way in which Kant sees the Categorical Imperative as authoritative for us. What he says in Section 3 of the *Groundwork* is that when we are deciding what to do we must *see* the Categorical Imperative as our highest level principle of practical reasoning insofar as we see ourselves as acting at all. If we take any other principle to be fundamental for us, then we cannot see ourselves as acting but only as the slaves of factors acting on us. This claim depends in turn on Kant's argument, in Section 2 of the *Groundwork*, that there can be only one categorical imperative (that is, that any principle other than the one he has presented could influence an agent only though its appeal to his or her inclinations.) Thus, in Kant's view it is only if one takes the Categorical Imperative as the fundamental principle of practical reasoning that one can see oneself as *deciding* what to do rather than merely being determined by one's inclinations.

Turning now from the authority of the Categorical Imperative to its content, the Formula of Universal Law says that one should act only on a maxim that one could will to be a universal law. I believe that the best interpretation of

what Kant means by a maxim's being a universal law is for everyone to believe it to be permissible to act on that maxim, and to act on it when they are so inclined. The crucial questions in determining what this formula requires are thus: (1) what, in Kant's view, would prevent a maxim from even being a universal law in this sense, and (2) what would make it the case that a maxim could not be willed to be such a law.[60]

---

[60] Parfit discusses these questions in sections 40 and 41 respectively. My interpretations of these Kantian ideas differ slightly from his. The claim that it is wrong to act on a maxim that one could not rationally will to be a universal law in the sense I have just described is similar to what Parfit calls the *Law of Nature Formula* except that it substitutes for the phrase "and acts on it when they can" the phrase "and acts on it when they are so inclined." My version of the claim differs from what Parfit calls the *Moral Belief Formula* because it requires one to be able

Kant's idea seems to be that a maxim "cannot be a universal law" in the sense he has in mind if the plan of action it describes would be incoherent in the event that people's attitudes were of the kind that this universal law describes. The "contradiction" that he is appealing to is thus between the presuppositions of the plan of action that the maxim describes and the conditions that would obtain if this maxim were a universal law. The most plausible example of this is Kant's case of the lying promise: making a promise would not be an effective way of getting the money one desires if everyone believed that having made such a promise was no constraint on anyone's future conduct. Parfit may be right that the terms "contradiction" and "cannot be a universal law" are not the best way to put this point. But I think it is reasonably clear what Kant has in mind.

Parfit's understanding of the idea of something's being rationally willed to be a universal law is different from Kant's as I interpret him. When Parfit asks, in interpreting the various formulae he discusses, whether an action or principle is one that someone could rationally will, he understands this as a question about the reasons that person has, and their relative strengths. One can rationally will something, on his view, if one has sufficient reason to do so; one cannot rationally will it if one's reasons not to will it are stronger than one's reasons to will it (p. 000). Kant's idea of what one can will is different. When he considers the question of whether a given maxim could or could not be willed to be a universal law Kant seems not to appeal at all, or at least not in a fundamental way, to reasons or their relative strength.[61] Indeed, the idea of a reason and of the strength of a reason have at most a derivative role in Kant's account of rational action and morality.[62]

When Kant says that a maxim could not be willed to be a universal law, what he means is that willing such a law (willing that everyone act on the maxim should he or she be so inclined and believe that others will do this as well) would be incompatible with viewing oneself as a rational agent. For example, Kant claims that a maxim of developing one's talents only insofar as one finds this pleasant or

---

to will not only that everyone believes it to be permissible to act on the maxim in question, but that they also act on it when they are so inclined.

[61] To act on a maxim is to act for a certain reason. So in asking whether one could will that people act on, or be permitted to act on a maxim, the idea of a reason for action figures in what one is asking *about*. What I am saying is that for Kant such questions are not to be *answered* by appeal to the reasons an agent has.

[62] In an earlier version of the manuscript that became this book, Parfit expressed surprise that Kant seemed not to employ the idea of a reason in the normative sense in which Parfit understands it. My point here is that this observation was correct in a way, but less surprising than it might at first appear.

attractive, or a maxim of helping others only if it happens to please one, could not be willed to be universal laws, because in willing these laws one would be willing that one give, and that others give, no intrinsic weight to the existence of general conditions that are necessary to the pursuit of our ends. To be a rational agent, however, is to have ends, and one cannot (without being irrational) have ends yet be indifferent to the conditions necessary for their pursuit. The "contradiction" that Kant has in mind is thus grounded in the same thing that (as I maintained earlier) Kant believes grounds the authority of the Categorical Imperative itself, namely the views one must take insofar as one sees oneself as a rational agent.

Kant's claims about what the Formula of Universal law requires are thus not based on claims about what reasons individuals have, or about the relative strength of these reasons. When his claim is that a certain maxim could not *be* a universal law (as in the case of the lying promise), the question of what one can will does not even arise. When his claim is that we cannot *will* a maxim to be a universal law (such as a maxim of indifference to the development of our talents, or to the needs of others), his claim is not that the reasons we have not to will such laws are stronger than those in favor of doing so. What Kant says is rather that insofar as we see ourselves as rational agents we cannot see the development of our talents or the needs of others as considerations that in themselves count for nothing. The claims that provide the basis for Kant's arguments are claims about rationality—about the attitudes we must hold insofar as we are not irrational—not claims about the reasons we have.[63] Accordingly, the *conclusions* of these arguments are also claims that we must, insofar as we are not irrational, *see* these things—the development of our talents and the needs of others—as providing reasons for action rather than substantive claims about the reasons we have.

I should note, however, that as I have interpreted Kant's arguments about what one can will to be a universal law, their conclusions make only the most minimal claim about the strength we must see certain considerations as having. The claim is just that we cannot take these considerations—the development of our own talents and the needs of others—as counting for nothing (apart from their appeal to our inclinations.) If this interpretation is correct, and this minimal conclusion is all that Kant's argument yields, then it is left up to each person to determine (depending, I suppose, on his or her inclinations) how much weight to give to these considerations. But perhaps Kant's argument actually yields a stronger

---

[63]I discuss this distinction further in "Reasons: A Puzzling Duality?" in R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith, eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (New York: Oxford University Press, 2004), pp. 231-246, and in "Structural Irrationality," in Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith, eds., *Common Minds: Essays in Honor of Philip Pettit*, (Oxford: Oxford University Press, 2007).

conclusion. Perhaps Kant could establish that a person who sees him or herself as a rational agent cannot consistently will a maxim of not helping others or doing what is required to develop his or her talents when these aims come into conflict with certain considerations of convenience or comfort.

It might seem that in order to establish such a conclusion Kant would have to appeal to premises about the relative strength of reasons: that is, it would have to rest on a claim that the possibility of enjoying the forms of convenience or comfort in question is not a sufficient reason for failing to develop one's talents in certain ways, or for failing to aid someone else in a certain way. But from the Kantian point of view as I am interpreting it this would be to get things backwards. Claims about reasons (more exactly, about what a person must see as reasons) must be grounded in claims about rational agency, claims about what attitudes a person can take, consistent with seeing herself as a rational agent. Justification never runs in the other direction, from claims about reasons to claims about what rationality requires.

This view, which I will call Kantian constructivism about reasons, seems to me to be a fundamental feature of Kantian ethical theories, distinguishing them from other views that resemble Kant's in some ways. In particular, as I have said, it distinguishes Kant's view from all of the moral views that Parfit discusses in Part Three of *On What Matters*. All of these views, including those described as Kantian, appeal to an idea of "what one can rationally will" that presupposes an independently understandable notion of the reasons that a person has and their relative strength. So there is one sense in which none of these views is Kantian: none of them accepts Kantian constructivism about reasons. This divergence raises questions facing in two directions. Negatively, why *not* accept Kantian constructivism about reasons? Positively, what can be said in defense of the alternative conception of reasons that Parfit employs, and that I myself would also favor?

On the negative side, Parfit raises objections to what he calls Kant's Impossibility Formula, according to which it is wrong to act on maxims that could not even *be* universal laws.[64] These objections mainly take the form of arguments that Kant's remarks about what could not be a universal law cannot be interpreted in a way that avoids intuitively implausible implications about moral right and wrong. I agree with many of the points Parfit makes here, although I would put them in a somewhat different way.

---

[64] See, for example, p. 228.

The "contradiction in conception" test[65] is intuitively appealing because it seems to capture the idea that it is wrong to exempt oneself from the moral requirements that apply to everyone else. Many wrongs do fit this pattern: if certain constraints are needed to provide some essential public good (or to prevent some serious "public bad"), and people are generally complying with them, then it is wrong to free ride on their compliance by exempting oneself from these constraints. But Kant's test does not track this idea in a reliable way.

The class of actions that Kant's test captures are ones in which an agent's plan of action presupposes that others believe that everyone is bound by constraints that rule out action of the kind that the agent is going to perform. The problem is that by focusing on the relation between an agent's action and what that action presupposes about the beliefs and intentions of others this test bypasses the question of whether the constraints in question are indeed justified. (This may be part of the appeal of Kant's test: it seems to provide a criterion of wrongness that can be applied without asking messy questions about the relative strength of reasons.) But the question of justification is essential. If the constraint that others take to be binding is in fact groundless (a mere taboo, for example) then it may not be wrong to violate this constraint, even if the success of one's action depends on the fact that most others take that constraint seriously. On the other hand, when constraints are necessary and justified, then it is wrong to violate them whether or not the success of *this very action* depends on the fact that others take these constraints to be binding and generally observe them. Everything depends on the need for the constraints in question, not merely on whether the success of one's action depends on their being generally observed.

What is commonly called Kant's "contradiction in the will" test might be called upon to answer this question of justification. The idea would be that to determine whether a constraint is justified we should ask whether one could will that it be generally believed to be permissible to violate this constraint when this suits one's purposes. As Parfit says, this criterion of justifiability is similar to the version of contractualism that I myself have proposed.

One way in which Kant's criterion appears to differ from mine, and Parfit's, is in focusing simply on whether *the agent* could will a principle permitting what he or she proposes to do, rather than on whether there is anyone who could reasonably reject a principle permitting such actions, or whether everyone could will the universal acceptance of such a principle. The question here is how a mode of thinking about right and wrong is to be sensitive to the interests of other people. Different theories solve this problem in different ways.

---

[65] Parfit refers to this test as "Kant's actual version of his Impossibility Formula" (p. 161.)

I believe that on the best interpretation of the way Kant understands his Formula of Universal Law, when we ask whether an agent could will his maxim to be a universal law what we are asking is whether he could will that people be universally permitted to act on such a maxim, where this universality includes situations in which the agent occupies any of the positions involved—for example, situations in which the agent is a person in need of help as well as ones in which he or she is the one called upon to give it. Assuming that this idea is intelligible, and that if the agent were in one of these other positions he or she would have the same reasons as a person who is actually in that position, this test would seem to lead to the same result as asking, as Parfit suggests, whether *everyone* could will this universal permission. Even if this is so, however, I agree with Parfit that it makes things clearer to avoid counterfactuals about the agent's being in different positions and to keep clearly in view the fact that we are dealing with different persons, by asking what everyone in these other positions could will, or could reasonably reject.

Another possible divergence from Kant arises when we consider how the idea of what someone could rationally will is to be understood. One might object to Kant's account of this idea on the ground that its implications about the reasons we have are inadequate or implausible. I have mentioned two objections of this kind. The first is that Kant's account yields only conclusions about what individuals must see as reasons, insofar as they are not irrational. It seems to me, however, that there are true substantive claims about the reasons we have that are different from claims of this kind and cannot be derived from them. Second, leaving aside the difference between these two kinds of claims, I do not believe that the idea of rational agency is rich enough to yield all the claims about reasons that seem evidently correct.

Going beyond objections of this kind, however, if we are going to reject Kant's account we need to consider the deeper question of where his argument for the Categorical Imperative as the limiting ground of the reasons we have goes wrong, if it does go wrong. Here I would cite Kant's claim that accepting the Categorical Imperative as one's highest level principle of practical reasoning is the only way in which one can see oneself as acting independent of inclination. This claim strikes me as untenable. I do not see why an agent cannot see him or herself as "active" in making judgments about which considerations constitute reasons.[66]

---

[66] It might be suggested that one can avoid these problems, and also provide the basis for a more extensive set of reasons, by appealing to Kant's Formula of Humanity—that is, to the idea that each person must regard his or her own rational nature (and that of others as well) as an end in itself. I do not believe that this line of argument is any more successful than the one I have sketched, but it would take me too far afield to examine it here.

Kant offers a top-down conception of reasons (or at least of our states of taking things to be reasons.) In his view, claims about reasons are grounded in the requirements of rational agency. If this account is rejected, the alternative might seem to be a "bottom up" conception, according to which practical reasoning begins with claims about particular reasons and their relative strengths and proceeds "upward" from there to conclusions about what we have most reasons to do or to think, taking all the relevant reasons into account. A desire-based theory of reasons for action would at least appear to be of this form. Such a view holds that if doing X would promote the satisfaction of some desire that an agent has, then that agent has at least a *pro tanto* reason to do X. What an agent has most reason to do all things considered is determined by balancing these various, and possibly conflicting, reasons.

Parfit considers and rejects desire-based theories in his Chapters 3 and 4. What provides us with reasons for action, he says, are not desires but the various facts about certain aims and acts that make them relevantly good, or worth achieving. Reasons are provided by considerations such as the fact that doing X would injure someone, or would save someone's life. This seems right to me. But when we focus simply on such considerations, considered individually, as ultimate reason-providers, a bottom-up view can be made to seem implausible. Do we really want to claim, it might be asked, that such considerations, in addition to their physical and psychological properties can have the additional normative property of providing a reason of a certain strength, and that the basis of practical reasoning lies in detecting these properties? Put in this way, this does seem odd. But the oddness results, I believe, from the fact that this way of putting things ignores several crucial aspects of reasons.

One thing that seems odd about this atomistic formulation is that it leaves out the relational character of reasons, and their dependence on context. A certain consideration does not provide a reason of a certain sort, full stop. It provides a reason for an agent, in a certain situation, to take a certain action, or to have a certain attitude. The same consideration can provide different reasons in this fuller sense depending on the agent, situation, and attitude involved. Similarly, the "strength" of a reason—that is to say, the way in which one consideration can override, undermine, or be overridden or undermined by other considerations—depends on the context within which a decision is being made.

A desire-based theory gains some of its plausibility from the fact that it has a certain relational structure built in. A desire is a desire *for* a certain content, but it is also the desire *of* a particular agent, a desire of a particular strength, and it provides reason for different actions depending on that agent's situation. One weakness of a desire-based theory is that the relational structure that it provides is too limited. Insofar as a desire is just a desire of a certain strength for a certain outcome, it provides reasons for actions that would promote that outcome. But not all reasons

are goal-directed in this way, and we have reasons for things other than actions. An adequate account of reasons needs to accommodate these facts.

The contrast with the atomistic realism I mentioned earlier brings out another feature of desire-based theories that should be noted, which is that their "bottom up" character is more apparent than real. Desires derive their reason giving force because they are the desires of some desiring agent. In this respect a desire-based theory is similar to the Kantian view, but it focuses of a different aspect of agency and, at least as I have formulated it, yields conclusions about the reasons that an agent has, rather than about what an agent must see as a reason insofar as he or she is rational.

But even if a desire-based theory offers a top down account of the source of reasons, its account of the process of practical reasoning remains bottom up: it sees practical reasoning as beginning with our experience of individual desires and their strength. An atomistic realism about reasons that preserved this bottom up character would share this implausibility. We do not experience considerations one by one as reasons with a certain strength. Rather, to regard one consideration is a stronger reason than another is to see it as more important *in regard to a certain type of decision in a certain context.* For example, whether the fact that it would be fun to make a certain remark counts as a strong reason for making it depends on the context, on what my aims and responsibilities are, and on my relation with the others present. Moreover, judgments about reasons and their importance are subject to requirements of consistency: if I judge A to be a reason for some action in one context, and a stronger reason than B, then I must judge this to be so in other contexts and for other agents as well, unless I can cite some relevant difference between these situations.

This discussion suggests several conclusions about what an adequate account of reasons must be like: It must preserve the idea that questions about reasons arise for, and are about, agents facing certain decisions. Second, it must be holistic in the way just described: judgments about particular reasons and their relative strengths depend on an overall view of the reasons we have. The strength of the Kantian view lies in its recognition of these important points. But an account of reasons must be substantive: it must include claims about the reasons that agents have, rather than merely about what they must see as reasons. And these claims cannot be derived solely from the agents' desires or from the mere fact that they are rational agents. If I am correct about this, then an adequate account of reasons will be a kind of substantive holism.

I turn now to Parfit's striking claim, in his Chapter 16, that Contractualism and Rule Consequentialism converge or, more exactly, that what he calls Kantian

Contractualism will coincide with Rule Consequentialism. I hope that an examination of his careful arguments will help to bring out what is distinctive about a Contractualist theory of the kind I have proposed, and how such a theory would differ from Rule Consequentialism even if the two were to support the same principles.

I will begin with what Parfit calls *The Kantian Contractualist Formula*:

> Everyone ought to follow the principles whose universal acceptance everyone could rationally will (p. 000).

As I have said, Parfit understands the question of what someone could rationally will as a question about what is supported by the overall balance of reasons that that person has. In his view, an agent can rationally will that certain principles be universally accepted just in case he or she has sufficient reason to will this. So the interpretation of the Kantian Contractualist Formula depends, as Parfit says, on claims about reasons and rationality. This formula will yield definite answers about what we ought to do in a given case only if there is a single principle (applicable to our situation) which everyone has sufficient reason to will to be universally accepted. Parfit calls this the uniqueness condition (p. 000) Given some views of the reasons a person has, this condition will not be fulfilled because there will be no principles that everyone has sufficient reason to will. Perhaps Rational Egoism is an example of such a view. [67]

Different moral theories deal with this problem in different ways. Rawls assumes that people will lack concern for how others fare (they will be "mutually disinterested"), but requires that they choose principles behind a veil of ignorance. My own version of contractualism deals with the problem by making particular stipulations about the reasons that are relevant to the choice of principles and the ways that these are to be considered.[68] The view that Parfit calls Kantian Contractualism makes neither of these moves. On this view, what we ought morally to do depends on what everyone could rationally will, with full information about their situation and taking into account all the reasons they in fact have. Parfit believes that the uniqueness condition is fulfilled "sufficiently often" (p. 000) because the reasons people have include impartial reasons as well as personal and partial ones.

Impartial reasons, he says, are reasons we see that we have when we consider matters from an impartial point of view—that is to say, without considering our own place in

---

[67] As Parfit argues (227-228). David Gauthier might disagree.

[68] Restricting these to what I call "personal reasons." See *What We Owe to Each Other*, 218-223.

a situation. We take such a view when, for example, we are, or suppose ourselves to be, merely an outside observer of what happens rather than one of the people whose well-being, or that of others to whom they have close ties, will be affected by it. Central among these impartial reasons are reasons to care about the well-being of others, but our impartial reasons may also include reasons to care about things other than individuals' welfare. Parfit argues that we have these same impartial reasons when we consider matters from our own personal perspective. (68) What the shift to the personal perspective does is merely to add personal and partial reasons to the impartial ones.[69]

A decision about what someone can rationally will must take all of these reasons into account. In some cases, the impartial reasons may predominate: one would not have sufficient reason to do something that would lead to the death of many people just to avoid scratching one's finger. In other cases the opposite will be true: one would not have sufficient reason to sacrifice one's life to prevent the scratching of one other person's finger (or, I would say, any number of persons' fingers.) But Parfit believes that there are many cases in which neither kind of reasons predominate in this way. In such cases, he writes,

> When one possible act would be impartially best, but some other act would be best either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either way (p.000).

Parfit believes that the uniqueness condition is fulfilled "sufficiently often" because there are certain principles that everyone has sufficient impartial reason to will to be universally accepted, even though they may have personal and partial reasons to prefer other principles.

Parfit defines the idea of "best outcome" in terms of the idea of impartial reason. We should call an outcome "best," he writes, just in case it is "the outcome that, from an impartial point of view, everyone would have most reason to want" (p. 000). He does not say very much about which outcomes will be best in the sense he defines. In particular, he leaves it open to what degree this idea of bestness will be aggregative: will an outcome containing a greater sum of well-being be better than one which contains less aggregate well-being no matter how well-being is distributed in the two situations? For

---

[69] This brings out the fact that the idea of a "point of view" is merely an expository device, a way of focusing our attention. Impartial reasons are not the reasons we *have* from a certain point of view. They are reasons we have *independent of* our particular relation to their objects, in contrast to personal reasons (to care about ourselves) or partial reasons (to care about others to whom we stand in certain special relations.) When we "take up the impartial point of view" we ignore these relations, and thus are aware only of reasons that do not depend on them.

example, will a situation in which greater total well-being count as better if this total is produced by significant costs to a few people which however bring small benefits to a very great number? As Parfit sets things up, this will depend on whether people have impartial reasons for favoring one of these states over the other. This leaves open the possibility that conception of best outcome he is defining is in important respects non-aggregative.

Using the notion of best outcome, Parfit defines universal acceptance rule consequentialism as the view that

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

He argues that this view is a direct consequence of

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

His argument for this proceeds as follows:[70]

> Kantians could argue:
>
> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.
>
> (B) Anyone could rationally choose whatever they would have sufficient reasons to choose.
>
> (C) There are some optimific principles whose universal acceptance would make things go best.
>
> (D) These are the principles that everyone would have the strongest impartial reasons to choose.
>
> (E) No one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons.
>
> Therefore
>
> (F) Everyone would have sufficient reasons to choose these optimific principles.
>
> (G) There are no other significantly non-optimific principles that everyone would have sufficient reasons to choose.

---

[70] On pp. 246-247.

Therefore

(H) It is only these optimific principles that everyone would have sufficient reasons to choose, and could therefore rationally choose.

Therefore

These are the principles that everyone ought to follow.

I do not dispute Parfit's conclusion about the relation between his Kantian Contractualism and Rule Consequentialism. What I want to concentrate on here is what this connection shows about the ways in which the structure of his Kantian Contractualism differs from the version of contractualism presented in my book.

Parfit says that according to Kantian Contractualism, in order to decide whether an action is permissible we must assess a principle that would permit it by conducting number of thought experiments, one for each person. In each of these we ask whether one of these persons could rationally will a principle that would permit such an action. This question is to be answered by considering both the person's personal and partial reasons and his or her impartial reasons. Suppose that the person's impartial reasons support accepting the principle. If the person has personal or partial reasons for not accepting the principle, the question we are to ask is whether, despite these reasons, the person nonetheless has sufficient reason to choose that everyone accept the principles that impartial reasons favor. As we have seen, Parfit holds that this might be true even if the person has sufficient reason to choose the principle that his or her personal and partial reasons favor.

According to my version of contractualism, deciding whether an action is right or wrong also involves a series of thought experiments. These consist in asking, in the case of each person considered, whether that person could reasonably reject a principle that would permit the action in question.[71] As in the previous case, suppose that one such person, call her *Green*, has personal reasons for rejecting the principle in question because of the burdens it would require her to bear. According to my version of contractualism, to decide whether Green could

---

[71] Parfit and I may take different views about the correct characterization of the "individuals" whose reasons are to be considered. Although he does not say so explicitly, some of what he does say suggests that he has in mind actual persons affected by the action, or by the acceptance of the principle. In my case what we consider are not the reasons of actual persons but the "generic" reasons that someone would have in virtue of occupying a certain role in regard to the principle in question, such as being the person who has relied on the assurance of others, or a person in need of help, or a person called upon to give it. I discuss this issue in *What We Owe to Each Other*, pp. 202-206.

reasonably reject the principle we need to consider the opposing reasons that others, considered individually, have for wanting the principle to be accepted. This involves a further series of thought experiments, corresponding to the various ways that people might be affected by the principle in question. In each case we are to ask whether, given the reasons that a person in the position in question would have for wanting the principle to be accepted, it would still be reasonable for J to reject it. The reasons that we consider here, in opposition to Green's personal reasons for rejecting the principle, *correspond* to reasons that Green would have if she took an impartial view of the situation, but there is a significant difference. In the form of contractualism that I have proposed, what we are to consider are not two kinds of reasons that Green might have (such as personal reasons and impartial ones) but, rather, the reasons that individuals in two different positions have: Green's reasons and those that a person would have who would be affected by the principle in a different way than Green would be.

The difference between these two ways of interpreting the reasons that someone might have for accepting a principle, or not rejecting it, can be illustrated by considering the way in which Parfit deals with a potential objection to his argument that Kantian Contractualism leads to Rule Consequentialism. Imagine a lifeboat case in which one is faced with the choice between saving five strangers and saving one's own child. Parfit believes that in such a case one would have decisive reason to save one's child. It may appear that optimific principles would require one to save the five strangers. If this were so then one might have decisive reason to reject these optimific principles, despite the impartial reasons in favor of willing their universal acceptance, contrary to premise (E) of Parfit's argument in the passage I have quoted above. Parfit responds as follows:

> The optimific principles would *not*, however, require you to save the strangers rather than your child. If everyone accepted and many people followed such a requirement, things would go in one way better, since more people's lives would be saved. But these good effects would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would save several strangers rather than one of our own children, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and some other similar facts, the optimific principles would often permit us, and often require us, to give some kinds of strong priority to our own children's well-being (p. 000).

This line of argument is familiar from the literature on consequentialism.[72] It has a distinctively consequentialist flavor because it appeals to what would be best overall—the

[72] See, for example, Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," *Philosophy & Public Affairs*13 (1984), pp. 134-171.

kind of outcome that everyone has most impartial reason to prefer. I make a similar point within my version of contractualism, but with an important difference.[73] Rather than appealing to the idea of the best outcome—what everyone has impartial reason to prefer—my argument was based on what each individual has reason to want for him or herself. A principle requiring us always to give the needs of strangers the same weight as those of friends and family members would be one that each of us could reasonably reject, because it would make impossible special relationships that we have strong reasons to want to have. Even if these two arguments lead to the same conclusion, and assign normative significance to the same facts about human life, they take these facts into account in different ways.

As I said above, according to my version of contractualism the considerations that we need to consider in order to decide whether it would be reasonable for J to reject a principle take the form of reasons that others would have to want that principle to be accepted. In Parfit's Kantian Contractualism these considerations enter in the form of impartial reasons that J has to want the principle to be accepted. But these are only some of the impartial reasons that could count in favor of Green's accepting the principle according to Parfit's Kantian Contractualism. Two differences are particularly significant. First, in addition to reasons corresponding to the reasons that other individuals have to want things to go better for them, Green's impartial reasons as Parfit would describe them can include impartial reasons that Green has for wanting more people to be benefited rather than fewer, or for the aggregate benefit to be as great as possible. According to the version of contractualism described in my book, however, what is to be taken into account in assessing the reasonableness of a person's rejecting a principle are only the reasons that *each* affected person has for wanting that principle to be accepted. Aggregative considerations are not directly relevant. Second, my view excluded impersonal reasons such as those associated with the value of natural objects or works of art, considered apart from the benefits to individuals of being able to experience these things. But impartial reasons as Parfit describes them could include reasons of this kind.

These two differences may be seen as improvements over the view stated in my book, which seemed implausible to many because it excluded aggregative arguments and because it gave no weight to impersonal values in determining what is right or wrong. These objections could be dealt with by allowing reasons of

---

[73] See *What We Owe to Each Other*, pp. 160-161.

these two kinds to be considered in determining whether a principle could be reasonably rejected.[74]

It is worth saying a little more here about the way in which the problem of aggregation is dealt with in Parfit's Kantian Contractualism, and therefore would be dealt with on this revised version of my view. The problem of aggregation is this. There are many cases in which what we should do, and even what it is permissible to do, seems to depend on the number of people who would be affected by the courses of action available to us. It seems that an adequate account of moral argument should make aggregative considerations relevant in these cases but do this in a way that does not support implausible aggregative arguments such as ones what would justify the killing or enslaving of a few people to make a huge number of people better off, each in a very small way.

Parfit's proposal, as I understand it, is to deal with this as a problem about which outcomes are indeed "best" (that is to say, ones that everyone has impartial reasons to prefer.) So he would say that in a case of the kind I have just considered the fact that aggregate well-being would be increased by enslaving a few people in order to benefit a great many people in small ways does not mean that a situation in which this was done would be one that we have impartial reason to prefer: the idea of "best outcome" is sensitive to numbers, but is not strictly aggregative. I leave aside the question of how such an account of impartial reasons and "best outcome" might be spelled out.

I have been discussing different views about the reasons that should be taken into account in deciding whether a principle is one that everyone could will to be universally accepted, or whether it is one that could reasonably be rejected. Let me turn now to the importance of the difference between these two ways of understanding the question we should ask in carrying out the thought experiments on which the rightness or wrongness of an action depends. According to Parfit's Kantian Contractualism one is to ask whether each person could rationally will that a principle permitting that action be universally accepted. On my view one is to ask whether every such principle would be one that someone could reasonably reject. How might the differences between these questions lead to different answers about which actions are right?

As we have seen, Parfit allows that there are many cases in which a person has sufficient impartial reasons to accept a principle but also sufficient self-interested reasons to refuse to do so. It seems possible that in some cases of this kind it

---

[74] Parfit has previously urged that I should make this change by giving up my "Individualist Restriction" on reasons for rejection. See his article "Justifiability to Each Person", in *On What We Owe To Each Other*, Philip Stratton-Lake, ed., (Blackwell, 2004), pages 67-8.

would be reasonable for the person to reject the principle in question. It might be that the universal acceptance of the principle would involve a cost that the person would have sufficient reason to accept (it would not be like a case of losing one's life because this would prevent the scratching of someone else's finger.) But this would also be a cost that a person could reasonably refuse to make. If there are cases of this kind, then Kantian Contractualism would involve higher costs than my version of contractualism would.

It will be helpful to divide possible cases into two types. In cases of the first type, although following the optimific principle would involve a major cost to someone, another person would suffer an even graver loss if the optimific principle were not followed. In cases of the second type this is not so: the sacrifice required of one person by the optimific principle is greater than the loss that any other individual would suffer if everyone were to follow some non-optimific principle.

Here is a possible case of the first type. Suppose that, in

> *Case One*, by giving some organ of his for transplant, *Grey* would be shortening his life by a few years. But by doing this he could give White, whom he does not know, many more years of life.

If this is so, then Grey would have sufficient impartial reason to donate the organ, and the outcome, if he were to do so, would be better in Parfit's impartial reason-involving sense. But Grey would also have sufficient self-interested reason not to make this donation. Moreover, it seems plausible to say that it would be reasonable for someone in Grey's position to reject a principle requiring this person to make such a donation.

Cases of the second type would involve two principles, P, which is optimific and imposes a high cost on people in the position of *Blue*, and Q which does not impose that high a cost on anyone (there is no one who would lose as much by a shift from universal acceptance of P to universal acceptance of Q as someone in Blue's position would gain from such a shift.) If P is optimific, and everyone has impartial reasons to prefer its universal acceptance to the universal acceptance of Q, this is most likely because the aggregate benefits to various people in P is accepted outweigh the costs to people in Blue's position. Perhaps Q would permit us to save Blue's life at the cost of failing to prevent a large number of people from being paralyzed, whereas P would require the opposite. Or perhaps P would require us to prevent many people from losing a leg rather than saving Blue's life, as Q would permit. In order to know which of these cases would fit the pattern I have described, one would have to know how Parfit's notions of impartial reasons and "best outcome" deal with aggregation. As I have said, this is not obvious. But presumably there will be some cases that fit the abstract pattern I have described.

These reflections have a bearing on Parfit's argument for the convergence of Rule Consequentialism and the two forms of Contractualism that he discusses. In this argument, he claims that everyone would have strong impartial reasons to choose that optimific principles be universally accepted, and that, because these reasons are not decisively outweighed by any conflicting reasons, everyone could rationally choose these principles. He then argues that, because there are no other significantly non-optimific principles that everyone could rationally choose, these optimific principles are the only ones whose universal acceptance everyone could rationally choose. When Parfit turns to my version of Contractualism, he then says that if certain optimific principles are the only ones whose universal acceptance everyone could rationally choose, this means that there are stronger objections to every other set of principles, and that if this is so then these optimific principles could not reasonably be rejected.

Suppose that optimific principles would require that we save many other people from smaller burdens rather than saving Blue's life. Though someone in Blue's position may have sufficient reasons to will the universal acceptance of these optimific principles, this person may also have sufficient reasons to will the acceptance of some non-optimific principle which would permit or require us to save Blue's life.

It might be that, taking only impartial reasons into account, everyone has stronger reason to will the acceptance of these optimific principles than to will the acceptance of some non-optimific principle that would require us to save Blue's life. This might also be put by saying that (considering only impartial reasons) there are "stronger objections" to this alternative than to the optimific principle. But taking *all* reasons into account, someone in Blue's position might have a stronger objection to the optimific principle that would impose such a sacrifice on Blue than anyone would have to some non-optimific principles that did not impose such a sacrifice. If this is correct, then the fact that these alternative principles are open to stronger (impartial) objections need not mean that they are open to decisive objections and hence need not entail that the optimific principles could not be reasonably rejected.

If what I have just said is correct, then shifting from the question "could anyone reasonably object to these principles being universally accepted" to the question "could everyone rationally will that they be universally accepted" produces a moral theory that requires us to make significantly greater sacrifices, and permits or requires others to impose such greater sacrifices on us.

This move would, however, also solve a difficulty that arises for a contractualist view like mine in cases of the first type.[75] If someone in Grey's position could reasonably reject a principle requiring him to make the organ donation, why would it not follow that someone in the position of the proposed recipient could reasonably reject a principle permitting Grey not to make the donation. After all, the personal reason that this person has for objecting to such a principle seems at least as strong as Grey's reason for rejecting the more demanding principle, and the cost to Grey is less. This would seem to lead to a moral standoff, in which there is no right answer to the question of what one should do. Shifting to the "what everyone could rationally will" (or concluding, with Parfit, that the reasonable rejection standard in fact collapses into this one) would solve this problem, albeit at a certain cost.[76]

Let me close by expression my agreement with a point that Parfit makes in his conclusion. Given its emphasis on impartial reasons and optimific principles, the Triple Theory that he proposes in his conclusion sounds (at least on first impression) more like consequentialism than my version of contractualism does. So one may question whether his Triple Theory is essentially a contractualist theory or a consequentialist one.

Parfit is correct, I believe, in saying that this theory is contractualist. Any plausible moral view makes what is right or wrong in many cases depend on the harms and benefits to individuals. A theory is consequentialist only if it takes the value of producing the best consequences to be the foundation of morality. Parfit's combined theory does not do this. According to that theory it matters whether the principles that would permit an action would be optimific. But this matters only because these are the principles that everyone has reason to will, and taking what can be justified to others—what they have reason to will—as the most fundamental moral idea is the essence of contractualism, at least as I have described it.

Recognizing the idea of justifiability to others as basic opens up a possibility that Parfit does not discuss, but which I think should not be neglected. Many people may be drawn to consequentialism because they see that there are some situations in which it the morally correct way to decide what to do is to figure out what would produce the best consequences overall. Decisions by public officials about what kind of hospitals to build may be a good example. Because producing the best

---

[75] Thomas Nagel raises this problem in *The View From Nowhere* (New York: Oxford University Press, 1986), pp. 50-51, 172.

[76] That is to say, it would solve the problem if in such situations there always is some principle that everyone could rationally will to be universally accepted (if the "uniqueness condition" is fulfilled.) This depends on the relative strength of impartial and self-interested reasons.

consequences seems so obviously to be the right standard in these cases, people then infer that this idea is always morally basic. This seems to me to be a mistake: producing the best consequences might be the correct standard in these cases not because it is the basis of morality but because it is what is owed to people in situations of that kind, by agents who stand in a certain relation to them. Recognizing the contractualist idea of justification to others as morally basic allows us at least to raise the possibility that although what is owed to others in some situations is to follow the principles that would produce best consequences, impartially understood, this need not always be the case. In other cases our responsibilities and obligations may be different.

Of course it needs to be asked why this should be so, if it is so. And it might be responded that the cases in which it appears to be the case are in fact misleading: they are cases in which, because of the burdens of being impartial, *optimific* principles would permit people decide what to do on a basis other than what would be impartially best. But, as I said earlier in discussing Parfit's treatment of partiality toward one's friends and relatives, there are two was of describing such cases. Is partiality morally permitted because permitting it is impartially best? Or is it permitted because principles that demanded a higher level of impartiality would be ones that individuals could reasonably reject (for reasons that are not impartial)? The latter seems to me more plausible. In any event, this is a point where the residual tension between Rule Consequentialism and my version of contractualism seems to show itself.

PART FIVE RESPONSES

CHAPTER 18 ON HIKING THE RANGE

65 Actual and Possible Consent

Susan Wolf makes several claims that seem to me both true and important. And we disagree, I believe, less than she thinks.

When Kant explains the wrongness of a lying promise, he writes:

> he whom I want to use for my own purposes with such a promise cannot possibly agree to my way of treating him.

Kant then refers to this remark as 'the principle of other human beings'. Kant's principle, I suggest, is

> (A) It is wrong to treat people in any way to which they could not rationally consent.

Wolf objects that, by interpreting Kant in this way, I abandon the Kantian idea of respect for autonomy, which often condemns treating people in ways to which they do not *actually* consent. But I do not abandon this idea. Many acts, I claim, are wrong, even if people could rationally consent to them, if these people do not in fact consent. To cover such acts, I suggest, we could plausibly appeal to

> *the Rights Principle*: Everyone has rights not to be treated in certain ways without their actual consent. [531]

Nor, I believe, do I misinterpret Kant's remarks about consent. These remarks seem intended to cover all cases. Kant seems to be claiming

> (B) It is always wrong to treat people in ways to which they cannot possibly consent.

This cannot mean

> (C) It is often wrong to treat people in ways to which they do not actually consent.

That is why, when I propose the Rights Principle, I do not claim to be interpreting

Kant. According to some writers, Kant means

> (D) It is always wrong to treat people in ways to which they cannot possibly consent because we have not given them the power to choose how we treat them.

But, as Wolf agrees, this claim is false, and is unlikely to be what Kant means. On my proposed interpretation, more fully stated,

> (E) It is always wrong to treat people in ways to which they could not rationally consent, if these people knew the relevant facts, and we gave them the power to choose how we treat them.

This claim is plausible and might be true. (E) might be called *the Principle of Possible Rational Consent*, but I used the shorter and perhaps misleading name: *the Consent Principle*.

Wolf claims that this principle 'would allow us to do things to a person even if she explicitly refuses consent to it'. This claim could be misunderstood. As Wolf notes, the Consent Principle does not claim to cover all wrong acts, so when this principle fails to condemn some act, it does not thereby *allow* or *permit* this act in the sense of implying that this act would not be wrong. This principle also condemns many such acts, since it would often be irrational to consent to being treated in some way without our actual consent. And on some plausible assumptions, this principle could not conflict with the Rights Principle. If it would be wrong to treat someone in some way without this person's actual consent, the Consent Principle would not require this act.

66 Treating Someone Merely as a Means

According to some of Wolf's other claims, which can be summed up as

> *Wolf's Principle*: If we harm people, without their consent, as a means of achieving some aim, we thereby treat these people merely as a means, in a way that is always to be regretted, and that, if other things are equal, makes our act wrong. [532]

As Wolf notes, I argue against a similar principle. But Wolf does not discuss my proposed alternative. According to my proposed

> *Harmful Means Principle*: It is wrong to impose harm on someone as a means of achieving some aim, unless

> (1) our act is the least harmful way to achieve this aim,
>
> and,
>
> (2) given the goodness of this aim, the harm we impose is not disproportionate, or too great.

To compare these principles, consider

> *Fifth Earthquake*: You and your child are trapped in slowly collapsing wreckage, which threatens both your lives. You could save your child's life by using *Black's* body as a shield, without Black's consent, in a way that would destroy one of her legs. You could also save your own life, by causing Black to lose her other leg. But you believe that this act would be wrong, since it is only the saving of a child that could justify imposing such an injury on someone else. Acting on this belief, you save your child's life by causing Black to lose one leg.

According to Wolf's Principle, since you are harming Black without her consent as a means of achieving one of your aims, you are treating Black merely as a means. Given what is meant by 'merely' and 'as a means', this claim seems to me false. If you were treating Black merely as a means, you would save your own life as well as your child's, by causing Black to lose both legs. We cannot be treating someone merely as a means if, in acting in some way, we are letting ourselves die rather than imposing some lesser injury on this person.

We treat people merely as a means, Wolf also claims, if we use these people in some way that 'neglects or ignores' their 'purposes and plans'. But this claim does not support Wolf's Principle. When you save your child's life by destroying one of Black's legs, you may not be ignoring Black's purposes and plans. You may believe that you ought not to destroy Black's other leg because this second injury would make it even harder for Black to achieve some of her purposes and plans. This may be why you choose to die rather than imposing this injury on Black.

Most of us would believe that, in saving your child's life by destroying one of Black's legs, you would be acting wrongly. This, I assume, would also be Wolf's view. But Wolf's Principle supports this view only if we can truly claim that you are treating Black *merely* as a means. And as I have said, that claim is false, since you are giving up your life for Black's sake.

To defend our belief that your act is wrong, we could appeal instead to my proposed Harmful Means Principle. We could claim that, though there are some lesser harms that you could justifiably impose on Black if that were the only way to save your child's life, it is wrong to achieve this aim by imposing on Black an injury as great as losing a leg. Your act is wrong, we can add, even though you are *not* treating Black merely as a means.

Return next to

> *Bridge*, in which you could save five people's lives by using remote control to cause me to fall in front of a runaway train.

Wolf claims that this act would 'very definitely' treat me merely as a means. In some versions of this case, I argued, you would *not* be treating me merely as a means. But this fact, I also claimed, would not justify your act.

Similar claims apply to other cases. Some of Wolf's remarks suggest that, on my view, there is no objection to harming someone as a means of saving others from greater harms. But that is not my view. I make the different claim that, if it would be wrong for us to impose certain harms on people *as a means* of achieving certain aims, these acts would be wrong *whether or not* we would also be treating these people *merely* as a means. If we appeal to Wolf's Principle rather than the Harmful Means Principle, it would be *harder* to defend the belief that such acts are wrong. On Wolf's view, it would not be enough to appeal to the claim that such acts harm certain people as a means, since we must also defend the claim that these acts treat these people merely as a means. On the view that I suggest, to condemn harming people as a means, we do not need to defend that further and often more doubtful claim.

67 Kantian Rule Consequentialism

Wolf challenges my argument that Kantian Contractualism implies Rule Consequentialism. In giving this argument, Wolf claims, I fail to 'appreciate the value of autonomy and its power to generate reasons'.

We respect people's autonomy, Wolf writes, by

> refraining from interfering with their choices for themselves, and from imposing burdens on them that they would not themselves endorse.

We impose a burden on someone, in Wolf's intended sense, if we act in some way that harms this person without this person's consent. Such acts may be wrong, Wolf claims, even if they would also save several other people from similar or greater burdens. Principles that condemn such acts we can call *autonomy-protecting*. Principles that require or permit some such acts we can call *autonomy-infringing*.

According to what I call the Kantian Contractualist Formula, we ought to follow the principles whose universal acceptance everyone could rationally will, or choose. Such principles are *optimific* if their universal acceptance would make things go best

in the impartial-reason-implying sense. Wolf assumes that certain autonomy-infringing principles would be optimific, since their acceptance would save more people from death or other burdens. Wolf also claims that, when we consider such cases,

> (F) everyone could rationally choose that everyone accepts some other, *non*-optimific autonomy-protecting principle.

In Wolf's words, we could rationally prefer some principle that preserves everyone's autonomy, even if that would reduce our 'overall security against the loss of life and limb'. Wolf calls this a *preference for autonomy over welfare*. Wolf then objects that, since everyone could rationally choose such a non-optimific principle, my argument fails to show that Kantian Contractualism requires us to follow the optimific Rule Consequentialist principles.

To assess this objection, we can again suppose that in

> *Tunnel*, you could redirect some runaway train so that it kills me rather than five other people.

Wolf's autonomy-protecting principles would condemn your saving the five in this way, since this act would impose a great burden on me. According to Wolf's objection,

> (1) everyone could rationally choose that everyone accepts some such principle,

even though

> (2) this principle would not be optimific.

But these claims could not both be true. When we apply the Kantian Contractualist Formula, asking which principles everyone could rationally choose, we suppose that everyone knows the relevant, reason-giving facts. On this assumption, people could rationally choose only what they would have sufficient reasons to choose. If the autonomy-protecting principles would not be optimific, their acceptance would make things go worse in the impartial-reason-implying sense. That is what it means to claim that these principles are not optimific. So everyone would have impartial reasons *not* to choose any such principle. And some people would also have strong personal reasons not to choose any such principle. In *Tunnel*, for example, the five people would know that, if they chose one of Wolf's autonomy-protecting principles, you would fail to save their lives by redirecting the runaway train. Nor would the five have any relevant and strong reason to choose such a principle. Since the five would have both impartial reasons and strong personal reasons *not* to choose any such principle, and they would have no similarly strong

opposing reason, these people would not have sufficient reasons to make this choice. They could not rationally choose any principle that would *both* be significantly non-optimific *and* would require you to let them die.

Wolf might object that, in making these claims, I have overlooked the rationality of a preference for autonomy over welfare. She writes:

> in failing to notice or address the challenge to his argument that is posed by [this] preference . . . Parfit reveals once again a failure to recognize and appreciate the value of autonomy. . .

I did fail to consider what would be implied by the rationality of this particular preference. As I have just argued, however, if this preference were rational, that would be no challenge to my argument. If everyone could rationally choose some autonomy-protecting principle, as Wolf claims, this principle must be *optimific*, since this must be one of the principles that, from an impartial point of view, everyone would have most reason to choose. Unless the five had strong impartial reasons to choose this principle, they would have decisive personal reasons *not* to choose this principle, since that choice would lead you not to save their lives. But Wolf might be right to claim that the five *would* have such strong impartial reasons to choose this optimific autonomy-protecting principle.

Wolf also claims that, given the fundamental value of autonomy within the Kantian tradition, it is doubtful that any Kantian could accept Rule Consequentialism 'without abandoning the spirit that led him to be a Kantian in the first place.' After claiming that everyone could rationally choose some *non*-optimific autonomy-protecting principle, Wolf writes that some Kantians might go further, claiming that the choice of such a principle would be '*uniquely* rational'. On this view, she comments,

> Kantian Contractualism not only fails to imply what Parfit calls Kantian Rule Consequentialism, it implies principles that are very likely, if not certain, to conflict with it.

For similar reasons, however, this view could not be true. For it to be uniquely rational for everyone to choose that everyone accepts some autonomy-protecting principle, everyone must have decisive reasons to make such a choice. And these could not all be *personal* reasons. Some people would have strong personal reasons not to choose any autonomy-protecting principle, since that choice would lead others to let them die, or let them bear some other great burden. So, if we all had decisive reasons to choose that everyone accepts some autonomy-protecting principle, these reasons would have to be impartial. And if we had such reasons, these principles would be optimific, since they would be the principles whose acceptance would make things go best in the impartial-reason-implying sense.

These autonomy-protecting principles would be some of the Rule Consequentialist principles that, as I argue, Kantian Contractualism would require us to follow.

When Wolf challenges my argument, she may be using 'optimific' in some sense that differs from mine. Wolf may assume that, in the cases we are considering, principles would be optimific if their acceptance would best promote people's well-being in certain familiar ways, by giving them the longest life-expectancy or minimizing their risk of being injured. But we should not make this assumption. If we could all rationally prefer to live in a world in which we had more autonomy, though with less 'security against the loss of life and limb', this might be truly claimed to be a world in which our lives would on the whole go better. In preferring this world, we may not be, as Wolf claims, preferring autonomy *over welfare*. Nor should we assume that principles are optimific only if their acceptance would on the whole best promote everyone's well-being. The goodness of outcomes may in part depend on other facts, such as facts about how benefits and burdens are distributed between different people, or facts that are not even about people's well-being. If everyone could rationally choose that everyone accepts some autonomy-protecting principle, this might be one of the principles whose acceptance would make things go best, even if this principle's acceptance would not on the whole best promote everyone's well-being. Rule Consequentialism need not take this Utilitarian form, or any other wholly *welfarist* form.

Wolf may not intend her claims to apply to cases like *Tunnel*. Of those who reject Rule Consequentialism, many would believe that, in *Tunnel*, you would be morally permitted to redirect the train so that it kills me rather than the five. But Wolf does discuss *Bridge*, in which you could save the five only *by* killing me.

Most of us would believe that, in *Bridge*, it would be wrong for you to save the five in this way. According to Wolf's autonomy-protecting principles, it is wrong to impose great burdens on people without their consent. Wolf's principles would not distinguish between *Tunnel* and *Bridge*. In both cases, if you save the five, your act would impose a great burden on me, by killing me without my consent. Wolf also writes:

> many people have a strong preference for being in control of their own lives. . . . 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.

These claims also fail to distinguish between *Tunnel* and *Bridge.* In both cases, I and the five would all have strong reasons to prefer to be the ones calling the shots, deciding what would happen to our bodies, and whether we would live or die.

If we believe that your saving the five would be wrong in *Bridge,* but permissible in *Tunnel*, we cannot appeal to Wolf's autonomy-protecting principles. We must appeal to something like my suggested Harmful Means Principle. In both cases, if you save the five, your act would also kill me, without my consent, but only in *Bridge* would you be killing me as a *means* of saving the five.

I assumed that, in *Bridge*, the optimific principles would require you to save the five by killing me. Wolf questions this assumption. She suggests that, if everyone accepted 'something close to the Harmful Means Principle', this might 'lead to better results' and 'be optimific in the long run'. As before, this suggestion might be correct. As Wolf claims, it can be hard to judge whether some principle would be optimific, since it can be hard to predict the effects of the acceptance of different principles, and hard to decide how good or bad these effects would be. When I discussed *Transplant*, I made a similar claim. The optimific principles, I argued, would require doctors never to kill or injure their patients even when they could thereby save more people's lives. If Wolf's suggestion were correct, because the optimific principles would condemn your saving the five, in *Bridge*, by killing me, this would be no objection to my argument that Kantian Contractualism implies Rule Consequentialism. It would merely make Rule Consequentialism in one way easier for some of us to accept. This view would not here conflict, as I assumed, with most people's moral intuitions.

68 Three Traditions

Wolf does not discuss other moral principles or kinds of case. But she makes some wider comments. In my attempts to develop a Kantian theory, she claims, I depart from Kant's 'explicit positions' in a way that is 'both interpretively implausible and normatively regrettable.'

Wolf is partly referring here to my claim that, on Kant's view, we ought to treat people only in ways to which they could rationally consent. I believe that, for the reasons that I gave above, this claim is neither interpretively implausible nor regrettable.

I also claim that, in several passages, Kant must be appealing to what I call the *Moral Belief Formula*, which condemns our acting on some maxim unless we could rationally will it to be true that everyone believes such acts to be permitted. This claim is not, I believe, interpretively implausible. I then argue that this formula should be revised, so that it does not refer to maxims in the sense that covers policies, and so that it appeals, not to what the agent could rationally will, but to what everyone could rationally will. Since I am here *revising* Kant's formula, these claims cannot be *interpretively* implausible. According to my proposed

revision, we ought to follow the principles whose universal acceptance everyone could rationally will. This revised formula differs little from some of Kant's 'explicit positions'. Kant appeals, for example, to 'the idea of the will of every rational being as a will giving universal law. ' [533]

When Wolf calls some of my claims 'normatively regrettable', she is also referring to my claim that Kantian Contractualism implies Rule Consequentialism. There may be other people who would regret this claim. But we are doing philosophy. We should ask, not whether this claim is regrettable, but whether it is true. I believe that, in Sidgwick's words,

> the real progress of ethical science. . . would be benefited by an application to it of the same disinterested curiosity to which we chiefly owe the great discoveries of physics. [534]

Even if we hope that Kantian Contractualism does not imply Rule Consequentialism, my argument for this conclusion may be sound.

Wolf also writes that, in my development of a Kantian theory, some of what seems to her 'most compelling and distinctive about Kant's own moral perspective gets diluted.' Wolf is partly referring here to the idea of respect for autonomy. But the Kantian Contractualist Formula would, I believe, require us to follow some version of my proposed Rights Principle, according to which we have rights not to be treated in certain ways without our actual consent. For some of the reasons that Wolf describes, this would be one of the optimific principles whose universal acceptance everyone could rationally choose. So this part of Kant's perspective would not, I believe, get diluted.

Wolf may also be thinking of my claim that, in *Bridge*, the Kantian Formula would require you to save the five by killing me. As we have seen, Wolf questions this claim, since she suggests that the optimific principles might condemn such acts. Though I believe that the optimific principles would require doctors never to kill some patient as a means of saving several other people's lives, I am still inclined to believe that in what I call *non-medical emergencies*, like *Tunnel* and *Bridge*, the Kantian Formula would require us to do whatever would save the most lives. This formula would then imply that, in *Tunnel*, you ought to redirect the runaway train so that it kills me rather than the five. Like most other people, I can accept that conclusion. But this formula would also imply that, in *Bridge*, you ought to save the five *by* killing me. And like Wolf, I find this claim implausible. Intuitively, this act seems to me wrong.

This intuition is not, however, strong. There are facts which seem to me to count the other way. Compared with being killed as a side-effect, in *Tunnel*, it would be no worse for me to be killed as a means in *Bridge*. And the Kantian Formula

provides an argument against this intuition. If we were choosing the principles that, in such non-medical emergencies, everyone would follow, we would have more reason, I believe, to choose principles that required you to save the five. Though I am still inclined to believe that it would be wrong for you to kill me as a means, this intuition is not strong enough to convince me that we ought to reject the Kantian Formula. [535]

We have strong reasons, I believe, to accept this formula, and to act on the optimific principles of Kantian Rule Consequentialism. As I have said, however, there might be other cases in which this moral theory conflicts more strongly with our moral intuitions. If that were true, we might justifiably reject this theory.

Wolf makes another, wider claim. 'Like Parfit', Wolf writes, 'I see the Kantian, Consequentialist, and Contractualist traditions as each capturing profound and important insights about value.' When she discusses my argument that these three kinds of systematic theory can be combined, Wolf takes me to be trying to show

> that there is a single true morality, crystallized in a single supreme principle that these different traditions may be seen to be groping towards, each in their own separate and imperfect ways.

Wolf doubts that there is any such principle. Nor, she claims, do we need such a principle. In her words:

> 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.

If there is no single supreme principle, that, I agree, would not be a tragedy. But it *would* be a tragedy if there was no *single true morality*. Conflicting moralities could not both be true. In trying to combine these different kinds of moral theory, my main aim was not to find a supreme principle, but to find out whether we can resolve some deep disagreements. As Wolf claims, it would not matter greatly if morality *turned out* to be less unified, because there are several true principles, which cannot be subsumed under any single higher principle. But if we cannot resolve our disagreements, that would give us reasons to doubt that there are *any* true principles. There might be nothing that morality *turns out to be*, since morality might be an illusion.

CHAPTER 19 ON HUMANITY AS AN END IN ITSELF

69 Kant's Formulas of Autonomy and of Universal Law

I have learnt a great deal from Allen Wood's fascinating books, and I am delighted and relieved by the fact that, in his commentary, Wood expresses agreement with some of my claims. I shall try here to resolve some of our remaining disagreements.

Though Wood believes that Kant at least roughly describes 'the supreme principle of morality', he also believes that Kant's principle cannot provide a *criterion of wrongness*, in the sense of a way of deciding which acts are wrong. Of Kant's various formulations of his supreme principle, Wood has the lowest opinion of Kant's Formula of Universal Law. Wood calls this the 'least adequate' of Kant's formulas, and the formula that most clearly fails to provide a criterion of wrongness. [536] He also writes:

> Self-appointed defenders of Kant. . . will probably never abandon the noble, Grail-like quest for an interpretation of the universalizability test that enables it to serve this purpose, despite the history of miserable failure that has always attended the quest. I regard their attempts as worse than a waste of time, since they encourage critics of Kant's ethics to continue thinking, falsely, that something of importance turns on whether there is a universalizability test for maxims that could serve as such a general moral criterion. [537]

These Kantians, he adds,

> desperately seek ever more creative interpretations of Kant's test in a passionate effort (as they see it) to save Kantian ethics from oblivion. [538]

Since I have tried to show that Kant's Formula of Universal Law can give us a plausible criterion of wrongness, I may seem to be one of these self-appointed defenders of Kant whose noble, Grail-like quest Wood regards as worse than a waste of time. But I cannot claim such nobility. I accept Wood's view that no new *interpretation* of Kant's formula, however creative, could make this formula provide a criterion of wrongness. We ought, I argue, to *revise* this formula. According to my proposed

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

In revising Kant's formula, my aim is the same as Wood's aim in his latest book, *Kantian Ethics.* We are both trying to produce what Wood calls 'the most defensible' Kantian moral theory. To achieve this aim, as Wood notes, we may have to revise some of Kant's claims. [539]

The Kantian theories that Wood and I propose are also, I believe, more similar than Wood assumes. Wood appeals to Kant's Formula of Autonomy, which Kant sums up as 'the idea of the will of every rational being as a will giving universal law. ' [540] This formula, Wood writes,

> tells us to think of ourselves as members of an ideal community of rational beings, in which each of us should strive to obey the moral principles by which we would choose that members of the community should ideally govern their conduct. [541]

In a briefer statement, which we can call

> FA: Each of us should try to follow the principles that we would all choose as the principles that would govern everyone's conduct.

Wood calls FA 'the most definitive form' of Kant's supreme principle, and the formula that we ought always to 'use for moral judgment.' [542] But as Wood also claims, FA is not a reliable criterion of wrongness. [543] If we ask which are the principles that people *would in fact* choose, we could not predict which principles other people would choose. Nor could we assume that everyone would choose the same principles.

We ought, however, to revise FA, so that this formula refers to the principles that it would be *rational* for all of us to choose. This revised formula would better express Kant's idea of the will of every *rational* being as giving universal law. And this revision is also clearly needed, since there are countless bad principles that we might all irrationally choose, and these cannot be the principles that we should try to follow. So FA should become

> FA2: Each of us should try to follow the principles that it would be rational for all of us to choose as the principles that would govern everyone's conduct.

This claim is another statement of my Kantian Contractualist Formula. [544] Though my proposed Kantian theory revises Kant's Formula of Universal Law, and Wood's proposed theory revises Kant's Formula of Autonomy, these revisions both lead us to what I have called Kantian Contractualism. That is not surprising given Kant's

assertion that these different 'ways of representing the principle of morality are, fundamentally, only so many formulas of precisely the same law'. [545]

Return now to Wood's claim that nothing of importance turns on whether there is some 'universalizability test' that provides a criterion of wrongness. This claim would be justified only if either (1) we already have some other, wholly reliable criterion, or (2) we would not be helped by having some such criterion, since we can always reliably judge, without using any criterion, whether some act would be wrong. Wood does not claim either (1) or (2). So Wood, I believe, should agree that it matters whether Kantian Contractualism provides a good criterion of wrongness. And that, I have argued, may be true.

70 Rational Nature as the Supreme Value

Wood also discusses Kant's Formula of Humanity, which is clearly *not*, as Kant asserts, a different way of stating 'precisely the same law'. When Kant presents this formula, I suggest, Kant claims it to be wrong to treat people in ways to which they could not rationally consent. This claim, I argue, is both plausible and defensible. I am glad that, in his commentary, Wood seems to agree. Kant's Formula of Humanity includes the different claim that we must never treat rational beings merely as a means. Though this claim is also plausible, I argue that it needs to be revised, and that it adds little to Kant's view. Though it is wrong to *regard* anyone merely as a means, whether our *acts* are wrong seldom if ever depends on whether we are treating people merely as a means. Wood ignores this part of Kant's formula, because he believes that it adds nothing to Kant's view. [546]

Wood restates Kant's formula as

> FH: We should always respect humanity, or rational nature, as an end-in-itself.

This version of Kant's formula, I claim, is too vague to provide a criterion of wrongness. Wood agrees. [547]

Unlike me, however, Wood believes that FH is the most important of all Kant's statements of his supreme moral law. This formula, Wood claims, 'is fundamentally the articulation of a basic value'. He even writes:

> Perhaps the most fundamental proposition in Kant's entire ethical theory is that rational nature is the supreme value. . . [548]

This supreme value, Wood suggests, gives us our 'rational ground or motive' to obey the moral law. If there are categorical imperatives, Kant argues, we must

have a reason to obey them. This reason would have to be provided by something that is an end-in-itself, having supreme and absolute worth. And this end-in-itself, Kant claims, is humanity or rational nature. With these claims, Wood writes, Kant gives us 'a deeply true account of the foundations of ethics'. On this interpretation of Kant's view, which I shall call

> *Wood's Foundational Thesis*: Humanity or rational nature has the supreme value that both grounds morality and gives us our reason to obey the moral law.

Herman similarly writes:

> Kant's project in ethics is to provide a correct analysis of 'the Good', understood as the determining ground of all action.

No moral theory could succeed, Herman claims, 'without a grounding concept of value'. On Kant's theory, it is the value of rational nature that gives morality its 'end or point', thereby showing how morality's demands on us 'make sense'. [549]

These claims need to be further explained. When Kant uses the words 'humanity' or 'rational nature', he is sometimes referring to rational beings, or persons. All persons, Kant claims, have *dignity*, which he defines as absolute, unconditional, and incomparable value or worth. [550] So the supreme value which Kant claims to ground morality might be the dignity of all persons.

Kantian dignity, many writers assume, is a kind of supreme *goodness*. For example, Herman calls the dignity of rational nature a value that is 'absolute in the sense that there is no other kind of value or goodness for whose sake rational nature can count as a means'. [551] Wood calls rational nature 'the underivative objective good'. [552] Kerstein similarly writes that humanity is 'absolutely and incomparably good', [553] and Korsgaard writes that, on Kant's view, humanity must be treated 'as unconditionally good'. [554] [555]

As I pointed out, however, some rational beings or persons are not good. Hitler and Stalin were two examples. Wood comments:

> 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 same worth as even the morally best people. . . 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*. [556]

If the dignity of persons were a kind of supreme goodness, and Hitler and Stalin had this kind of goodness, that would imply that Hitler and Stalin were supremely good. Since that is clearly false, as Kant would have agreed, we should conclude that, at least when had by persons, dignity is not a kind of goodness. As Wood, Hill, and others claim, the dignity of persons is a kind of 'moral status', or a 'value to be respected'. Though Hitler and Stalin were not good, they had dignity in the sense that, as rational beings, they had the moral status of being entities who ought always to be treated only in certain ways.

Return now to Wood's Foundational Thesis. If we take 'rational nature' to refer to rational beings, or persons, this thesis implies that

> (1) our reason to treat all persons only in certain ways is provided by the fact that persons have supreme value.

This supreme value, as we have just seen, is not a kind of goodness but a kind of moral status. So we can restate (1) as

> (2) our reason to treat all persons only in certain ways is provided by the fact that persons have the moral status of being entities who ought to be treated only in these ways.

This less appealing statement of Wood's Thesis could not be claimed to ground morality's requirements in what Herman calls 'a correct analysis of the Good'. (2) claims only that our reason to follow these requirements is provided by the fact that morality requires these acts. This claim does not give morality what Herman calls an end or point, showing how morality's demands make sense.

Wood suggests another version of his thesis. Kant sometimes uses 'humanity' and 'rational nature' to refer to

> our *non-moral rationality*, which Kant describes in part as our 'capacity to set an end---any end whatsoever', and which also includes, Wood claims, both instrumental and prudential rationality, and various other rational abilities. [557]

These kinds of rationality, Wood writes, have 'the absolute worth that grounds morality'. [558]

In defending this version of his thesis, Wood once claimed that, according to Kant:

> When we use our capacity to set an end, by choosing to try to fulfil some desire, we thereby make this end good.
>
> The source of something's goodness must itself be good.

Therefore

> Our capacity to set an end is good. [559]

This argument involves, Wood wrote,

> an inference from the objective goodness of the end to the unconditional objective goodness of the capacity to set the end. [560]

Wood even suggested that, on Kant's view, the 'rational choice of ends is the act through which objective goodness enters the world'. [561]

This is not, I believe, Kant's view. Kant did not believe that our capacity to set ends is the source of all goodness, such as the goodness of good wills, or deserved happiness. And Wood now rejects, and believes that Kant rejects, this argument's first premise. Wood accepts a value-based objective theory both of reasons and of the goodness of our ends, and he calls these views 'good Kantianism'. [562]

Our non-moral rationality may have some kinds of value, to which I shall return. But such rationality cannot be defensibly claimed to have, as Wood suggests, the supreme goodness or absolute worth that grounds morality, by giving us our reason to obey the moral law.

There is another possibility. Kant writes

> morality, and humanity insofar as it is capable of morality, is that which alone has dignity. [563]

In this and some other passages, as Wood notes, Kant ascribes dignity to rational nature 'not in its capacity to set ends, but only in its capacity of giving (and obeying) moral laws.' Surprisingly, Wood also writes

> It is the *capacity* for morality. . . not its successful exercise, that has dignity. [564]

The *unexercised* capacity for morality, as had by people like Hitler and Stalin, cannot be claimed to be supremely good, or to be what grounds morality. [565]

Wood's Foundational Thesis might appeal instead to the *exercised* capacity to give and to obey moral laws, which is roughly what Kant calls a 'good will'. Kant claims, much more plausibly, such good wills are supremely good. So Wood's Foundational Thesis might become the claim that

> (3) Kant grounds morality on the goodness of good wills.

Wood considers and rejects this claim. He reminds us that, on Kant's view, we cannot be certain that any actual person has a good will. [566] Wood then writes: 'If

only the good will had the dignity of an end in itself. . . the existence of such an end, and consequently the validity of categorical imperatives, would be doubtful. ' [567]

This argument is not, I believe, sound. For something's goodness or good features to give us a reason for acting, which might be decisive and categorical, this thing need not exist. Many of our acts are intended to achieve some merely possible good end. So, if Kant had stated a version of Wolf's Foundational Thesis, Kant might have claimed that

> (4) the supreme goodness of good wills gives us our reason to try to have such a will, and to act rightly. [568]

For us to have such a reason, it must be possible for us to have good wills, and to act rightly. But Kant believes we know that to be possible.

Remember next Kant's claim that the Highest or Greatest Good would be a world of universal virtue and deserved happiness. Everyone, Kant claims, ought always to strive to promote this ideal world. And Kant also writes,

> the moral law commands me to make the greatest possible good in a world the final object of all my conduct. [569]

These claims overlap with (4). What would make this the best possible world would be the fact that everyone had good wills and acted rightly, thereby deserving their happiness. If these claims are true, that would be enough to give to morality what Herman calls an 'end or point', so that morality's demands 'make sense'.

71 Rational Nature as the Value to be Respected

Wood gives another argument against the view that Kant grounds morality on the goodness of good wills. On Kant's theory, Wood writes, 'all reasons for acting are based, directly or indirectly, on the objective value of rational nature'. 'What morality demands most fundamentally is that we show respect for that value', and acts that are wrong 'all involve treating that value. . . with disrespect'. [570] These claims would not be plausible, Wood argues, if the value of rational nature was the goodness of good wills. When we ask what makes it wrong to injure, coerce, deceive, or otherwise mistreat people, the answer does not seem to be that such acts show disrespect for the goodness of such wills. As Wood points out, from Kant's claim that good wills are supremely good, we cannot draw any conclusions about what we ought morally to do. This claim, Wood concludes, has only 'marginal' importance in Kant's moral theory. [571] Kant's ethics is grounded, *not* on the goodness of good wills, but on what Wood calls the 'absolute worth of rational nature'. [572]

Though this argument has more force, its conclusion is, I believe, too simple. In discussing Kant's theory, we can distinguish between what gives morality its end or point, and the properties or facts that make acts wrong. Wood's argument, I believe, does not count against the view that Kant's ethics is grounded on the goodness of good wills and deserved happiness. This part of Kant's theory may not be intended to help us to decide which acts are wrong. It is a separate question whether, as Wood claims, our acts are wrong when and because they show disrespect for the value of rational nature.

Kant uses 'rational nature' to refer both to rational beings and to the rationality of these beings. The value of rational nature therefore consists in part in the dignity of all rational beings, or persons. As we have seen, this dignity is not a kind of goodness, but is the moral status of being entities who ought to be treated only in certain ways. The claim that persons have this status does not help us to decide how persons ought to be treated.

When Wood refers to the supreme value of rational nature, he is more often referring to the value of non-moral rationality, such as prudential rationality. Though Wood no longer claims that our capacity to set an end confers goodness on what we choose, he still takes Kant to be claiming truly that 'the correct exercise of one's rational capacities. . . must be esteemed as unconditionally good'. [573] On Kant's view, Herman similarly writes, 'the domain of "the Good" is rational activity and agency: that is willing'. [574]

These claims are not, I believe, justified. Some kinds of rational activity may have great intrinsic value as achievements, and this would support Kant's claim that we ought to develop and use our various rational abilities. But unlike good wills, non-moral rationality cannot be claimed to be supremely good. The rational agency of Hitler and Stalin was not good. Nor, I believe, would Kant have made this claim. On Kant's view, as Herman notes, what is good is only *good* willing. [575]

Even if rational agency is not supremely good, such agency might be claimed to have what Wood calls 'the basic value to be respected'. [576] Our acts are wrong, Wood suggests, when and because they fail to respect the value of non-moral rationality. Herman makes similar claims. On Kant's view, she writes,

> Failure to assign correct value to rational agency---discounting the conditions of human willing---is the 'content' of morally wrong action. [577]

Most wrong acts are wrong, Herman suggests, because of the ways in which these acts destroy, obstruct, or misuse rational agency. Coercion is wrong, for example, because it involves 'an attack on agency', deception is wrong because it frustrates rational agency, and violence is wrong because it attacks agency's 'conditions.'

These claims are, I believe, misleading. On Kant's view, Herman also writes:

> killing is not wrong because it brings about death, and mayhem is not wrong because it brings about pain or harm. . . The kind of value. . . I have as an agent is not lost or compromised in dying. [578]

What makes killing wrong is instead 'some erroneous valuation'. I can justifiably resist aggression, Herman writes, because

> the aggressor acts on a maxim that involves the devaluation of my agency. . . I am not acting to save my life as such, but to resist the use of my agency. . . [579]

Rational agency seems here to have the kind of value that some people claim for chastity, and self-defence to be like the protection of our chastity---whose value, women were often told, is not lost or compromised in dying. I doubt that this is really either Kant's or Herman's view. Aggressive violence *is* wrong, I believe, not because it devalues rational agency, but because it brings about death, pain, or other harms.

Similar claims apply to deception and coercion. What makes these acts wrong is not, I believe, their 'failure to assign correct value to rational agency'. People can act rationally when they are being deceived and coerced. Such acts are wrong for other reasons, such as the fact that people could not rationally consent to them, or the fact that such acts treat *people*, not their *agency*, with disrespect.

Return next to Wood's claim that the capacity to set ends, and the other components of non-moral rationality, have 'the absolute worth that grounds morality'. To show respect for this value, Wood writes, we must help other people to achieve their permissible ends. But if it was other people's non-moral rationality that had such worth, that would give us no reason to help these people to achieve their ends. Other people could act just as *rationally*, even if less *successfully*, without our help. Wood similarly claims that concern for alleviating human suffering is 'grounded' in the 'fundamental value' of non-moral rationality. These claims are, I believe, misleading. Our concern to relieve people's suffering should be grounded, not in the value of these people's rationality, but in the ways in which suffering is bad for these people, by being a state that they have strong reasons to want not to be in. We have similar reasons to relieve the suffering of those abnormal human beings who have no rational abilities, and the suffering of non-rational animals. As Bentham said, the question is not 'Can they reason?' but 'Can they suffer?'

Wood also claims:

> to act morally is always to act for the sake of a person, or more precisely, for the sake of humanity in someone's person. [580]

> the fundamentally valuable thing. . . is a rational being, a person – or, more precisely, rational nature in a person.

These more precise claims are, I believe, mistaken. We ought to act for the *person's* sake, not for the sake of her non-moral rationality. And it is the *person*, not her rationality, who has the high moral status that Kant calls dignity.

Wood is aware of this objection. Some of Kant's readers, Wood writes, may

> worry about the injunction to respect humanity (or rational nature) in someone's person. They fear that it means respecting only an abstraction and not the persons themselves. Kant's answer to these worries, of course, is that rational nature is precisely what makes you a person, so that respecting it *in* you is precisely what it means to respect *you*. [581]

This suggested answer is not, I believe, true. Respecting your non-moral rationality is not respecting *you*. Wood also writes that, on Kant's view,

> respect for the dignity of humanity is identical with respect for law grounding morality in general.

Kant does claim that respect for a person is, strictly speaking, respect for the moral law. [582] But these are not the claims that have rightly made Kant's Formula of Humanity so widely accepted and loved. Respect for persons should be, precisely, respect for *them*.

CHAPTER 20 ON A MISMATCH OF METHODS

72 Does Kant's Formula Need to be Revised?

In some of her brilliant discussions of Kant's Formula of Universal Law, Barbara Herman claimed that this formula cannot provide a criterion of wrongness. [583] Despite 'a sad history of attempts', she wrote, '. . . no one has been able to make it work'. [584] Herman, I have argued, was right. In its present form, Kant's Formula cannot succeed. But if we revise this formula, I claimed, we can make it work. Herman would agree, I hoped, that her 'sad history' has a happy ending.

My hopes were dashed. In her commentary, Herman seems to argue that Kant's Formula does not need to be revised. She also argues that my proposed revision could not, even if it were needed, achieve Kant's aims. [585]

One of my arguments can be summed up as follows:

> According to Kant's Formula, it is wrong to act on any maxim that we could not rationally will to be universal.
>
> There are many maxims that we could not rationally will to be universal, though acting on these maxims would often not be wrong.

Therefore

> When applied to such maxims, Kant's Formula would often mistakenly condemn acts that were not wrong.

To illustrate these claims, I imagined that some Egoist has only one maxim 'Do whatever would be best for me'. For self-interested reasons, this man pays his debts, keeps his promises, puts on warmer clothing, and risks his life to save a drowning child, hoping to get some reward. I then argued:

> (A) When this man acts in these ways, his acts have no moral worth, but he is not acting wrongly.
>
> (B) This man is acting on an Egoistic maxim that he could not rationally will to be universal.

Therefore

> Kant's formula falsely implies that this man *is* acting wrongly.

In some passages, Herman seems to reject premise (A). Kant's Formula, she suggests, *truly* implies that this man is acting wrongly.

In defending this suggestion, Herman claims that, on Kant's view,

> (C) we act wrongly when we act for the wrong motive, or our decision about how to act was made in some morally defective way.

When my Egoist saves the drowning child because he hopes to be rewarded, this man's selfish motive, Herman suggests, makes his act wrong. [586] And my imagined ruthless gangster acts wrongly, Herman also suggests, when this man buys his cup of coffee from a coffee seller whom he regards as a mere means.

Herman remarks that, in suggesting that these acts are wrong, she may seem to be ignoring Kant's

> famous distinction between morally worthy and duty-conforming actions, the former requiring that the action be done from a moral motive, the latter motive-indifferent.

She also writes:

> The doctrine of moral worth is not the only place where Kant is taken to be offering a motive-independent notion of wrongness; also noted are his views of perfect duties and duties of justice.

But she then claims:

> Neither view supports the general thesis of motive-independent wrongness. In both cases, the error in thinking that they do is instructive.

Kant, I believe, *does* use 'a motive-independent notion of wrongness', so there seems to be no error here. It will be enough to consider what Kant calls 'duties of justice'. Kant claims that, unlike duties of virtue, which require us to act for the right motive, duties of justice can be fulfilled whatever our motive. [587] As Herman writes, these duties

> are indeed about external actions only; motives are not relevant to their correct performance.

Kant includes, among duties of justice, duties to pay our debts and keep our promises. When my Egoist acts in these ways for self-interested motives, he fulfils these duties. So Kant, I believe, would accept my claim that these acts are not wrong.

Herman concedes that these acts are in one sense permissible. But on Kant's view, she claims,

> avoiding impermissibility and avoiding wrongness are not the same thing; actions can be "not impermissible" and yet wrong.

She also writes:

> duties of justice are not one of the classes of moral duties, on all fours, as it were, with duties of aid or respect or friendship. They are institution-based duties. . . they only come into existence through the legislative activity of a state.

Herman elsewhere suggests that we ought not to 'model wrongness on a legal notion of impermissibility.' And Kant himself writes that, when we fulfil duties of justice for selfish motives, that gives our acts 'legality' not 'morality'.

Kant's remark could be misunderstood. Duties of justice *are*, on Kant's view, moral duties. As Kant writes

> all duties, just because they are duties, belong to ethics. [588]

When Kant claims that our failure to fulfil duties of justice makes our acts 'illegal', he does not mean only that such acts are against the criminal, state-based law. He often means that such acts are against the *moral* law. Kant often uses 'illegality' to refer to the kind of wrongness, or *moral* impermissibility, that is involved in failing to fulfil duties of justice. This kind of wrongness is, in Herman's phrase, *motive-independent,* since we can fulfil such duties, thereby avoiding this kind of wrongness, whatever the motive on which we act. Kant's prudent merchant does his duty when he pays his debts, even though his motive is to preserve his reputation and his profits. Kant calls such acts 'right' or 'in conformity with duty', and our failure to fulfil such duties he calls 'wrong' or 'contrary to duty'.

Despite her remarks quoted above, Herman seems to agree that Kant sometimes uses 'wrong' in this motive-independent sense. Though we have only a duty of justice not to steal, Herman refers to the 'moral wrong of stealing'. And she writes:

> impermissibility is the mark of a class of wrongful actions that are wrong no matter what the agent's motive.

Herman's claim can at most be that Kant also uses 'wrong' in at least one other sense. And she does make such claims. On Kant's view, she writes:

> An externally conforming action that lacks moral worth is a behavior whose connection to moral correctness is conditional or accidental. It is in that sense *not* a correct action.

She also writes:

> An agent who ignores or fails to respond appropriately to the morally relevant features of her circumstances acts in a way that is wrong.
>
> Wrongness. . . arises from the principles of the deliberating agent and is about whether, through them, she has a sound route of reasoning to her action.

Herman might claim that, even when some act is morally permissible and in conformity with duty, this act may be in these other senses wrong.

If we distinguish these senses of 'wrong', my argument could become:

> (D) When my Egoist pays his debts, saves the drowning child, and puts on warmer clothing, his acts have no moral worth, but these acts are not wrong in the sense of being morally impermissible and contrary to duty.
>
> (E) According to Kant's Formula, it is in this sense wrong to act on any maxim that we could not rationally will to be universal.
>
> (B) When my Egoist acts in these ways, he is acting on an Egoistic maxim that he could not rationally will to be universal.

Therefore

> Kant's Formula falsely implies that these acts are in this sense wrong.

Though Herman seems to accept both (D) and (B), she might reject (E). She might claim that, in proposing his formula, Kant does not intend to provide a criterion of whether our acts are wrong, in the sense of being morally impermissible and contrary to duty. Herman has elsewhere made this claim.[589] But Kant often declares or assumes that his formula provides such a criterion. For example, he writes:

> to inform myself in the shortest and yet infallible way. . . whether a lying promise is in conformity with duty, I ask myself: would I indeed be content that my maxim. . . should hold as a universal law? [590]
>
> common human reason, with this compass in hand, knows very well how to distinguish in every case what is good and what is evil, what conforms with duty or is contrary to duty. [591]

As these and many other passages together show, Herman cannot defensibly reject premise (E). [592] My argument, I believe, is sound. When my Egoist acts in the ways that I have described, Kant's Formula falsely implies that these acts are wrong in the sense of being morally impermissible and contrary to duty. So this formula fails, and needs to be revised.

This objection, moreover, can take another form, to which several of Herman's claims do not apply. There are people who are conscientious, and who sometimes act in ways that they truly believe to be right, though these people are acting on maxims that they could not rationally will to be universal. One example would be Kant himself if, as we can suppose, he accepted the maxim 'Never lie'. Kant could not have rationally willed it to be true that no one ever tells a lie, not even to a would-be murderer who asks where his intended victim is. So Kant's Formula would imply that, whenever Kant acted on this maxim by telling anyone the truth, he would be acting wrongly. That claim is clearly false. Suppose next that we accept the maxims 'Never steal' and 'Never break the law'. We could not have rationally willed it to be true that no one ever steals or breaks the law, even when these are the only ways to save some innocent person's life. So Kant's Formula implies that, whenever we act on these maxims, by returning someone's property or keeping some law, we would be acting wrongly. These claims are also clearly false. As before, to avoid this objection, Kant's Formula must be revised.

73 A New Kantian Formula

We should revise Kant's Formula, I argued, by making this formula refer, not to *maxims* in the sense that covers policies, but to the acts that we are considering, described in the morally relevant ways.

Herman does not discuss my proposed revisions. But some of Herman's claims, quoted above, suggest some other ways in which Kant's formula might be revised. We might distinguish between

> an act's being wrong in the sense of being morally impermissible and contrary to duty,

and

> an act's being wrong in the sense that it
>
> > (1) lacks moral worth,
> >
> > (2) fails to respond appropriately to the morally relevant facts,

(3) is done for the wrong motive, or

(4) is only accidentally in conformity with duty.

We might then suggest that, on a different version of Kant's Formula, which we can call

> *the New Kantian Formula*: When we act on some maxim that we could not rationally will to be universal, our act is wrong in one or more of these other senses.

We ought, I believe, to reject this formula. Though it matters whether our acts have the properties described by (1) to (4), it would often be misleading to call such acts wrong. Nor would this formula be a good criterion of whether people's acts *are*, in these various senses, wrong.

As we have seen, Herman's remarks suggest that

> (1) when some morally required act lacks moral worth, this act is in one sense incorrect or wrong.

But this would not, I believe, be a defensible or useful sense of 'wrong'. When my Egoist pays his debts and keeps his promises for self-interested reasons, his acts have no moral worth, but that is no reason to call these acts wrong.

Even if we called such acts in this sense wrong, that would not give us a reason to appeal to the New Kantian Formula. Whether our acts have moral worth does not depend on whether we could will our maxims to be universal. Suppose that Kant tells someone the truth, at a great cost to himself, because he rightly believes this act to be his duty. As I have said, this would be more than enough to give this acts moral worth. It would be irrelevant whether Kant was acting on some maxim, such as 'Never lie', which he could not rationally will to be universal. So the New Formula should not assume that all such acts lack moral worth.

Consider next Herman's claim that

> (2) we act wrongly when we fail to respond appropriately to the morally relevant facts.

When my Egoist saves the drowning child, his act is not in this sense wrong. It is wholly appropriate to save drowning children. Nor is it in this sense wrong for my Egoist to pay his debts and keep his promises. These are wholly appropriate acts. Nor would Kant act inappropriately if he acted on the maxim 'Never lie' by telling someone the correct time of day. So the New

Formula should not claim that, when we act on some maxim that we could not rationally will to be universal, we are failing to respond appropriately to the relevant facts. That claim would often be false.

Some of Herman's remarks suggest that

> (3) in my imagined cases, my Egoist acts wrongly in the sense that he acts for the wrong motive.

Even when my Egoist responds appropriately to the relevant facts, he might be acting for the wrong motive. But (3) is also, I believe, false. We should distinguish here between this man's *maxim*, 'Do whatever would be best for me', and the self-interested *motive* on which this man acts. Though this man's maxim is morally defective, his motive is not always wrong. In my imagined case, since no one has a duty to risk their life to save the drowning child, no one would be acting for the wrong motive if they chose, for self-interested reasons, not to risk their life. So we should similarly claim that, when my Egoist chooses, for self-interested reasons, to risk his life in an attempt to save this child, he is not acting for the wrong motive. Nor does he act for the wrong motive when he fulfils his duties of justice, by paying his debts and keeping his promises. As Herman seems to admit, we can fulfil these duties whatever our motive. Nor should the New Formula claim that, whenever we act on some maxim that we could not rationally will to be universal, we have the wrong motive. Kant would not be acting for the wrong motive if he rightly told someone the truth because he believed this act to be his duty. It would be irrelevant whether he was acting on a maxim, 'Never lie', that he could not rationally will to be universal.

Herman also suggests that

> (4) when our acts are only accidentally morally permissible, or in conformity with duty, these acts are in one sense wrong.

This claim does not, I believe, describe a useful sense of 'wrong'. When some people follow certain traditional rules, or certain religious beliefs, they are acting on incorrect principles, and using unsound moral reasoning. In such cases, when these people do their duty, their acts would be only accidentally in conformity with duty. But we should not claim that these people's acts are all, in one sense, wrong. When these people act rightly, for the right motive, believing that their acts are right, their acts are not in any sense wrong.

Return next to my claim that, if Kant acted on the maxim 'Never Lie' by telling someone the correct time of day, Kant's Formula would falsely imply that this act was wrong. Herman might reply that Kant's act *would be* in one sense

wrong, since this act would be only accidentally in conformity with duty. Kant's maxim might have led him to act wrongly, as would be true in the possible case in which Kant told some would-be murderer where his intended victim was. But that is not enough to justify the claim that, when Kant tells someone the correct time of day, Kant's act is in one sense wrong. Our claim should be only that, *if* Kant had acted on his maxim in this other, very different possible case, that *different* act would have been wrong.

Return now to my imagined gangster, who regards other people merely as a means, and who pays for his coffee merely because he thinks it not worth stealing from the coffee seller. Herman imagines that this man is morally reborn, and looks back with horror at his earlier life. She then writes:

> It's easy enough to imagine him concluding that what he had done was wrong: it was a matter of sheer luck that there was a benign outcome. It would not be inapt for him to wish it had not happened: not the paying for the coffee, of course, but the entire episode. If a sign of wrongdoing is guilt, or a sense that apology might be in order, motive or attitude can suffice to trigger it, and a change in attitude is often integral to the work of moral repair for what was done.

As Herman here claims, however, this man has no reason to wish that he had not paid for his coffee. And that is all that this man *did*; so he should not conclude that 'what he had done was wrong', nor is it true that he should apologize for what he did. As I wrote:

> though this gangster treats the coffee seller merely as a means, what is wrong is only his *attitude* to this person. In buying his cup of coffee, he does not *act* wrongly.

Herman herself writes elsewhere:

> not all things required of the Kantian agent are required *actions*. . . we are also required to adopt a general policy: to be willing to help when the need is there. [593]

Since we are also morally required not to regard other people merely as a means, my gangster's attitude is wrong. And we might agree that, in having this wrong attitude to the coffee seller, this gangster in one sense wrongs this person, and should apologize later for having had this attitude. But there is no useful sense in which, when this main paid for his coffee, what he did was wrong.

In the passages that I have just been discussing, and several others, Herman very well describes, and makes several plausible and original claims about, some of the ways in which it can be morally important whether our acts have the

properties described by (1) to (4). But as I have tried to show, we should not claim either that all such acts are in one sense wrong, or that our acts have these properties when we act on maxims that we could not rationally will to be universal. The first claim would be at least misleading, and the second would often be false.

74 Herman's Objections to Kantian Contractualism

In the last two sections, I have tried to show that Herman's claims do not answer one of my objections to Kant's Formula of Universal Law, nor do these claims suggest an acceptable way to revise this formula.

I gave several other objections to Kant's Formula, none of which Herman directly discusses. These objections show, I believe, that Kant's Formula must be revised.

My proposed revision Herman calls a 'hybrid theory', which seems to her deeply un-Kantian. This revision, she writes,

> cannot capture what is most distinctive about Kant's theory. The mismatch of methods is too profound. . . If the separation of the two methodologies is so wide . . . there may not be much to be gained from a point-by-point comparison of the best classical Kantian arguments and Parfit's hybrid reconstruction. They are simply too far apart.

These remarks surprise me. Since I revise Kant's Formula in only two main ways, a point-by-point comparison is easy to make. According to one version of Kant's Formula, which I called

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

According to my proposed revision,

> MB5: It is wrong to act in some way unless everyone could rationally will it to be true that everyone believes that such acts are morally permitted.

One difference here is that

> (F) instead of appealing to what the *agent* could rationally will, my proposed formula appeals to what *everyone* could rationally will.

This revision does not make these two formulas 'too far apart' to be worth comparing. What *each* of us could rationally will, Kant and many Kantians assume, is the same as what *everyone* could rationally will. This assumption, I claimed, is not true. What could be rationally willed by some people who are men, rich, or powerful could *not* be rationally willed by some people who are women, poor, or weak. Kant's Formula therefore permits some acts that are clearly wrong. To avoid this objection, I argued, Kant's Formula should appeal to what everyone could rationally will. No Kantian could have a deep objection to this proposed revision.

The other difference is that

> (G) unlike Kant's Formula, which applies to maxims in the sense that covers policies, my proposed formula applies to certain kinds of act, described in the morally relevant ways.

This revision does abandon one of the distinctive features of Kant's moral theory, since only Kant and Kantians often use the concept of a maxim. But as I argued, this feature of Kant's theory is a mistake, which must be corrected. It is worth restating this argument in its most general form. When Kant first states his formula, he writes:

> I ought never to act except in such a way that I could also will that my maxim would become a universal law. [594]

In this and many other passages, Kant claims only that we act wrongly *if* we act on maxims that we could not rationally will to be universal. Taken strictly, this claim allows that there might be other ways in which some acts are wrong. But Kant's Formula is one statement of what Kant claims to be the supreme moral principle. So Kant clearly means that we act wrongly *if and only if*, or *just when,* we act on maxims that fail the test provided by Kant's Formula. We can now argue:

> According to Kant's Formula, we act wrongly just when we act on some maxim that fails a certain test.

Therefore

> Kant's Formula implies that, if some maxim fails this test, it is always wrong to act upon it, and that, if some maxim passes this test, it is always permissible to act upon it.
>
> There are countless maxims on which it is sometimes but not always wrong to act.

Therefore

> When applied to such maxims, Kant's Formula either mistakenly condemns some acts that are morally permissible, or mistakenly permits some acts that are wrong.

As this restatement shows, nothing turns on the content of Kant's test, or on the sense in which we could not will some maxims to be universal laws. Kant's Formula fails simply because it applies to maxims, in the sense that covers policies. For Kant's Formula to succeed, it would have to be true that, if it would *ever* be wrong to act on some maxim or policy, such acts would *always* be wrong. And that is clearly false. It is sometimes but not always wrong to act on the maxims 'Do whatever would be best for me', 'Never lie', and 'Never break the law'. And there are many other *mixed maxims* of this kind.

It might be objected that, if we revise Kant's Formula so that it does not refer to maxims, we lose Kant's concern with the *principles* on which we act. For this and other reasons, I restate my proposed revision as

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

Herman cannot claim, I believe, that this formula is a 'hybrid reconstruction', which is deeply un-Kantian. Kant himself refers to

> the idea of the will of every rational being as a will giving universal law. [595]

Herman's objections are not to my proposed formula, but to my way of *applying* this formula. She has the same objections to my way of applying Kant's own formula.

In stating these objections, Herman discusses the questions of why Kant's Formula condemns lying, and whether this formula implies that lying is always wrong. Herman compares two principles, of which one permits us to lie whenever that would be to our advantage, and the other permits us to lie only when some lie is necessary to save some innocent person's life. Like me, Herman believes that, when Kant's Formula is correctly applied, this formula condemns lying for our own advantage, but permits lying to save such a person's life. But Herman objects to my way of reaching this conclusion. She sums up my reasoning as follows:

> When advantage-lying is widespread, it undermines the trust conditions necessary for cooperative activity, itself a great good. Therefore, a principle

> of general permissiveness about lying would not be rational to will… But a principle that permitted lying when necessary to save wrongfully threatened lives would not be interfering with interests we have reason to protect and would have little or no undermining effect on trust. So advantage-lying is shown to be wrong; not all lying is wrong; and the rationale for the wrongness points not to the value of rational agency, but to the benefits of cooperation. In this way, the revisionist retains the Kantian (Contractualist) spirit and get a much more plausible moral view.

To my surprise, Herman rejects this way of applying Kant's Formula, which she claims to be too *Consequentialist*. She writes:

> The consequentialism figures in the revisionary account twice---in the values appealed to and in the treatment of the universality condition setting up a comparison between how we would fare were advantage-lying, as opposed to life-saving lying, permissible.

What Herman finds objectionable here is my appeal to certain values. On my account, she writes,

> since the values that inform rational willing are (for the most part) about what is non-morally best, the hybrid theory winds up having a strongly consequentialist cast.

Some possible outcome is non-morally best, in what I call the impartial-reason-implying sense, just when this is the outcome that, from an impartial point of view, everyone would have most reason to want, or to hope will come about. When some outcome would be in this sense *impersonally* best, that is often because of the ways in which this outcome would be *best for* particular people, in a similar reason-implying sense. When I appeal to these values, I am appealing to the facts that give us personal and impartial reasons to care about our own and other people's well-being, and to the facts give us other non-moral reasons to care about what happens.

There are two ways in which Herman might reject my appeal to these values and reasons. She might claim that

> (H) there are no such values, since no outcomes could be either impersonally good or bad, or good or bad for particular people, in these reason-implying senses.

Or she might claim that

> (I) though outcomes can be good or bad in these reason-implying senses, when we apply Kant's Formula or any other Kantian Formula, we should not appeal to such values or reasons.

Herman elsewhere makes some claims that seem to suggest (H). For example, she writes

> states of affairs are not possible bearers of value in Kantian ethics. [596]

But this remark is about *moral* value. As Herman writes elsewhere:

> Things that happen are not themselves morally good or bad, right or wrong: only willings are. [597]

When she discusses some outcome that involves 'loss and distress', Herman similarly writes

> There is no point of view from which the untoward outcome as such makes the world morally worse. [598]

We could all accept *these* claims. As Kant remarks when discussing the Stoics, it is not morally bad to be in pain. But pain is bad in the different, non-moral sense of being a state that we all have non-moral reasons to want not to be in. And as I claimed, outcomes can be non-morally good or bad, and good or bad *for* particular people, in such reason-implying senses. It is bad when an earthquake kills many people, though this event is not, like some mass-murderer, *morally* bad.

Herman seems to have similar beliefs about these kinds of value. For example, she writes:

> If everyone killed as they judged useful, we would have an unpleasant state of affairs. Population numbers would be small and shrinking; everyone would live in fear. These are bad consequences all right. [599]

She also writes that we could not rationally

> will a world where one's life can have no value in this reason-giving sense. [600]

If we accept some desire-based or aim-based subjective theory about reasons, we could not claim that we all have such reasons to care about our own and other people's well-being. But Herman seems to rejects such theories, and to assume that various facts can give us what I call value-based object-given reasons. [601]

Though Herman seems to believe that we can have this kind of reason to care about what happens, she claims that, when we apply Kant's Formula we should not appeal to such reasons. For example, when she describes my way of applying Kant's Formula, Herman writes that my reasoning would appeal

> not to the value of rational agency, but to the benefits of cooperation.

When Herman rejects this reasoning as too Consequentialist, she must mean that our reasoning should *not* appeal to the benefits of cooperation.

We can ask: Why not? When we apply Kant's Formula, we ask whether we could rationally will it to be true either that everyone accepts some maxim and acts upon it when they can, or that everyone believes such acts to be permissible. If such a world would be bad for us and other people, and we have reasons to care about our own and other people's well-being, these facts give us reasons not to will that this maxim be universal. When we ask what we could rationally will, why should we ignore such reasons? Why should we not appeal, for example, *both* to the value of rational agency *and* to the benefits of cooperation?

Kant himself, as I remarked, does *not* ignore such reasons. When he explains why lying is wrong, Kant writes that 'a lie. . . always *harms* another, even if not another individual, nevertheless humanity generally, inasmuch as it makes the source of right unusable'. [602] Consider next Kant's discussion of his imagined rich and self-reliant man, who has the maxim of not helping others who are in need. This man, Kant writes, could not rationally will that his maxim be a universal law,

> since many cases could occur in which one would need the love and sympathy of others, and in which, by such a law of nature arisen from his own will, he would rob himself of all hope of the assistance that he wishes for himself. [603]

Kant is appealing here, not to the value of rational agency, but to this man's reasons to care about his own future well-being. As Herman writes

> It is surely no crude mistake. . . to interpret this passage as making some kind of prudential appeal.

But Herman then argues that this claim *is* a mistake, since, when applying Kant's Formula, we should not appeal to reasons that are *prudential* in the sense of being concerned with our own future well-being.

Herman rightly rejects one bad argument for this conclusion. Schopenhauer suggests that, since Kant here appeals to prudential reasoning, Kant undermines his claim that we ought to do our duty for moral rather than prudential reasons.

That is not so. Kant does not argue that, if his imagined man helps other people in the actual world, that would in fact be better for this man, because he would thereby bring it about other people would help him. Kant makes the quite different claim that, if this man had the power to choose how everyone would act, he could not rationally choose to live in a world in which no one would ever help others. Kant would agree that, in the actual world, we do not always have prudential reasons to help others who are in need. On Kant's view, we ought to help others for moral reasons.

Herman gives a different argument for the claim that, when we apply Kant's Formula, we should not appeal to prudential reasons. If this is how we apply Kant's Formula, Herman claims, we may be unable to show that *everyone* ought to help others who are in need. There may be some rich and self-reliant people who *could* rationally will that the maxim of not helping be a universal law. In Herman's words:

> The problem then appears to be: can the argument in the example be construed in a way that makes it impossible for a rational agent to adopt the strategy of being willing to forgo help in order to keep his maxim of non-beneficence?
>
> if the reasoning is prudential, then it would also be appropriate to consider the likelihood of situations arising when he would prefer help more than he prefers the policy of non-beneficence. . . any person well situated in life and of a sufficiently self-disciplined temper might have good reason to feel that the price of increased security in having the help of others is too high.

The 'price' that Herman refers to here is the fact that, if we lived in a world in which everyone helps others who are in need, we would sometimes have to help others at some cost to ourselves. Herman continues:

> there seems to be no way . . . to show that people willing to tolerate risk have a duty to help others, if they would prefer not to help.
>
> To salvage the argument for beneficence then, it must be possible to show that such considerations cannot legitimately be introduced. As we have so far interpreted the argument, there seems to be no way to exclude them and so no way to show that people willing to tolerate risk have a duty to help others, if they would prefer not to help. [604]

This objection does not, however, show that we must *exclude* appeals to prudential reasons. This objection could show only that, in some cases, it may not be *enough* to appeal only to such reasons. [605]

When Herman tries to solve this problem, moreover, she does *not* exclude appeals to prudential reasons. According to the argument that Herman regards as too weak, because it may not apply to everyone, the costs of helping others would be likely to be much less than the benefits from being helped. Rather than disallowing this prudential argument, Herman tries instead to give a similar but stronger argument.

Herman first considers Rawls's proposed solution, which appeals to prudential reasoning from behind a veil of ignorance. If Kant's imagined man did not know that he was rich and self-reliant, Rawls claims, this man could not rationally choose to live in a world in which no one helped others who are in need. Herman rightly rejects this proposal, not because it involves prudential reasoning, but because Rawls's veil of ignorance abandons some of Kant's distinctive claims about moral reasoning.

Herman then suggests a way of applying Kant's Formula that makes no appeal to probabilities, or to the balance of likely costs and benefits. This argument claims that, even if we are rich and self-reliant, we could not rationally choose to live in a world of universal non-beneficence, in which no one helps others. No rational agent could will such a world, Herman writes

> if either of two conditions holds: (1) that there are ends that the agent wants to realize more than he could hope to benefit from non-beneficence and that he cannot bring about unaided or (2) that there are ends that it is not possible for any rational agent to forgo (ends that are in some sense necessary ends). [606]

Though Herman claims that this argument does not involve prudential reasoning, she means only that it does not appeal to *probabilities*, or to benefits that are merely likely. This argument does appeal to our reasons to care about our future well-being, as is shown by the phrase 'hope to benefit'.

Herman considers an objection to this argument, which appeals to an imagined Stoic who chooses to adopt only ends whose achievement could not possible require help from others. This imagined case, she argues, may be impossible, or incoherent, and she calls it 'a strength of Kant's argument that we are pushed to the edge of what we can imagine to find a potential exception'. [607]

If this argument succeeded, however, it would show only that, according to Kant's Formula, it is wrong *never* to help others who are in need. This would be far from a full defence of this formula. To find other objections to Kant's Formula, moreover, we are not 'pushed to the edge of what we can imagine'. There are, I argue, many actual cases in which Kant's Formula clearly fails.

The most important cases raise what I call the Non-Reversibility Objection. This objection can be summed up with a comparison to the Golden Rule. There are many wrong acts with which we benefit ourselves in ways that impose much greater burdens on others. As I wrote:

> The Golden Rule condemns such acts, since we would not be willing to have other people do such things to us. But when we apply Kant's formula to our acting on some maxim, we don't ask whether we could rationally will it to be true that *other* people do these things to *us*. We ask whether we could rationally will it to be true that *everyone* does these things to *others*. And we may know that, even if everyone did these things to others, *no one* would do these things to *us*.

To stay close to Kant's example, we can consider those rich people who act on the maxim 'Give nothing to the poor'. Kant's Formula condemns these people's acts only if they could not rationally will it to be true either that they and other rich people continue to give nothing to the poor, or that everyone, including the poor, believes that their giving nothing is morally permissible. Given the restrictions on the kinds of reason to which we can here appeal, we must admit, I argued, that these people *could* rationally will such a world. Similar claims apply to other wrong-doers, such as those men who benefit themselves by treating women as inferior, denying women certain rights and privileges, and giving less weight to women's well-being. These men could rationally will it to be true both that they and other men continue to treat women in this way, and that everyone, including women, believes their acts to be justified.

To answer this and similar objections, we cannot appeal to Herman's suggested non-probabilistic argument. Kant's Formula faces these objections because, when we apply this formula, we appeal to what the *agent* could rationally will. To avoid these objections, I believe, Kant's Formula should appeal instead to what everyone could rationally will.

We can now return to Herman's claims about my attempt to answer such objections to Kant's Formula of Universal Law. Herman objects to the way in which, when I apply both Kant's Formula and my proposed revision, I appeal to facts about what would be non-morally good or bad, and to our reasons to care abour our own and other people's well-being. My appeal to such values and reasons, Herman claims, makes my proposed Kantian Contractualism a 'hybrid reconstruction', which departs too far from the best elements in Kant's view. When we apply Kant's Formula, Herman writes, 'such considerations cannot be legitimately introduced'.

These claims are not, I believe, true. In the second half of her Commentary, Herman gives another brilliant demonstration of what can be achieved without appealing to claims about well-being. As we have seen however, when Herman applies Kant's Formula, she sometimes appeals to such claims. So does Kant, as is shown by some of the passages I quoted above. To give some other examples, Kant writes:

> if he lets his maxim of being unwilling to assist others. . . become . . . a universal permissive law, then everyone would likewise deny him assistance when he himself is in need. . . Hence the maxim of self-interest would conflict with itself if it were made a universal law. . . Consequently the maxim of beneficence towards those in need is a universal duty. [608]

And Kant said

> I cannot will that lovelessness should become a universal law, for in that case I also suffer myself. [609]

On Kant's view, Herman elsewhere writes, we cannot 'weigh' amounts of non-moral value, and we should reject 'principles that involving "counting heads"'. [610] But Kant writes:

> Then two of us suffer, though the trouble really (in nature) affects only one. But there cannot be a duty to increase the ills in the world. [611]

If we appeal to such claims about well-being, Herman writes, our theory cannot be Kantian. Anticipating Marx, Kant might have said 'Then I am not a Kantian'. [612]

CHAPTER 21 HOW THE NUMBERS COUNT

75 Scanlon's Individualist Restriction

Scanlon's Commentary starts with an illuminating discussion of Kant's Formula of Universal Law and Kant's views about rationality and reasons. Since I accept all of Scanlon's main claims, I shall add only two remarks. According to what Scanlon calls 'Kantian constructivism', claims about reasons must be grounded on claims about which attitudes are consistent with regarding ourselves as rational agents. Scanlon asks why we ought to reject this view, and appeal instead to what Scanlon calls 'true substantive claims about reasons'. We ought to appeal to such claims, I believe, because they are true. And for Kantian moral theories to succeed, they must appeal to substantive claims about reasons. It is not enough to appeal to claims about what we could will, or choose, in ways that are consistent with regarding ourselves as rational agents. Those claims would be too restricted, and too weak.

Scanlon then discusses my attempt to show that a revised version of Scanlonian Contractualism can be combined with Kantian Rule Consequentialism. Before responding to Scanlon's comments, I shall describe and defend my proposed revisions of Scanlon's view.

According to one statement of

> *Scanlon's Formula*: We are morally required to act in some way just when such acts are required by some principle that no one could reasonably reject.

Scanlon supposes that, in

> *Case One*, if Grey gave one of his organs to White, Grey would shorten his own life by a few years, but he would also give White many more years of life.

This case, as Scanlon points out, raises a 'difficulty' for his view. Most of us would believe that, though it would be admirable for Grey to give his organ to White, Grey is not morally required to make this gift. But if we accept Scanlon's Formula, this belief is hard to defend. This formula implies that

> (A) Grey is not required to make this gift if he could reasonably reject every principle that requires this act.[613]

If we accept (A), we cannot also claim that

> (B) Grey could reasonably reject every such principle because he is not required to make this gift.

These claims would go round in a circle, getting us nowhere. To defend our belief that Grey is not required to make this gift, we must suggest some other ground on which Grey could reasonably reject every principle that requires this act.

Scanlon makes several claims about what are reasonable grounds for rejecting some moral principle. According to what we can call the *Greater Burden Claim,* or

> *GBC*: 'it would be unreasonable. . . to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others. '[614]

Scanlon uses the phrase 'impose a burden' in a wide sense, which covers not only harming someone but also failing to give someone some possible benefit. If some principle required me, for example, to save some stranger's life rather than your leg, this principle would impose on you the burden of losing your leg. Suppose next that, in

> *Case Two*, I could use some scarce drug either to give Grey a few more years of life, or to give White many more years of life. Neither Grey nor White has any other claim to be given this drug.

Scanlon's view rightly requires me to use this drug to benefit White. As GBC implies, Grey could not reasonably reject every principle that required this act. Though such principles would impose on Grey the burden of losing a few years of life, any principle that did not require this act would impose on White the much greater burden of losing many years of life.

Case One involves the same possible benefits and burdens. Scanlon's GBC therefore implies that Grey could not reasonably reject every principle that required him to give his organ to White. As in *Case Two*, though such principles would impose a burden on Grey, any principle that did not require this act would impose a much greater burden on White. So Scanlon's view implies, implausibly, that Grey is morally required to shorten his life by giving his organ to White. [615]

There is another, more serious problem. White might appeal to some principle which permits or requires other people to take Grey's organ, without Grey's consent, and give it to White. GBC seems to imply that Grey could not reasonably reject this principle. But most of us would believe such an act to be very wrong.

Since it is GBC which raises these problems for Scanlon's view, we should ask whether Scanlon could reject this claim. The answer depends in part on whether Scanlon should revise his view in another, wider way.

76 Utilitarianism, Aggregation, and Distributive Principles

According to what we can call Scanlon's

> *Individualist Restriction*: In rejecting some moral principle, we must appeal to this principle's implications only for ourselves and for other *single* people.

In Scanlon's words:

> the justifiability of a moral principle depends only on *individuals'* reasons for objecting to that principle and alternatives to it. [616]

We can also call such reasons *personal grounds* for rejecting some principle. The strength of these grounds depends in part on how great the burdens are that this principle's acceptance would or might impose on us. This strength may also depend on certain other facts, such as how badly off we are, and whether we are responsible for the fact that either we or others will have to bear certain burdens. Some reasonable personal grounds for rejecting principles, Scanlon adds, may have nothing to do with our well-being. Such grounds might be provided, for example, by some principle's unfairness to us.[617] And any such list of grounds may be incomplete, since we may come to recognize other reasonable grounds for rejecting moral principles.

Scanlon's Individualist Restriction is given some support by one of Scanlon's most appealing ideas, that of justifiability to *each* person. Since we are asking which are the principles that *no one* could reasonably reject, we must consider each person's grounds for rejecting some principle, and we can plausibly claim that these grounds are provided by this principle's implications for *this* person.

Scanlon also defends this claim in another way. Like Rawls, Scanlon intends his Contractualism to provide 'a clear account of the foundations of non-Utilitarian moral reasoning'. [618] Act Utilitarians believe that it would always be right to impose great burdens on a few people, if we could thereby give small benefits to enough other people. In one of Scanlon's imagined cases,

> *Jones* has suffered an accident in the transmitter room of a television station. To save Jones from one hour of severe pain, we would have to cancel part of the broadcast of a football game, which is giving pleasure to very many

people. [619]

Within a single life, pain can be *hedonically outweighed* by pleasure. We might have decisive reasons, for example, to choose to endure one hour of pain for the sake of many hours of pleasure. This choice would benefit us, by giving us a positive net sum of pleasure minus pain. It makes no difference, Utilitarians believe, whether pain and pleasure come, not within a single life, but in different lives. On this view, it might be wrong for us to save Jones from his hour of pain. This act would be wrong if, by lessening the pleasure of the many watchers of the football game, we would reduce the total sum of pleasure minus pain. Scanlon rejects this Utilitarian conclusion, claiming instead that, whatever the number of people whose pleasure would be lessened, we ought to save Jones from his hour of pain. Many of us would agree.

Utilitarians reach such unacceptable conclusions, Scanlon suggests, because they mistakenly add together different people's benefits and burdens. By appealing to the Individualist Restriction, Scanlon writes, we can avoid such conclusions 'in what seems, intuitively, to be the right way'. [620] In his words:

> A contractualist theory, in which all objections to a principle must be raised by individuals, blocks such justifications in an intuitively appealing way. It allows the intuitively compelling complaints of those who are severely burdened to be heard, while, on the other side, the sum of the smaller benefits to others has no justificatory weight, since there is no individual who enjoys these benefits. . . [621]

On the simplest form of Scanlon's Individualist Restriction, benefits to different people cannot ever be morally summed. In applying Scanlon's Formula to any two conflicting principles, we should consider only the strongest personal objection that any one person would have to one of these principles, and the strongest objection that anyone else would have to the other principle. It makes no difference how many people would have these two strongest, conflicting objections, and we can ignore all other, weaker objections. Every such choice can thus be regarded as if it would affect or involve only two people. In *Scanlon's* phrase, *the numbers do not count*. [622]

Scanlon qualifies this view in two ways. He suggests that, when different possible acts would impose equal burdens on different people, numbers can break ties, since we ought to impose such burdens on as few people as we can. [623] Scanlon also suggests that, when one burden is not much smaller than another, the numbers count. To avoid these complications, we can discuss cases in which we could either save one person from some great burden, or save many other people from *much* smaller burdens.

Scanlon's Individualist Restriction is not, I believe, the right way to avoid unacceptable Utilitarian conclusions. Scanlon misdiagnoses how Utilitarians reach these 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 can suppose that certain people have painful diseases, and that as doctors who have scarce medical resources we must decide which of these people we shall treat. None of these people has any special claims, nor do they differ in any other morally relevant way. As before, people are *burdened* in the relevant sense if they fail to receive some possible benefit.

In some cases of this kind, if we don't intervene, some of the people whom we could benefit would be much worse off than the others. In such cases, we can say, the *baseline* is *unequal*. Suppose that, in *Case Three*, the only possible outcomes are these:

	Future days of pain	
	for Blue	for each of some number of other people
We do nothing	100	10
We treat Blue	0	10
We treat the others	100	0

If we do nothing, Blue will be much worse off than these other people, since Blue will suffer for ten times as long as each of them. Suppose next that each day of pain is an equal burden. Utilitarians would then claim that, if we could save eleven of these other people from their 10 days of pain, we ought to treat these people rather than Blue. We would thereby save these people from a combined total of 110 days of pain, which is a greater sum of benefits than the benefit to Blue of saving her from all of her 100 days of pain. Most of us would reject this Utilitarian claim, believing instead that we ought to save Blue from her great ordeal. We might even believe that we ought to save Blue from her 100 days of pain rather than saving *any* number of such other people from their much smaller burden of 10 days of pain.

Scanlon's Formula supports these beliefs. Given Scanlon's Individualist Restriction, Blue could reasonably reject every principle that required us to treat these other people, since this act would impose on Blue a burden that would be much greater than any burden that would be imposed on any other single person if

instead we treated Blue.

Though Scanlon's Formula gives a plausible answer here, it does not, I believe, support this answer in the right way. If we ought to treat Blue rather than these other people, that is not because we would be saving Blue from a much greater burden. It is because, if we don't save Blue from this burden, she would be much worse off than these other people, since she would suffer for many more days. To show this fact to be what matters, we can turn to a version of this case in which there is no such difference, so that the *baseline* is *equal*. We can also suppose that, rather than giving Blue a very great benefit, we could give equal though much smaller benefits to everyone. Suppose that, in *Case Four*, the only possible outcomes are these:

	Future days of pain	
	for Blue	for each of some number of other people
We do nothing	100	100
We do A	0	100
We do B	90	90

If we do nothing, Blue and the others would all be equally badly off, since they would all have 100 days of pain. If we do B, we would give equal benefits to all these people. According to Scanlon's Individualist Restriction, benefits to different people cannot be morally summed, so we ought again to do A, thereby saving Blue from all of her 100 days of pain. We would thereby give Blue a much greater benefit than we could give to any of the other people by saving this person from only 10 of her 100 days of pain. On Scanlon's view, it makes no moral difference how many of these other people we could save from 10 of their days of pain. We ought to give Blue her 100 pain-free days rather than giving 10 pain-free days to Blue and as many as a *million* of these other people.

These claims are clearly false. If we gave Blue her 100 pain-free days, we would not merely be failing to save the other people from a total of ten million days of pain. This vastly greater sum of pain would be suffered by people who would all, without our help, suffer just as much as Blue. We ought instead to give 10 pain-free days to each of these many people.

In cases of this kind, Scanlon's view conflicts with all plausible views about the distribution of benefits and burdens. According to one such view,

> *Telic Egalitarianism*: It would always be in one way better if benefits and burdens were more equally distributed between different people.

This view implies that, compared with Blue's being saved from all of her 100 days of pain, it would be better if Blue and nine other people were saved from 10 of their 100 days of pain. The same total sum of benefits would then be shared equally between Blue and these other people. Since there are no other morally relevant facts, this would be the outcome that we ought to produce. It might also better if a *smaller* sum of benefits were shared more equally between different people. But such cases raise questions that we can here ignore. Egalitarianism can also take a purely deontic form, which claims only that, in many cases, we ought to distribute benefits more equally between different people.

According to another, less familiar view, which we can call

> *the Telic Priority View*: It would always be in one way better if benefits came to people who are worse off.

This view also implies that, compared with Blue's being saved from all of her 100 days of pain, it would be better if Blue and nine other people were saved from 10 of their 100 days of pain. But this outcome would be better, not because there would be no inequality, but because more of these benefits would come to people who were worse off. Suppose that we first ensure that Blue will be saved from 10 of her 100 days of pain. On the Priority View, since the other people would then face a longer ordeal than Blue, we would do more good by giving 10 further pain-free days, not to Blue, but to any of these other people. Compared with reducing any of these people's burdens from 100 days of pain to 90, we would do less good by reducing Blue's burden from 90 days to 80, and even less good by making a further reduction from 80 to 70, and so on. [624] Since there are no other morally relevant facts, we ought to save Blue from only 10 of her days of pain, so that we can also give the same benefit to these nine other people. As before, this view could take a non-telic, deontic form, which claimed only that, in many cases, we ought to give priority to benefiting those who are worse off.

It may help to vary our example. Suppose that Blue and several other people are all aged 25, and have life-shortening medical conditions. With our scarce medical resources, we cannot treat all these people. In *Case Five*, the only possible outcomes are these:

	Blue will live to the age of	Each of some number of other people will live to
We do nothing	30	70

We treat Blue	70	70
We treat the others	30	75

Scanlon's view implies that we ought to give Blue her 40 more years of life, whatever the number of the other people to whom we could instead give 5 more years. If the number of the other people would be very large, this view would, I believe, be too extreme. But it would be fairly plausible to claim that we ought to give Blue her 40 more years of life rather than giving 5 more years to each of eight, twelve, twenty, or even more of these other people.

What makes this claim plausible, however, is the fact that, without her extra 40 years, Blue's life would be so much shorter than the lives of all these other people. As before, to show this fact's importance, we can change this feature of this case. We can again suppose that, rather than giving Blue her great benefit, we could give equal though much smaller benefits to everyone. Suppose that, in *Case Six*, the only possible alternatives are these:

	Blue will live to the age of	Each of some number of other people will live to
We do nothing	30	30
We do A	70	30
We do B	35	35

On Scanlon's view, we ought to give Blue her 40 more years of life rather than giving 5 more years to Blue and to as many as a *million* of these other people. As before, that is clearly false. And what makes it false is not merely that, compared with 40 more years, 5 million more years of life would be a vastly greater total sum of benefits. These benefits would also be more fairly distributed between different people. It would be clearly better if, rather than Blue's living to the age of 70 rather than 30, Blue and a million other people each lived to 35 rather than 30. This second outcome would be better, I believe, even if these 5 extra years came to as few as seven, or six, or perhaps even fewer of these other people.

Because Utilitarians believe that the goodness of outcomes depends only on the total net sum of benefits, they deny that it would be in itself better if benefits were more equally distributed, or if benefits came to people who were worse off. Though this view is, I believe, mistaken, Utilitarians are at least neutral between different patterns of distribution. In some cases, as we have just seen, Scanlon's Formula favours the *less* equal distribution. In such cases, this formula has a built-in bias against equality, and against giving priority to benefiting those who are worse off. That is not what

Scanlon intends. And as Scanlon would agree, we ought to reject these conclusions. [625] In *Cases Four* and *Six*, rather than giving Blue her great benefit, we ought to give a greater sum of benefits that would be shared equally between Blue and many other people who are just as badly off.

These cases show, I believe, that Scanlon ought to drop his Individualist Restriction. [626] It might be suggested that, even if Scanlon kept this restriction, he could revise his view in some other way. But it is clearly the Individualist Restriction which is making Scanlon's Formula go astray. Suppose that, in a different version of *Case Six*, we could either enable Blue to live to 70 rather than 30, or enable only *one* other person to live to 35 rather than 30. Scanlon's Formula would then rightly imply that we ought to give Blue her much greater benefit. But if instead we could enable a hundred or a million other people to live to 35 rather than 30, that would be what we ought to do. For Scanlon's Formula to give the right answer in these cases, Scanlon must allow that these many other people could reasonably reject any principle that did not require us to give these benefits to them. Since the benefits to *each* of these people would be much smaller than the benefit that we could give Blue, these people must be allowed to appeal to the fact that, as well as being as badly off as Blue, *they together* would receive a much greater total sum of benefits, in significant amounts of five years per person. Each of these people could appeal to this fact, speaking on behalf of this group.

As these cases also show, it is not only Utilitarianism that gives weight to the numbers of people who might receive benefits or burdens. So do all plausible distributive principles. We should reject Utilitarianism, not because this view gives weight to numbers, but because it ignores distributive principles.

Scanlon claims that his Individualist Restriction

> is central to the guiding idea of Contractualism, and is also what enables it to provide a clear alternative to Utilitarianism. [627]

This claim implies that, if Scanlon dropped this restriction, Scanlon's view would cease to provide a clear alternative to Utilitarianism. But that is not so. Even without the Individualist Restriction, Scanlonian Contractualism could provide such an alternative.

Here is one of the many ways in which that is true. According to what we can call

> *the Contractualist Priority View*: People have stronger moral claims, and stronger grounds to reject some moral principle, the worse off these people are.

Unlike the Telic Priority View, this view is not about the goodness of outcomes. In his earliest statement of his theory, Scanlon appealed to this view. When we consider a principle, Scanlon wrote,

> our attention is naturally directed first to those who would do worst under it. This is because if anyone has reasonable grounds for objecting to the principle it is *likely* to be them.[628]

In his book, however, Scanlon applies this view only to certain cases, and he gives little priority to the claims of people who are worse off. [629] As well dropping his Individualist Restriction, Scanlon ought to return, I believe, to a stronger version of the Contractualist Priority View.

With these two revisions, Scanlonian Contractualism could be successfully applied to all of the cases that we have been discussing. In these cases, we could either save a single person from some great burden, or save many people from much smaller burdens. Scanlon claims that, in such cases, the numbers don't count, so that we ought to save the single person from her great burden. When applied to some of these cases, this claim may seem acceptable. We can agree that, in *Case Three*,

> (A) we ought to save Blue from her 100 days of pain rather than saving each of eleven other people from all of their 10 days of pain.

But Scanlon's view also implies that, in *Case Four*,

> (B) we ought to save Blue from her 100 days of pain rather than saving Blue and a million other people from 10 of their 100 days of pain.

And (B) is clearly false. Instead of claiming that the numbers don't count, Scanlon should say that people have stronger moral claims, and stronger grounds to reject some principle, the worse off these people are. This version of Scanlon's view would still rightly imply (A). Because Blue would suffer much more than each of the eleven other people, Blue has a much stronger claim to be saved from most of her days of pain. And this view would not mistakenly imply (B). Since these million other people are as badly off as Blue, facing the same great ordeal, these people's claims to be saved from any of their days pain are as strong as Blue's. So they could reasonably reject any principle that did not require us to save them from a total of ten million of their days of pain.

Similar claims apply to *Cases Five* and *Six*. This revised version of Scanlon's view would also have more plausible implications in many other kinds of case. That is in part because, unlike the claim that benefits to different people cannot be morally summed, the Contractualist Priority View can respond to differences of degree. On this view, when we compare the strength of people's grounds for rejecting some moral principle, we ought to give slightly more weight to the moral claims of people who are

slightly worse off, and much more weight to the claims of people who are much worse off.

If Scanlon drops his Individualist Restriction, he might appeal instead to a similar but weaker view. Scanlon suggests one such view, according to which numbers count only when we are comparing benefits and burdens that are *close enough* in size. But this *Close Enough View* would have unacceptable implications. Suppose this view claims that, for some benefit to be morally outweighed by many lesser benefits, these other benefits must be at least a quarter as great. Suppose next that, in

> *Case Seven*, we could give extra years of life to people who would otherwise die at 30. We could either
>
> (1) give 40 more years to Blue,
>
> or
>
> (2) give 15 more years to each of a thousand other people,
>
> or
>
> (3) give 5 more years to each of a million other people.

On this view, the great benefit to Blue would be outweighed by the lesser benefits to the thousand other people, since these benefits are close enough in their size. The benefits to the thousand would in turn be outweighed by the benefits to the million, since these benefits are also close enough. But Blue's great benefit would *not* be outweighed by the benefits to the million, since these benefits are *not* close enough. So the Close Enough View implies that we ought to do (2) rather than (1), and that we ought to do (3) rather than (2), and that we ought to do (1) rather than (3). Whatever we do, we would be acting wrongly, since we ought to have done something else instead. Even if there might be cases in which we could not avoid acting wrongly, that is not plausible here. And it is clear that we ought to do (3) rather than doing either (1) or (2).

Rather than appealing to the Close Enough View, Scanlon's claim should at most be that significant benefits and burdens cannot be morally outweighed by any number of other benefits and burdens that are insignificant, or trivial. He might, for example, claim that

> (C) we ought to give one person one more year of life rather than lengthening any number of other people's lives by only one minute,

and that

> (D) we ought to save one person from a whole year of pain rather than saving any number of other people from only one minute of similar pain.

Though these claims are very plausible, they can have unacceptable implications. A year contains about half a million minutes. Suppose that, in

> *Case Eight*, we are in a community of just over a million people, each of whom we could benefit once in the way described by (C). Each of these acts would give to one of these people half a million more minutes of life rather than giving one more minute of life to each of the million other people.

Since these effects would be equally distributed, these acts would be worse for everyone. If we always acted in this way, we would give everyone only one more year of life. If instead we always gave all the other people their extra minutes, we would give everyone a total of *two* more years of life. Suppose next that these people are often in pain, and that we could benefit each person once in the way described by (D). Each of these acts would save one of these people from half a million minutes of pain rather than saving each of the million other people from one such minute. As before, these acts would be worse for everyone. If we always acted in this way, we would save all these people from only one rather than two years of pain.

There are several ways in which claims like (C) and (D) can seem to be obviously true. Most of us are bad at judging the significance of large numbers. We may assume that, if it matters little whether one person would bear some burden, it also matters little whether a million people would bear such burdens. We may also assume that, if some people would bear much greater burdens than others, or would lose much greater benefits, these are the people who would be worst off. But that may not be true. And when it isn't, one great loss may be morally outweighed by many small benefits. Suppose that, if I gave a million dollars to some aid agency, my gift would be divided equally between ten million of the world's poorest people, so that each of these people would get only ten cents. If I was giving away most of my wealth, the burden to me of losing a million dollars would be much greater than the average benefit that ten cents would give to each of these other people. But these million benefits would *together* be much greater than my burden. Since this sum of benefits would both be much greater, and would come to people who are much worse off than me, it is morally irrelevant that the average benefit to each of these people would be very small. My million dollars, even when giving these people such small benefits, would do much more good.

A third mistake is to consider only single acts. Some acts give ourselves significant benefits in ways that impose tiny burdens on very many other people. That is true, for example, of many of the acts that add to the pollution of many people's air, food, or water. When we consider any one such act, the tiny effects on the many other people

may seem trivial. It may seem not to matter if such an act imposes costs on others of less than ten cents, or reduces the life-expectancy of others by less than one minute. But when many people act in such ways, these small effects add up. And when such effects are roughly equally distributed, these acts are worse for almost all of the affected people. In the world as it is now, such acts together impose great burdens on many people.

Though we should not always ignore trivial benefits and burdens, we are often justified in doing that. That might be true in Scanlon's case in which, to spare Jones from an hour of severe pain, we would have to interrupt the pleasure of millions of watchers of a football game. It might be reasonable for Jones to reject any principle that would require or permit us to let him suffer his hour of pain. The million watchers might object that, though each of them would lose little, they together would lose a sum of pleasure that would hedonically outweigh Jones's hour of pain. But Jones would be much worse off than all these people. Given this fact, Jones might plausibly reply, his claim to be spared his pain morally outweighs their combined claims.

We can now turn to a different question. When a great benefit to one person might be morally outweighed by several lesser benefits to other people, we must ask whether the importance of these benefits would be *proportional* to their size. That would be true, for example, if some benefit to one person would have the same importance as two benefits to other people that were half as great.

Scanlon suggests that, rather than saving one person's life, we ought perhaps to save a million people from total paralysis.[630] For most people, becoming completely paralysed would be at least a twentieth as bad as dying. If the moral importance of these burdens were proportional to their size, one person's death would be morally outweighed by as few as thirty or forty people's becoming completely paralysed. Since Scanlon chooses the much larger number of a million people, he seems to give these lesser burdens much less weight. On what we can call this

> *Disproportional View*: The moral importance of lesser benefits and burdens is less than proportional to their size.

This view is a weaker version of the Individualist Restriction. On that restriction, a great benefit or burden to one person cannot be morally outweighed by any number of lesser benefits or burdens to other people. On the Disproportional View, this great benefit or burden could be morally outweighed, but the lesser benefits or burdens should not be simply added together, as Utilitarians claim. Though such lesser benefits or burdens can be added together, they should be given disproportionately less weight.

Scanlon ought, I believe, to reject this view. Though a great burden to one person should often be given disproportionately greater weight, that is true, I believe, only

when and because this burden would make this person much worse off than other people. When this person would *not* be worse off, the Disproportional View is mistaken. Suppose that, in

> *Case Nine*, we could either
>
> > (1) save Blue from all of her 100 days of pain
>
> or
>
> > (2) save each of ten other people from 10 of their 100 days of pain.

Suppose next that, because each day of pain is an equal burden, (1) would give a benefit to Blue that is ten times as great as the benefits that (2) would give to each of these ten other people. If the importance of these lesser benefits was less than proportional to their size, we ought to give Blue her 100 pain-free days. But the *opposite* is true. It is Blue's *greater* benefit whose moral importance is *less* than proportional to its size. As the Priority View claims, benefits have less moral weight when they come to people who are worse off. Compared with the claims of the other people to have their days of pain reduced from 100 to 90, Blue would have a weaker claim to have her days of pain reduced from 90 to 80, an even weaker claim to have a further reduction from 80 to 70, and so on. That is why, rather than giving Blue her 100 pain-free days, we ought to give 10 pain-free days to as few as nine, or eight, or even fewer of these other people.

In some cases, as Larry Temkin suggests, there is an argument the other way. Temkin claims that, though we always have more reason to spread *burdens* over many different people, we may sometimes have reasons to concentrate *benefits*, by giving them all to a single person. In *Case Seven*, for example, we may have a special reason to give Blue her extra 40 years of life, since that would allow at least one person to live a full life. Temkin here appeals to what we might call a *qualitative* reason to give benefits to a single person.

Though we may sometimes have such reasons, Temkin's view is different from and does not support the Disproportional View. Consider, for example,

> *Musical Chairs*: A hundred people will later be at a hundred levels of well-being. There are only two possibilities:
>
> > (A) *Person One* is at level 1, *Person Two* at level 2, *Person Three* at level 3, and so on.
> >
> > (B) Person One is at level 100, and everyone else is one level lower down.

On the Disproportional View, we ought to choose (B). If greater gains and losses had an importance that was more than proportional to their size, the single great gain to Person One of being ninety nine levels higher would clearly morally outweigh the ninety nine small losses of the other people. That is not plausible. Person One has no claim to be at the top.

Scanlon, I conclude, should not appeal to any weaker version of his Individualist Restriction. If Scanlon appeals instead to a strong version of the Contractualist Priority View, his view would provide a clear alternative to Utilitarianism, and would avoid all of the objections that we have been considering.

We can now return to an earlier objection. Remember that, in

> *Case One*, if Grey gave one of his organs to White, Grey would shorten his own life by a few years, but he would also give White many more years of life.

There is no other way, we can add, in which White's life could be saved, since Grey is the only other person who has an organ of the right tissue type. As we have seen, Scanlon's present view implies that Grey ought to shorten his life in this way, since Grey could not reasonably reject every principle that required him to give his organ to White. This case raises a problem, Scanlon writes, because he is inclined to believe that Grey is *not* required to make this gift. That is also what most people would believe.

As I have said, there is another, more serious problem. If some principle requires Grey to give his organ to White, this principle could also claim that Grey has a right to decide what happens to his body. Grey would then have a right to act wrongly, by deciding not to give his organ to White. But we can next consider a more extreme principle which *denies* that Grey has such a right, since this principle permits or requires other people to take Grey's organ, without Grey's consent, and give it to White. This principle conflicts even more deeply with most people's moral beliefs.

Scanlon's Formula would support these beliefs if Grey could reasonably reject this principle. When discussing a similar case, Scanlon writes

> It is not unreasonable to refuse to regard one's own life and body as 'on call', to be sacrificed whenever it is needed to save others who are at risk. [631]

As we have seen, however, Scanlon also claims

> GBC: It would be unreasonable to reject some principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others.

If we accept this claim, it may be hard to argue that Grey could reasonably reject every principle that permitted or required other people to take Grey's organ, without Grey's consent, and give it to White. Even if some other people acted in this way, Grey would lose only a few years of life, and that is a much smaller burden than the many years of life that, without Grey's organ, White would lose. And if Grey could not reasonably reject this principle, Scanlon's Formula would imply that it would be right for other people to take Grey's organ without Grey's consent and give it to White. Since that is much harder to believe, this implication would provide a much stronger objection to Scanlon's view.

It might be suggested that, since Grey has a right to decide what happens to his body, Grey could reasonably reject every principle that permitted others to take his organ without his consent. But in claiming that Grey has this right, we would be claiming that it would be wrong for others to act in this way. And when we are asking what Scanlon's Formula implies, we cannot appeal to our beliefs about which acts are wrong. We can appeal to these beliefs only at a later stage, when we are deciding whether, given its implications, we ought to accept this formula.

There is, however, another way in which, when we apply Scanlon's Formula, we might defend the claim that Grey has a right to decide what happens to his organ. If Scanlon drops his Individualist Restriction, as I have argued that he should, he could also reject GBC. According to this revised version of Scanlon's view, we could reasonably reject some principles by appealing to the combined force of the grounds for rejection that we and other people *together* have. We might then claim that we could reasonably reject any principle that permitted or required others to take Grey's organ without Grey's consent and give it to White. We all have reasons to want not to live in a world in which, when people in Grey's position refuse to give their organs, these people are hunted down by the police, and have their organs taken from them by force. Each of us would know that there would be only a small chance that we ourselves would be treated in this way. Given this fact, our reasons to want not to live in such a world would be *individually* much weaker than White's reason to want not to lose many years of life. But it might be true that *we together* have stronger grounds for rejecting any principle that would permit or require some people's organs to be forcibly removed and given to others.

It may be objected that, though we might later be in Grey's position, and would then lose a few years of life if some organ were forcibly taken from us, we would be just as likely to be in White's position, and we would then gain many more years of life if someone else's organ were given to us. Since our possible benefit in White's position would be much greater than our possible loss in Grey's position, it may seem that we could *not* reasonably reject every principle that permitted or required such acts. We could plausibly reply, however, that our grounds for rejecting these principles would not be provided only by the ways in which the acceptance of these principles would affect our own and other people's life-expectancies. Since such cases would be rare,

these effects would be small. If in all such cases some people's organs would be forcibly reallocated, everyone's predictable life-expectancy might rise by only a few hours or minutes. Our reasons to want such benefits might be clearly outweighed by our reasons to want not to live in a world in which the police hunt some people down and take their organs by force.

Here is another, partly similar question. When we know that the lives of certain people are in danger, as would be true, for example, if some group of miners are trapped underground, we have reasons to want great efforts to be made to save these people's lives. Some economists point out that we would do more to increase people's life-expectancy if, rather than spending huge sums on trying to save known particular people in such emergencies, we spent this money on more cost-effective safety measures that would prevent a greater number of statistically predictable future deaths. But we could reasonably deny that this fact is morally decisive. We have strong reasons to want great efforts to be made to save the lives of known particular people who are in danger. By making or supporting such efforts, for example, we reaffirm and express our solidarity with, and concern for, everyone in our community. That is less true of acts that merely prevent the statistically predictable future deaths of unknown people.

We have similar reasons to want it to be true that no one would be hunted down and have their organs removed by force. Though such acts would be done to save the lives of certain known particular people, these acts would produce much anxiety, conflict, and mistrust. We would have to admit that, compared with White's reasons to want to have many more years of life, and the similar reasons of those few other people who would be in White's position, the rest of us would have only weaker reasons to want to avoid such anxiety and mistrust. But even if these reasons were individually much weaker, the combined force of all these reasons would, I believe, give us reasonable grounds to reject any principle that required or permitted people's organs to be taken from them by force. So, if Scanlon dropped his Individualist Restriction, he could answer the objection that his view requires or permits such acts.

We can next ask whether, if Scanlon drops his Individualist Restriction and his Greater Burden Claim, he could also argue that Grey could reasonably reject any principle which required him, in *Case One*, to *give* his organ to White. This principle allows that Grey has the right to decide what happens to his body, and the right to act wrongly by refusing to give his organ to White. Given this fact, Scanlon could not reject this principle with the claims that I have just made. If we all accepted this principle, no one would be hunted down and have their organs removed by force. We might claim that we all had reasons to want not to be morally required, if we were in Grey's position, to give up a few years of life. But we would have to admit that, if we were in White's position, we would all have stronger reasons to want to be given many more years of life. [632]

There may, however, be other grounds on which we could reasonably reject this principle. We can reasonably reject some principles, Scanlon claims, on grounds that do not appeal only to the size of the burdens that these principles would impose on us or others, and to our level of well-being, or to claims about fairness. Of such other grounds, some might appeal to certain facts about human nature. Though most of us could follow moral requirements not to kill or seriously injure other people even when such acts would save our own lives, most of us would find it very hard to give up several years of life, merely to add many more years to some stranger's life. We might claim that, given these and similar facts, it is unreasonable to expect or require people to make this kind of sacrifice for strangers. In making such claims, we would not be violating the Moral Beliefs Restriction, since we would not be appealing to the belief that no one is morally required to make this kind of sacrifice. We would instead be claiming that these facts about human nature provide reasonable grounds for rejecting principles that require such acts. [633]

CHAPTER 22 SCANLONIAN CONTRACTUALISM

77 Scanlon's Claims about Wrongness and the Impersonalist Restriction

There are, I believe, two other ways in which Scanlon should revise and thereby strengthen his version of Contractualism.

In his book, Scanlon claimed that, rather than describing the facts that can *make* acts wrong, his theory gives an account of wrongness itself, or of *what it is* for some act to be wrong. This claim, I have argued, was a mistake. [634] According to one statement of

> Scanlon's Formula: An act is wrong just when such acts are disallowed by some principle that no one could reasonably reject.

If Scanlon was here using 'wrong' in a Contractualist sense, to mean 'disallowed by such an unrejectable principle', he could truly claim that his formula gives an account of this Contractualist kind of wrongness, or of what it is for acts to be wrong in this Contractualist sense. But Scanlon's Formula would then be a concealed tautology, whose open form would be

> SF2: An act is disallowed by some principle that no one could reasonably reject just when such acts are disallowed by such an unrejectable principle.

We could all accept this trivial claim, whatever our moral beliefs. Scanlon's claim should instead be that, if some act is disallowed by such an unrejectable principle, this fact makes this act wrong in one or more other, non-Contractualist senses. Scanlon might for example claim

> SF3: When some act is wrong in this Contractualist sense, that makes this act wrong in the justifiabilist, blameworthiness, and reactive-attitude senses.

These four senses of 'wrong' are all definable abbreviations of longer phrases. So this version of Scanlon's Formula could be more fully stated as

> SF4: 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. [635]

We can turn next to another of Scanlon's claims about what are reasonable grounds for

rejecting moral principles. According to what we can call Scanlon's

> *Impersonalist Restriction*: In rejecting some moral principle, we cannot appeal to claims about the impersonal goodness or badness of outcomes.

In Scanlon's words,

> impersonal values are not themselves grounds for reasonable rejection.[636]

All reasons for rejecting principles, Scanlon also claims, must be *personal*. Though Scanlon does not explicitly say that we cannot appeal to claims about the impersonal goodness of outcomes, that is implied by these other claims. Of these who reject such appeals, some claim that there is no sense in which outcomes can be impersonally good or bad. That is not Scanlon's view. Scanlon believes both that outcomes can be good or bad in the impartial-reason-implying sense, and that we can have strong reasons to try to produce or prevent such outcomes. [637]

Scanlon gives, as one example, reasons provided by the suffering of animals. He writes

> like the pain of humans, the pain of non-human animals is something we have reason to prevent and relieve, and failing to respond to this reason is a moral fault.

Scanlon then imagines someone saying:

> If there are impersonal reasons of this kind, why should they not count as possible grounds for reasonably rejecting principles?

He replies:

> In answering this question, it is important to bear in mind the limited range of the part of morality we are trying to characterize. The Contractualist formula is meant to describe one category of moral ideas: the requirements of 'what we owe to each other'. Reasons for rejecting a principle thus correspond to particular forms of concern that we owe to other individuals. By definition, impersonal reasons do not represent forms of such concern. [638]

When Scanlon claims that certain acts are *owed to others*, he means that failing to act in these ways would be wrong in his Contractualist sense, because there is some principle requiring such acts that no one could reasonably reject. Since Scanlon himself defines this Contractualist sense, he is entitled to claim that, when we ask which acts are in this sense wrong, we should not appeal to impersonal reasons, since by definition such reasons are irrelevant. But Scanlon now claims that, when acts are in this sense wrong, that makes these acts wrong in other, non-Contractualist senses. And Scanlon could not say that, when ask which acts are wrong in these other senses, claims about what is

good or bad in the impartial-reason-involving sense are *by definition* irrelevant.

Scanlon also suggests that, when we ask what we owe to others in his Contractualist sense, we can appeal to the importance to us of being able to respond to certain impersonal values. For example, we could reasonably reject some principle that required us to keep some fairly trivial promise rather than saving some animal from great pain. [639] As Scanlon points out, however, what we owe to others sometimes conflicts with impersonal values. And when we ask which acts are wrong in non-Contractualist senses, we could not defensibly claim that what we owe to others always has priority over such values.

Consider for example some *Retributive Principle* which requires us to give criminals the punishment that they deserve, even when such punishment would benefit no one. When we appeal to Scanlon's Formula, this principle is hard to defend. Criminals might reasonably object that such punishment would be bad for them and good for no one. We *owe it to them*, they might claim, not to punish them in a way that benefits no one. Scanlon would reject this Retributive Principle, I believe rightly. But Retributivists might reply that it would be in itself good if people get the punishment that they deserve. In rejecting this reply, Scanlon might claim that what we owe to others has priority over such facts about the goodness of outcomes. But that, I believe, would not be an adequate reply. We must reject the Retributive Principle in some other way, by denying that deserved punishment is in itself good, or by arguing that no one could deserve to suffer.

Since what we owe to others cannot be claimed to have absolute priority over facts about the goodness of outcomes, Scanlon's view could take two forms. If Scanlon keeps his Impersonalist Restriction, he might retreat to the view that, when some act is wrong in his Contractualist sense, that makes this act *prima facie* wrong in other, non-Contractualist senses. Such acts would be wrong unless they could be justified by appeals to claims about the goodness of outcomes. On this version of Scanlon's view, his formula would claim to describe only one of the facts that can make acts wrong in other senses. This version of Scanlon's view might seem disappointingly weak. But that might not be true. Scanlon might be able to defend the claim that, when acts are wrong in his Contractualist sense, that very often makes these acts wrong in other senses. And Scanlon's Formula might condemn most wrong acts. This formula might then describe one of the most important facts that can make acts wrong, and in a way that helps to explain why many other, more particular facts can also make acts wrong.

Suppose next that Scanlon drops his Impersonalist Restriction. On this version of Scanlon's view, when we claim that we could reasonably reject some principle, we are allowed to appeal to our beliefs about the goodness of outcomes. Given this revision, Scanlon might make the bolder claim that acts are wrong in other senses *just when* they are wrong in Scanlon's Contractualist sense. If that were true, Scanlon's

Contractualism would unify, and help to explain, all of the more particular facts that can make acts wrong. That gives Scanlon a strong reason to make this bolder claim.

78 The Non-Identity Problem

Scanlon has other reasons, I believe, to drop his Impersonalist Restriction. When Scanlon asks what we owe to others, he intends these *others* to include all future people. In his words:

> contractualism provides no reason for saying that people who do not now exist but will exist in the future have no moral claims on us. . .

He also writes: 'a restriction to presently existing human beings seems obviously too narrow'. [640] In deciding what we owe to future people, we must answer some questions that Scanlon does not discuss. So I shall now discuss these questions, returning only later to Scanlon's theory.

When our acts will affect certain people, it may be morally irrelevant that these people do not yet exist. If I leave some broken glass in a wood, and ten years later a five year old child is injured by this glass, my negligence may straightforwardly harm this child. It may be true that, if I had not left this broken glass where I did, this child would have later walked out of this wood unharmed. If that is true, my act would be just as wrong even though, when I acted, this child did not exist.

Suppose next that we must choose whether our community will continue to deplete certain scarce unrenewable resources, or continue to overheat the Earth's atmosphere. If we choose

> *Depleting* or *Overheating*, these policies would raise the quality of life of existing people, but the long-term effects, more than a century from now, would significantly lower the quality of future people's lives.

Such bad effects, we may assume, are like the bad effects that our policies might have on presently existing people. As Scanlon writes, 'It matters that there are, or will be, people out there with lives that will be affected by what we do.' [641]

There is, however, a problem here that is often overlooked. As well has having effects on the quality of future people's lives, our acts and policies may affect *who it is* who will later live. Which children we have depends on the slightest details of our private lives. Many of our acts affect such details in our own and other people's lives, and these effects spread, like ripples in a pool, over more and more lives. Unlike ripples, moreover, these effects never fade away. Over time, there will be more and more people of whom it is true that, if we had acted differently, these people would never

have been conceived. If the motor car had not been invented, for example, it is likely that none of the readers of this book would ever have existed. When we choose whether to continue policies like *Depleting* or *Overheating*, our choice may affect the identity of most of the people who will live more than a century from now. For these reasons, we can often know 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,

but 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 should ask whether and how (B) makes a difference. I have called this *the Non-Identity Problem*. [642]

Some people believe that

> (C) 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.

On this *Narrow Person-Affecting View,* even if such acts would greatly lower the quality of life in the further future, we have no reason not to act in these ways.

Most of us would rightly reject this view. We would believe that

> (D) it would be in itself worse if some of the lives that will be lived will be less worth living,

and that

> (E) we have reasons not to act in ways that would have such effects, and if these effects would predictably be very bad, and we could avoid them at little cost to ourselves, such acts would be wrong.

There are now two possibilities. On one view,

> (F) it makes no difference whether, because these future lives would be lived by the same people, these acts would be worse for these people.

We can call this *the No Difference View*. On what we can call

> *the Two-Tier View*: This fact does make a difference. Though we always have 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 Non-Identity Problem must be either practically or theoretically important. If the Two-Tier View is true, this problem is practically important, since our reasons and our obligations would in part depend on whether our acts would be worse for future people. If the No Difference View is true, this problem is theoretically important, since many moral theories imply that this view cannot be true. On these theories, it must always make a difference whether our acts would be worse for people.

In discussing these views, it will help to define a new phrase. Suppose that *Jane*, a 14-year-old girl, declares that she intends to have a child. In trying to persuade Jane to wait, we might say:

> It would be worse for your child if you have him now, while you are so young. If you have your child later, that would be better for him, since you would be able to give him a better start in life.

When we make such remarks, we may not be using the words 'your child' and 'him' to refer to a particular person. Suppose that Jane has a child now, whom she calls *Johnny,* and whom she fails to bring up well. We may know that, if Jane had waited before having her first child, that would not have been better for Johnny, since Johnny would never have existed. It would have been a different child to whom Jane would have later given a better start in life. Such uses of 'your child' and 'him' refer, not to a particular person, but to what we can call a *general person*. This phrase is merely an abbreviation. Like *the Average American*, a general person is not a person. A general person is a large group of possible people, one of whom will be actual. Things would go worse for the general person who is *Jane's first child* if the particular person who is actually Jane's first child has a life that is less worth living than the life that would have been lived by the different particular person who, if Jane had waited, *would* have been Jane's first child.

We can now say that, on the No Difference View, we have equal reasons to avoid doing what would be worse either for particular people, or for general people. On the Two-Tier View, we have stronger reasons to avoid doing what would be worse for particular people. We can here suppose that, on this view, these reasons would be twice as strong, so that, compared with benefits or burdens to particular people, benefits or burdens to general people matter morally only half as much. Other versions of the Two-Tier View would give either more or less priority to the interests of particular people.

When I consider policies like *Depleting* or *Overheating*, I accept the No Difference View. We always have reasons, I believe, not to act in ways that would lower the future quality of people's lives, and these reasons would be just as strong whether or not, because these lives would be lived by different people, these acts would be worse for any

particular people. When other people first become aware of the Non-Identity Problem, many respond like me, by accepting the No Difference View. After further thought, however, some of these people turn to the Two-Tier View.

In asking which view we ought to accept, it will help to consider some other cases. Suppose that, in

> *the Two Medical Programs,* we are doctors who must make decisions about the future policies of some National Health Service. We have planned two screening programs. In Program A, millions of women would be tested during pregnancy, so that we can identify those women who have a certain rare disease. By curing these women, we would prevent their disease from causing their unborn children to have some life-shortening condition. In Program B, millions of women would be tested when they intend to have a child, so that we can identify those women who have some other rare disease. By curing these women, we would prevent their disease from causing any children that they conceive to have a similar life-shortening condition. Since these women would be warned to postpone having a child until they had been cured, this delay would lead them to conceive different children.

Suppose next that, because our Government cuts Health Service funds, we must cancel one of these programs, and we must choose between them. We can predict that these programs would achieve results in as many cases. If we carry out either program, we would enable the same number of women to have a child who does not have some life-shortening condition. These would be different women, on the two programs. But since the numbers would be the same, the effects on these women and on most other people would be morally equivalent. If there is a moral difference between these programs, this difference must depend on how these programs would affect these children.

In considering these effects, we need not ask what is the moral status of a foetus or unborn child. Nor do we need to ask whether we have greater obligations to existing people than we have to future people. We can suppose that it would take at least a year before either medical program could begin, so that, when we choose between these programs, none of these future children has yet been conceived. And all of the children who will be conceived will be born and become adults. So our questions are about how our choice between these programs might affect these future people. We can also suppose that these people's lives, even if they would be shorter than most people's lives, would be happy, and well worth living.

This example could filled out in different ways. Suppose first that, in *Case One*:

> If we choose Program A, a thousand people would be conceived who would live for 70 rather than 50 years.

> If we choose Program B, a thousand people would be conceived who would live for 70 years, rather than a thousand different people who would live for 50 years.

On the No Difference View, these programs would be equally worthwhile. Though Program A would benefit particular future people, and Program B would benefit general people, these two kinds of benefit matter morally just as much. On the Two-Tier View, we ought to choose Program A. This program would give to a thousand particular people the benefit of an extra 20 years of life. Program B would give this benefit only to as many general people, and such benefits matter less.

Suppose next that, in *Case Two*, the predictable effects would be in one way different. If we cancel Program B, the people who would be conceived would live for only 40 years, so this program would give to a thousand general people the greater benefit not of 20 but of 30 extra years of life. On the No Difference View, we ought here to choose Program B. On the Two-Tier View, since benefits to general people matter only half as much, we ought again to choose Program A.

When I consider these examples, I accept the No Difference View, as do many other people. But some people accept the Two-Tier View. It must make a difference, these people believe, that only Program A would give more years of life to the same particular people, thereby benefiting these people.

In some other kinds of case, the Two-Tier View is harder to accept. Suppose first that, in *Case Three*, we have only these alternatives:

If we do A	Tom will live for 70 years,	Dick will live for 50 years,	and Harry will never exist.
If we do B	Tom will live for 50 years,	Dick will never exist,	and Harry will live for 70 years.

This case is a smaller version of *Case One*. On the No Difference View, we have equal reasons to act in either of these ways. On the Two-Tier View, it would be wrong to do B, since B would be worse for Tom, and A would not be worse for either Dick or Harry. If we do A, that would be worse only for the general person who would here partly consist of Dick and Harry.

Suppose next that, in *Case Four*, another outcome would be possible. Our alternatives are these:

If we do A	Tom will live for 70 years	Dick will live for 50 years	_____

If we do B	Tom will live for 50 years	_____	Harry will live for 70 years
If we do C	_____	Dick will live for 70 years	Harry will live for 50 years

'_____' means 'will never exist'.

The Two-Tier View again implies that it would be wrong to do B rather than A, since B would be worse for Tom, and A would be worse only for a general person. This view also implies that it would be wrong to do A rather than C, since A would be worse for Dick, and C would be worse only for a general person. It would be similarly wrong to do C rather than B, since C would be worse for Harry and B would be worse only for a general person. These claims imply that, whatever we do, we would be acting wrongly, since we ought to have done something else instead.

These are unacceptable conclusions. Even if there are some cases in which we cannot avoid acting wrongly, that is not true in *Case Four*. These three acts and outcomes are clearly morally equivalent. If we accept the Two-Tier View, we must revise this view, so that it ceases to have these implications. [643]

In revising this view, we should try to change this view's implications in cases like *Four*, while preserving its implications in the much more common cases that are like *Case Three*. If we did not preserve those implications, we would be abandoning this view. If the Two-Tier View made claims about the intrinsic goodness of outcomes, it could not be revised in this selective way. We could not coherently claim both that

> (G) the outcome produced by B would be worse than the outcome produced by A if these are the only possible alternatives,

and that

> (H) these outcomes would be equally good if C is also possible.

Whether one of two outcomes would be intrinsically worse cannot depend on which other outcomes are possible. But the Two-Tier View might make claims that are only about which acts are wrong. When we ask whether one of two acts would be wrong, the answer may sometimes depend on which other acts are possible.

Suppose for example that, in *Extra Risk*, two people's lives are in danger. These people are strangers to me. I could either

> X: Do nothing

or

Y: Save one of these people's lives at a great risk to myself.

We can plausibly believe that, if these are my only possible acts, I would be morally permitted to act in either way. Since Y would involve a great risk to me, this heroic act would go beyond the call of duty. Suppose instead that I could also

Z: Save both these people's lives, at no extra risk to myself.

If I knew that Z was possible, doing Y would be wrong. If I decide to run this risk, I ought to save both these people. But I would still be morally permitted to do X, since I would have no duty to run this risk. Whether it would be wrong for me to do Y rather than X therefore depends on whether Z is possible. We can explain why that is true by appealing to these facts about the risk to me.

If we accept the Two-Tier View, we might similarly claim that

> (I) it would be wrong for us to do B rather than A in *Case Three*, when these are the only possible acts, but doing B would *not* be wrong in *Case Four*, in which C is also possible.

On this suggestion, the Two-Tier View should not be applied to cases like *Four*. We would need, however, to defend this claim. We cannot merely say that the Two-Tier View should not be applied to cases in which this view has unacceptable implications. To illustrate this point, we can return to *Transplant*, in which the Act Utilitarian principle implies that a doctor ought secretly to kill one of his patients, if this person's transplanted organs would then be used to save five other people's lives. Most of us regard this implication of AU as counting strongly against this principle. In responding to this objection, Act Utilitarians cannot merely claim that their view should not be applied to cases of this kind. These people would have to claim that, if we understand their view correctly, we shall see why it does not apply to cases like *Transplant*. And Act Utilitarians would not be able to defend that claim.

We may also be unable to explain why the Two-Tier View should not be applied to cases like *Four*. This problem is in one way like the *Paradox of Voting*. According to

> *the Majority Criterion*: We ought to follow one of two policies when this policy is preferred by a majority of the relevant people.

It is often true that, of three possible policies, a majority of the relevant people rationally prefer B to A, C to B, and A to C. [644] In such cases, the Majority Criterion mistakenly implies that we ought to follow B rather than A, and ought to follow C rather than B, and ought to follow A rather than C. On this view, whichever policy we follow, we have acted wrongly, since we ought to have followed some other policy instead. As such examples show, we should reject the Majority Criterion, which cannot be a fundamental moral principle.

It might be suggested that, when we are choosing between only two alternatives, the Majority Criterion *is* acceptable, and that the Two-Tier View could also be restricted to such cases. But this restriction would be hard to defend. If we ought to give priority to avoiding what would be worse for particular people, that must be true in cases that involve more than two alternatives. *Case Four* provides, I believe, a strong objection to the Two-Tier View.

Suppose next that, in *Case Five*, we have a larger range of alternatives:

A	Adam lives for 70 years	Bernard lives for 50 years	_____	_____	_____
B	_____	Bernard lives for 65 years	Charles lives for 45 years	_____	_____
C	____	_____	Charles lives for 60 years	David lives for 40 years	_____
D	_____	_____	_____	David lives for 55 years	Ezra lives for 35 years
E	Ezra lives for 50 years	Frank lives for 30 years	_____	_____	
F	_____	Frank lives for 45 years	George lives for 25 years	_____	
G	_____	_____	George lives for 40 years	Herbert lives for 20 years	

If we do A rather than B, that would be worse for Bernard, by denying Bernard 15 more years of life. If we do B rather than A, that would be worse for the general person who would here in part consist of Adam and Charles, since this general person would be denied 25 more years of life. On the Two-Tier View, since losses to general people matter only half as much, this loss would matter less than Bernard's loss. This view therefore implies that we ought to do B rather than A. For similar reasons, we ought to do C rather than B, D rather than C, and so on down to G. When we are comparing several possible acts in a single case, the relation *ought to do rather than* is transitive. So the Two-Tier View here implies that we ought to do G rather than A. Rather than causing two people to exist, of whom one would live to 70 and the other to 50, we ought to cause two other people to exist, of whom one would live to 40 and the other to 20. That conclusion is clearly false. [645]

Though the Two-Tier View mistakenly implies that we ought to do G rather than A, this view also implies that we ought to do A rather than G. A would be much better than G for two general people, and worse for no one. So this view again implies that we cannot avoid acting wrongly. Whatever we do, we ought to have done something else. That is another unacceptable conclusion.

It is worth noting how this view goes astray. If we do B rather than A, that would be better for someone. If we do C rather than B, that would be better for someone else. These facts may suggest that, if we do C rather than A, that would be better for *two* people. But that is not true. If we do C rather than A, that would be better for no one.

We should again ask whether this objection could be met by revising the Two-Tier View. It is not obvious what this revised view should claim. *Case Five* is not like *Four*, in which all of the possible acts and outcomes are clearly morally equivalent. If A to G are the only possible acts, as we can suppose, the Two-Tier View in one way implies that doing G would be best. G is the only act that would not be worse for someone than some other possible act. It is clear, however, that G is not the *best* of these acts, but the *worst*. In its revised form, the Two-Tier View should imply that doing G would be wrong.

On the No Difference View, we ought to do A, since the two people who would then exist would have the longest lives. For the revised Two-Tier View to remain distinct from the No Difference View, the Two-Tier View must still imply that we ought to do B rather than A. This view might claim:

> (J) It would be wrong to do C rather than A, since C would be worse for two general people and better for no one. It would be similarly wrong to act in any of ways D to G rather than A, since these acts would all be worse than A for two general people, and better for no one. Of the remaining acts, A would be worse than B, since A would be worse for Bernard, and B would be worse for a general person by less than twice as much. So we ought to do B.

This may be the best revised version of the Two-Tier View. But (J) is not plausible. For another example, suppose that, in *Case Six*, our alternatives are these:

A	Adam lives for 70 years	Bernard lives for 50 years	_____	_____
B	_____	Bernard lives for 65 years	Charles lives for 45 years	_____
C	____	_____	Charles lives for 70 years	David lives for 45 years

As before, (J) implies that we ought to do B. But C is clearly better than B. C would give Charles an extra 25 years of life, which is a greater benefit than the 20 extra years that B would give to a general person. As the No Difference View implies, we ought to do C rather than B, and A rather than C.

Such examples show, I believe, that we should reject the Two-Tier View and accept the No Difference View. We should believe that

> (D) it would be in itself worse if some of the lives that will be lived will be less worth living,

and that

> (E) we have reasons not to act in ways that would have such effects, and if these effects would be very bad, and we could avoid them at little cost to ourselves, such acts would be wrong.

We should also believe that

> (F) it makes no difference whether, because these future lives would be lived by the same people, these outcomes would be worse for these people.

79 Scanlonian Contractualism and Future People

We can now return to Scanlonian Contractualism. Scanlon intends his formula to cover all of the acts with which we could affect future people. When applied to such acts, I shall argue, Scanlon's view needs to be revised.

According to Scanlon's Impersonalist Restriction, we cannot reject principles by appealing to claims about the goodness of outcomes. All reasons for rejecting principles must be *personal*. Scanlon also calls these reasons 'generic'. This word may suggest that such reasons could appeal to claims about what I have called *general people*. But that is not what Scanlon means. These generic personal reasons, Scanlon writes, are the reasons 'that any person would have in virtue of standing in one of the positions in a situation of the kind to which the principle applies'. [646] And he writes

> These must be reasons that such a person would have 'on his or her own behalf'.[647]

He also writes: 'This interpretation. . . rules out, as grounds for rejecting a principle, appeals to impersonal values. . . What it allows are reasons arising from the way a person would be affected by following the principle'.

Suppose that, in *Case Seven*, we must choose between two other medical programs. The predictable results would be these:

If we do A:	A thousand X-people would be conceived and live for 41 happy years,	and a thousand Y-people would be conceived and live for 40 happy years.
If we do B:	The same X-people would be conceived and live for 40 happy years,	and a thousand different Z-people would be conceived and live for 80 happy years.

Given Scanlon's claims about admissible grounds for rejecting principles, Scanlon's Formula seems here to require us to choose Program A. The X-people would have reasons on their own behalf to reject any principle that permitted us to choose B, since this choice would impose on the X-people the significant burden of being denied one more year of happy life. None of the other people would have reasons on their own behalf to reject any principle that requires us to choose A, since this choice would not impose any burden on any of these people. Our choice of A would not be worse for the Y-people, since if we had chosen B these people would never have existed. Nor would our choice of A be worse for any of the Z-people, since these people would never exist. Given these facts, it seems, the X-people could reasonably reject any principle that permits us to choose B, and could claim that we *owed it to them* to choose Program A.

If Scanlon's Formula requires us to choose A, as I have just claimed, that would be an objection to Scanlon's view. We ought to choose B. This choice would be required, not only by the No Difference View, but also by any plausible version of the Two-Tier View. Program B would give to a thousand general people 40 extra years of life. That is a very much greater benefit than the single extra year that Program A would give to the thousand particular X-people. Though we may believe that benefits to particular people matter more than benefits to general people, we could not plausibly believe that these benefits matter 40 times as much.

Scanlon might reject my claims about what his formula implies. I have assumed that, for one of two acts to impose a burden on someone, this act must be worse for this person than the other act would have been. We can call this *the comparative account* of benefits and burdens. Some writers claim that, when we consider acts that would cause certain people to exist, we should appeal instead to a *non-comparative* account. On this view, if we cause someone to exist who will be in some way badly off---by being deaf, for example, or having some life-shortening condition---that is enough to make it true that we are burdening or harming this person. We are imposing a burden on this person even if our act is not worse for this person, because this person's life is worth living, and having such a life is not worse than never existing.

If Scanlon appealed to this non-comparative account of burdens, he might claim that, in *Case Seven*, his formula does not require us to choose Program A. The X-people might claim that we owed it to them to choose A, since choosing B would have imposed on the X-people the burden of living for only 40 years. But if we choose A, Scanlon might say, that would impose the same burden on the Y-people, since these people would also live for only 40 years. On this non-comparative account, it is irrelevant that, while choosing B would be worse for the X-people, by denying them one extra year of life, choosing A would *not* be worse for the Y-people. On this view, it is a burden to live for only 40 years, and people have equal claims not to have this burden imposed on them whether their alternative would be living for longer, or never existing.

In some cases, this non-comparative account is plausible. Some acts can be claimed to harm people, even though these acts are not worse for the people who are harmed. But no such claim is plausible here. If the Y-people live for only 40 happy years, that is a burden only in the sense that it would be better for these people if they lived for more than 40 happy years. We would not be imposing a burden on these people, or be harming them, if we choose A, thereby failing to prevent these people from ever existing.

Some Scanlonian might now argue:

> If we choose B, we would impose on the X-people the burden of being denied one extra year of life. If we choose A, we would impose on the Z-people the burden of being denied 80 years of life. Since that is a much greater burden, the Z-people could reasonably reject any principle that does not require us to choose B.

Scanlonians cannot, however, make such claims. When Scanlon appeals to the principles that no one could reasonably reject, he uses 'no one' to mean 'none of the people who ever exist'. On this argument, it would be wrong for us to choose Program A, because the Z-people could reasonably reject any principle that permits this choice. But if we choose A, these Z-people would never exist. We cannot defensibly claim that some act is wrong because any principle that permits such acts could be reasonably rejected by some people who never exist. We could not, for example, claim that it would be wrong for any of us to choose not to have children, because any principle that permits this way of acting could be reasonably rejected by the merely possible children whom we do not have. [648]

Though *Case Seven* is artificial, and unrealistically precise, many actual cases are relevantly similar. Many of our possible acts or policies would predictably cause some future people to be much worse off than the different future people who, if we had acted differently, *would* have existed. My examples are acts or policies that would deplete certain scarce resources, or overheat the Earth's atmosphere. When we could avoid such acts at little cost to ourselves, these acts would be wrong. If we act in these ways, however, these different future people would never exist. When we

apply Scanlon's Formula in a way that appeals only to personal reasons, we are forced to ignore the fact that, if we had acted differently, these other people would have existed, and would have been much better off. These are morally relevant facts, which might make such acts wrong. To allow us to appeal to such facts, Scanlon must revise his claims about what are admissible grounds for rejecting principles.

Scanlon might suggest that, though all reasons for rejecting principles must be, in one sense, personal, these reasons could take two forms. In most cases, we could appeal to the burdens that some principle's acceptance would impose on us, as particular people. These burdens would give us reasons on *our own behalf*. In some other cases, however, we could appeal to the burdens that would be imposed on us, when regarded as *the person to whom some description applies*.

To assess this proposal, we can return to *Case Three*, in which our alternatives are these:

If we do A	*Mary* will have a child, Tom, who will live for 70 happy years	*Kate* will have a child, Dick, who will live for 50 happy years
If we do B	Tom will live for 50 happy years	Dick will never exist, but Kate will have another child, Harry, who will live for 70 happy years

On this revised version of Scanlon's view, we could deny that we owed it to Tom to do A. If we do B, that would be much worse for Tom, since our act would deny Tom an extra 20 happy years of life. But if we do A, that would be much worse for Dick, when Dick is regarded as *Kate's next child*. By doing A, we would also deny Dick, when so regarded, an extra 20 happy years of life.

Scanlon should not, I believe, make such claims. Phrases like 'your next child' are often used in this way, so that they refer to what I have called some *general person*. But it would be highly misleading for Scanlon to state his view in this way. Scanlon claims to be giving an account of

> the particular forms of concern that we owe to other *individuals*. [649]

General people are *not* individuals. A general person is a vast group of possible individuals, or people, one of whom will be actual. If we do A, and Dick lives for 50 happy years, Dick might agree that it would have been in one way better if we had done B, so that Dick would never have existed, and Kate would have had a different child who would have lived for 70 happy years. But there is no sense in which our doing A was worse for Dick. And if we fail to distinguish between Dick and Harry, regarding them as merely parts of a general person, we are ignoring the *separateness* of persons, which has been called 'the basic fact for ethics'. [650]

Return next to *Case Five*, in which three of our alternatives are these:

A	Adam lives for 70 years	Bernard lives for 50 years	_____	_____
B	_____	Bernard lives for 65 years	Charles lives for 45 years	_____
C	____	____	Charles lives for 60 years	David lives for 40 years

On this version of Scanlon's view, he would claim:

> It would be wrong to do either B or C, since any principle that permits these acts could be reasonably rejected by Charles, speaking on behalf of the general person who would here in part consist of Charles and Adam.

This claim would be implausible. If we do either B or C, Charles might later agree that we ought to have done A. But B or C would give Charles either 45 or 60 happy years, and if we had done A, Charles would never have existed. Charles is the person who has, not the *strongest*, but the *weakest* personal reasons to reject any principle that permits us to do either B or C. Nor would it help to appeal to Charles's reasons, not on his own behalf, but on behalf of the general person who consists in part of Charles and Adam. As I have said, there is no such person. Nor should we regard Charles and Adam as if they were the same person.

There is a better version of Scanlon's view. Scanlon should claim that, when we ask which are the principles that no one could reasonably reject, we should consider, and compare, two kinds of reason for rejecting principles. Each of us would have *personal* reasons for rejecting principles that permit or require certain acts. These reasons would be provided by the facts that such acts would impose burdens on us, or be unfair to us, or by other such facts. We would also have *impartial* reasons for rejecting principles that permit or require certain acts. These impartial reasons would be provided by the ways in which such acts would make things go worse, in the impartial-reason-implying sense.

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 morally 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.

Scanlonian Contractualism ought, I believe, to take this form. Before I defend this belief, it will help to consider why Scanlon's view does not already take this form.

One explanation is that, on Scanlon's view, all reasons for rejecting principles must be had by single people considered on their own, rather than as members of some group. Such *individuals'* reasons must also be personal reasons. If Scanlon dropped this *Individualist Restriction*, as I have argued that he ought to do, that would allow him to drop his restriction to personal reasons.

Scanlon also claims that, when we ask what we owe to each other, we need not consider certain *impersonal* reasons. Reasons are

> *impersonal*, in Scanlon's sense, when these reasons 'are not grounded in the moral claims or the well-being of individuals, either ourselves or others'.

We have such impersonal reasons, for example, to avoid acts that would inflict pain on animals, or would cause some species of animal to become extinct. Since these reasons have nothing to do with the moral claims or well-being of persons, Scanlon claims that such reasons are not relevant to what, as persons, we owe to each other.

These impersonal reasons may also be

> *impartial*, in the sense that we have these reasons whatever our personal point of view.

But we have other impartial reasons that are not, in Scanlon's sense, impersonal. We have such impartial reasons to care about the well-being of every individual or person. Scanlon says little about these impartial reasons. But when he claims that all reasons for rejecting principles must be personal, Scanlon thereby excludes, as irrelevant to what we owe to each other, not only impersonal reasons, but also those impartial reasons that are provided by facts about the well-being or moral claims of people. These impartial reasons, we might object, *are* relevant to what we owe to each other.

Scanlon might reply that, when our impartial reasons are provided by such facts about the well-being or moral claims of people, we have no need to appeal to these reasons. We all have impartial reasons, for example, to reject any principle that would impose burdens on certain people. But since these people would have personal reasons to

reject such principles, there is no need for us to appeal, as well, to these impartial reasons.

In most of the cases that Scanlon discusses, this would be a good reply. As this reply also shows, if Scanlon allowed us to appeal to impartial reasons, that would make no difference to most of the moral reasoning that his Contractualism describes. In most of our moral thinking, we can ignore the fact that our choice between different acts would affect the identity of future people. Most of our acts would not predictably cause some future people to be worse off than different future people would been. When our acts would predictably make things go worse, that is usually because these acts would be *worse for* one or more particular people. Since these people could appeal to the fact that such acts would be worse for them, we need not also appeal to the fact that such acts would make things go worse, in the impartial-reason-implying sense.

Things are different, however, when we consider some of the acts or policies with which we might affect future people. In some cases, we should consider what might happen to the different possible people who might later be actual. Some of these cases involve future people who would soon be actual. In deciding when to have children, for example, we ought to ask when we would be able to give such children a good start in life. That is why Jane ought not to have her first child when she is only 14. In other cases, such as those involving policies like Depleting or Overheating, we must consider possible effects on the many different people who might exist in the further future. When we apply Scanlon's Formula to such cases, it is not enough to ask which are the principles that no one would have sufficient personal reasons to reject. To explain why certain acts or policies 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. As we have seen, we cannot 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.

Return, for example, to *Case Seven*, in which our alternatives are these:

A:	A thousand X-people would be conceived and live for 41 years,	and a thousand Y-people would be conceived and live for 40 years.

B:	The same X-people would be conceived and live for 40 years,	and a thousand different Z-people would be conceived and live for 80 years.

We ought, I have claimed, to choose Program B. But the X-people would have personal reasons to reject all principles that required us to choose B, since this choice would have denied these people the significant benefit of one extra year of life. And we cannot claim that the Z-people would have stronger personal reasons to reject principles that required us to choose A. If we choose A, these people would never exist. But *we* could reasonably reject such principles. We could appeal to the fact that, if we choose A rather than B, things would go much worse in the impartial reason-involving sense. We would all have strong impartial reasons to want there to be a thousand people who would live for 80 years, rather than a thousand different people who would live for only 40 years. In cases of this kind, we need to appeal to such impartial reasons. If we could appeal only to personal reasons, we would have to ignore the fact that, rather than causing the X-people to live for only one year longer, we could cause there to be as many people who would live for 40 years longer.

If Scanlonian Contractualism allowed us to appeal to impartial reasons, Scanlon's Formula would be unchanged. This view would keep Scanlon's greatest contribution to our moral thinking: his appeal to principles that no one could reasonably reject. But Scanlon might have to qualify some of his other claims. Scanlon talks of what we *owe* to others, and he writes:

> The idea of justifiability to all possible beings. . . seems impossibly broad, and barely coherent. . . the beings whom it is possible to wrong are all those who do, have, or will actually exist. [651]

Such remarks suggest that

> (K) the acts with which we affect people cannot be wrong unless there exists, at some time, some actual person whom we have *wronged*, and to whom we *owed* it not to act in this way.

(K) implies that, in *Case Seven*, it would not be wrong for us to choose Program A, though we know that there would then be many people who would live for 40 years, rather than people who would have lived for 80 years. We would not have wronged the people who would live for 40 years, since we did not wrong these people by failing to prevent them from being conceived. Nor did we owe it to these people to cause them never to exist. Nor would we have wronged the people who would have lived for 80 years, since we could not have wronged people who never exist, nor could we have owed it to such people to cause them to exist.

Similar claims apply to many of the acts or policies with which we can affect those people who will live in the further future. If we choose policies like Depleting or Overheating, we may greatly lower the quality of future people's lives, for the sake of much smaller benefits to ourselves. But in many cases of this kind, (K) implies that it would not be wrong to cause this great lowering in the quality of future lives. If these lives would be lived by different people, our choice of these policies may not wrong any of these people, and we may not owe it to such future people not to choose these policies. When applied to such cases, (K) conflicts not only with the No Difference View, but even with the Two-Tier View. When we see *why* (K) has these implications, this claim ceases to seem plausible. We should expect that, in such cases, our acts or policies may be wrong, though there would not be any actual people whom we have wronged.

In making these claims, I am not assuming that we cannot be wronging someone if we know that our act would not be worse for this person. As I have claimed elsewhere, some of our acts may wrong future people even if we know both that these people's lives would be worth living, and that, if we had acted otherwise, these people would never have existed. For example, we might wrong some future people by choosing policies that risk causing some catastrophe, such as using nuclear energy and failing to ensure that radio-active wastes are stored safely. And Jane might be wronging Johnny by having him when she is only 14, so that she fails to give him a good start in life. Such acts might be wrong because they violate certain people's rights, or they cause people to exist with rights that cannot be fulfilled. [652]

Such claims, however, cannot wholly solve the Non-Identity Problem. First, we are not asking only which acts or policies would be wrong. We all have reasons to care about future generations, and about how our acts or policies might affect the quality of future people's lives. It is of great importance whether these reasons would be weaker if, because these lives would be lived by different people, these acts or policies would not be worse for these people. We cannot answer this question by appealing to claims about people's rights.

Second, if we appeal only to such claims, we shall have false beliefs about what we ought morally to do. We shall be led to ignore the fact that, if we had acted differently, the people who would have existed later would have had better lives. And if we ignore such facts, we may act wrongly. If everyone always acted in such ways, each new set of people would live worse lives. The world would be slowly wrecked.

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

We can now turn to the relation between Scanlonian and Kantian Contractualism. When we apply the Kantian Contractualist Formula, I argued, it is only the optimific principles whose universal acceptance everyone could rationally choose. Kantian Contractualism therefore implies Rule Consequentialism. Scanlon does not criticize this argument.

According to my Convergence Argument, since it is only the optimific principles that everyone could rationally choose, no one could reasonably reject these principles. If that is true, Kantian Rule Consequentialism could also be combined with Scanlonian Contractualism.

This second argument does not apply to the view stated in Scanlon's book, since this view includes both the Individualist and Impersonalist Restrictions. By appealing to these restrictions, Scanlon could reject some of my argument's premises. But Scanlon's view would be strengthened, I have argued, if he dropped these two restrictions, and if his view made claims, not about wrongness itself, but about what makes acts wrong. I shall now ask whether my Convergence Argument succeeds when applied to this revised version of Scanlon's view. Since this revision does not change Scanlon's Formula, but merely drops two restrictions, I shall be discussing what we might call *the Unrestricted Scanlonian Formula*.

It will be enough to discuss some of those Rule Consequentialist principles that are *UA-optimific*, in the sense that their universal acceptance would make things go best. According to one version of what I call

> *the Triple Theory*: Everyone ought to follow these optimific principles because these are the only principles whose universal acceptance everyone could rationally choose, and the only principles that no one could reasonably reject.

In considering this theory, we have four questions:

> Q1: What do these optimific principles require us to do?
>
> Q2: Are these the only principles whose universal acceptance everyone could rationally choose?
>
> Q3: Are these the only principles that no one could reasonably reject?

Q4: Are these the principles that everyone ought to follow?

Whether we could *rationally choose* one of two principles depends on the strength of all of our non-deontic reasons to choose these principles. Whether we could *reasonably reject* one of two principles depends instead on whether we have grounds to reject this principle that are relevantly stronger than anyone's grounds to reject the other principle. My argument for the Triple Theory is, in part:

> (A) If we could *not* rationally choose one of two principles, there must be facts that give us strong grounds for rejecting this principle.
>
> (B) If everyone *could* rationally choose the other principle, no one's grounds for rejecting this alternative could be as strong.

Therefore

> (C) We could reasonably reject the first principle, and no one could reasonably reject this alternative.

If we add certain further premises, this argument shows, I believe, that the Kantian and Scanlonian Formulas at least very often coincide, by requiring us to follow the same principles. But there may be some exceptions.

Scanlon describes one kind of possible exception. When Rawls and Scanlon propose their versions of Contractualism, they both appeal to the same kind of case. In what we can call

> *Rawls-Scanlon Cases*, we can either save one person from some great burden, or give much smaller benefits to many other people, who are all much better off.

We can call these people *Blue* and *the Many*. Suppose that, in one such case,

> (1) everyone could rationally choose some optimific principle that required us to give the small benefits to the Many,

and that

> (2) some people could not rationally choose any conflicting principle that required us to save Blue from her great burden.

If (1) and (2) were true, the Kantian Contractualist Formula would require us to give the small benefits to the Many. But Scanlon suggests that

> (3) in some of these cases, Blue could reasonably reject every such principle, and no one could reasonably reject some principle which required us to save Blue from her great burden. [653]

If (1) to (3) were true, the Scanlonian Formula would require us to save Blue from this burden. Kantian and Scanlonian Contractualism would here conflict.

Before deciding whether (3) is true, we must ask in which of these cases the optimific principles would require us to give the small benefits to the Many. To answer such questions, Scanlon writes, we would have to know 'how Parfit's notions of impartial reasons and "best outcome" deal with *aggregation*', or with how the goodness of outcomes might depend on the number of people who would receive benefits or burdens. My definition of this sense of 'best', he writes,

> leaves open the possibility that the conception of 'best outcome'. . . is in important respects non-aggregative.

This definition ought, I believe, to leave this question open. Some possible outcome would be best, in this impartial-reason-implying sense, if this outcome is the one that, from an impartial point of view, everyone would have most reason to want, or to hope will come about. It is a substantive question, which could not be answered by a definition, just when and how the strengths of everyone's impartial reasons would in part depend on facts about how many people might receive certain benefits or burdens.

When we ask which of two outcomes would be in this sense better, it would be very implausible to claim that the answer *never* depends on the number of people who might receive benefits or burdens. But we are here considering only Rawls-Scanlon Cases. For a more extreme example of this kind, we can suppose that, in *Case One*, the only possible outcomes are these:

A:	Blue will have 1,000 days of pain	Each of the Many will have no pain
B:	Blue will have no pain	Each of these people will have one brief period of pain

It is often assumed that, in all such cases, there must be some number of small benefits to the Many that would outweigh Blue's great burden, making outcome A better than outcome B. If the goodness of outcomes depended only on the net sum of benefits minus burdens, as Utilitarians believe, that would imply that it must be in this way possible for A to be better than B. But this conclusion is not implied by the impartial-reason-implying sense of 'better.' In our beliefs about the goodness of outcomes, we might reject this Utilitarian view. And if the benefits to each of the Many would be very small, we might plausibly believe that no number of these benefits could outweigh Blue's great burden. We might believe for example that, if Blue had her 1,000 days of pain, that would be worse than if *any* number of other people had one minute, or one hour, of pain. This belief would be true if we would all have stronger impartial reasons to want or hope that, in all such cases, the single person would be saved from

her great ordeal. It is an open question, I believe, whether we would have such reasons.

When we consider acts that would give to very many people *very small* benefits, or impose very small burdens, it is easy, I have claimed, to make moral mistakes. Given the technological developments of the last two centuries, such cases now have great importance. But we can ignore such cases here. These cases raise difficult problems which are not relevant to the question whether Scanlonian Contractualism might conflict with Kantian Rule Consequentialism. If the Scanlonian Formula would require us to ignore some such very small benefits or burdens, the same might be true of the optimific Rule Consequentialist principles. And we are looking for cases in which the optimific principles would require us to give the small benefits to the Many.

Since there are several views about which outcomes would be best, there are also several views about which principles would be optimific. The important question is whether Scanlonian Contractualism *necessarily* conflicts with Kantian Rule Consequentialism, or whether there are plausible versions of these theories that do not conflict, and could therefore be combined. So I shall suppose that, in their assessments of the goodness of outcomes, Kantian Rule Consequentialists accept a strong version of what I earlier called the *Telic Priority View*. That assumption makes this form of Consequentialism closer to Scanlonian Contractualism.

Suppose that, in *Case Two*, the only possible outcomes would be these:

A: Blue will have 100 days of pain	Each of the Many will have no pain
B: Blue will have no pain	Each of these people will have 10 days of pain

As before, and in all these cases, we should also suppose that each day of pain is an equal burden. On the Telic Priority View, people's burdens matter more, doing more to make the outcome worse, the worse off these people are. Since Blue would be much worse off in outcome A than each of the Many would be in outcome B, most of Blue's days of pain would matter more than the Many's days of pain. On a strong version of this view, for outcome B to be worse than outcome A, the numbers of the Many would have to be much greater than ten. For B to be clearly worse than A, we can here suppose, there would have to be more than a hundred or a thousand other people who, in B, would each have 10 days of pain.

Similar claims apply to *Case Three*, in which the only possible outcomes are these:

A: Blue will live to the age of 30	Each of the Many will live to 75

B:	Blue will live to 70	These people will live to 70

We can again suppose that, for B to be worse than A, the number of the Many would have to be more than a hundred or a thousand.

Let us say that, in such cases, moral principles are *Blue-protecting* if they require us to save Blue from her great burden, and *Blue-burdening* if they require us instead to save the Many from their much smaller burdens, thereby giving them much smaller benefits. On the views just described, the Blue-burdening principles would be optimific only when, compared with the benefit to Blue of being saved from her great burden, we could give to the Many a *much* greater total sum of benefits.

Return next to my argument that, in the thought-experiments to which the Kantian Formula appeals, it is only the optimific principles that everyone could rationally choose. My argument compares these principles with other possible principles that are *significantly* non-optimific, in the sense that their universal acceptance would make things go much worse. *Slightly* non-optimific principles raise some complications that would be best considered later.

Everyone would have strong impartial reasons to choose that everyone accepts the optimific principles, since that choice would make things go much better. And no one's impartial reasons, I argued, would be decisively outweighed by any relevant conflicting reasons. Since the optimific principles would impose great burdens on certain people, these people would have strong personal reasons *not* to choose the optimific principles. But these reasons would not, I claimed, be decisive.

Do these claims apply to the cases that we are now considering? Would Blue have sufficient reasons to choose that everyone accepts some optimific Blue-burdening principle? When I claimed that we could all rationally choose some optimific principle even if that choice would impose some great burden on us, I was discussing cases in which, by choosing such a principle, we would indirectly save many other people from *similarly great* burdens. In *Lifeboat*, for example, if I choose the Numbers Principle rather than the Nearness Principle, I would die, but my choice would indirectly save many other people's lives.

In Rawls-Scanlon Cases, no such claim is true. If Blue chooses some optimific principle, she would bear a great burden, and she would not indirectly save any number of other people from similarly great burdens. She would only save many people from *much smaller* burdens. It may seem that, given this fact, Blue would not have sufficient reasons to choose this principle. These are the cases in which it could most plausibly be claimed that some people could not rationally choose the optimific principles.

We ought, I suggest, to reject even this claim. Return to *Case Two*, in which we could

either

(1) save Blue from all of her 100 days of pain

or

(2) save some number of other people from all of their 10 days of pain.

For the reasons given above, we are supposing that, for (2) to make the outcome better, this number of other people would have to be more than a hundred or a thousand. If Blue chose some optimific principle that required us to do (2), Blue would have 100 days of pain, but her choice would save these other people from more than 1,000 or 10,000 days of pain. This choice would also have such effects in many other such cases. These facts would, I believe, give Blue sufficient reasons to make this choice. Blue would have sufficient reasons to choose to have her 100 days of pain, if her choice would save these other people from such a very much greater number of days of pain, in significant amounts of 10 days per person.

We can next ask whether, in any of these cases, everyone could rationally choose some significantly non-optimific Blue-protecting Principle. The answer, I believe, is No. The Many would have both impartial and personal reasons *not* to choose any such principle. And most of us would have these impartial reasons and would have no contrary reasons. So most people would not have sufficient reasons to choose such a principle.

These cases are not, I conclude, a strong counter-example to my argument for Kantian Rule Consequentialism. For these and some of the other reasons that I give in Chapter 16, when we apply the Kantian Formula to these cases, it is only the optimific Blue-burdening Principles that everyone could rationally choose.

We can now return to my argument that Kantian Rule Consequentialism can be combined with Scanlonian Contractualism. When applied to Rawls-Scanlon Cases, my argument would in part be this:

(D) Since the Many could *not* rationally choose any Blue-protecting principle, there must be facts that give these people strong grounds or reasons for rejecting these principles.

(E) Since Blue *could* rationally choose some Blue-burdening principle, Blue's grounds for rejecting these principles cannot be as strong.

Therefore

> (F) The Many could reasonably reject any Blue-protecting principle, and Blue could not reasonably reject every Blue-burdening principle.

In his commentary above, Scanlon rejects this argument. He suggests that

> (G) in some of these cases, though Blue could rationally choose some optimific Blue-burdening principle, Blue could also reasonably reject every such principle, and none of the Many could reasonably reject every non-optimific Blue-protecting principle.

If this claim is true, the Scanlonian Formula would sometimes require us to follow these Blue-protecting principles. Scanlonian Contractualism would here conflict with Kantian Rule Consequentialism.

Is (G) true? In *Case Two*, we could either

> (1) save Blue from all of her 100 days of pain

or

> (2) save some number of other people from all of their 10 days of pain.

We are supposing that, for the optimific principles to require us to benefit the Many rather than Blue, it would have to be true that we could save the Many from a total of more than 1,000 or 10,000 days of pain. Could Blue reasonably reject these principles, claiming that we ought instead to save Blue from her 100 days of pain? And would it be unreasonable for the Many to reject this claim?

It is not clear that our answers should be Yes. We can agree that, since Blue would be much worse off than any of the Many if she had her 100 days of pain, Blue's grounds for rejecting any Blue-burdening Principle have, in one way, much greater moral weight. But in our assessment of the goodness of these outcomes, the fact that Blue would be much worse off has already been taken into account. That is why, for the optimific principles to require us to give the smaller benefits to the Many, we would have to be saving more than a hundred or a thousand of these people from all of their 10 days of pain. In our assessment of the goodness of these outcomes, we have already given, to Blue's pain, as much as ten or a hundred times the weight that we give to the pains of the Many. It is not clear that Blue could reasonably claim that, in deciding how to act, we ought to give Blue's pain *more* than ten or a hundred times the weight that we give to these other people's pain. Nor would it be clearly unreasonable for the Many to reject this claim.

Return next to *Case Three*, in which we could either

> (3) enable Blue to live to 70 rather than 30,

or

(4) enable some number of other people to live to 75 rather than 70.

We are supposing that, for the optimific principles to require us to do (4) rather than (3), this number of other people would have to be more than a hundred or a thousand. Rather than giving to Blue her extra 40 years of life, we would then be giving to these other people more than 500 or 5,000 extra years. Could Blue reasonably reject principles which require this act? Could she reasonably claim that her 40 extra years are morally more important than these other people's total of 500 or 5,000 extra years? And would it be unreasonable for these other people to reject this claim? As before, it is not clear that our answers should be Yes.

It might be objected that, in my claims about these cases, I have taken some plausible beliefs about what we ought morally to do, or about the strength of people's moral claims, and mistakenly presented these beliefs as being about the goodness of outcomes. The Priority View, Scanlon suggests, should be regarded as making claims, not about the goodness of outcomes, but about the strength of different grounds for rejecting moral principles. These claims, Scanlon writes, are

> most naturally understood within the context of a view that makes conclusions about right and wrong depend on the relative strength of the reasons that inviduals can offer in the process of interpersonal justification. They are less plausibly interpreted as claims about what it is good or bad to have happen. [654]

Rawls similarly suggests that, in our assessments of the goodness of outcomes, we should not appeal to any distributive principles, since such principles make claims that are about, not what is good, but what is morally right. [655]

These suggestions are, I believe, mistaken. Though the Priority View can take purely deontic and Contractualist forms, it can also plausibly take a telic form, which makes claims about the goodness of outcomes. [656] There are some moral principles which cannot plausibly take such a form. Some examples would be those deontological principles which require us not to treat people in certain ways, such as harming one person as a means of benefiting others. Such an act is wrong, these principles claim, even if this act would make the outcome better by minimizing the number of acts of this kind. But distributive principles do not make any such claims. We can plausibly believe that it would be better if benefits or burdens were more equally distributed, or if more of the benefits and fewer of the burdens came to people who were worse off. We can believe for example that, if Blue has her 100 days of pain, that would be worse than if a hundred people each had only one day of pain. This outcome would be worse, I believe, in the sense that, if these people were all strangers to us, we would have more reason to hope that Blue avoids this great ordeal.

It might next be objected that, in our assessments of the goodness of outcomes, we might reject the Telic Priority View, or we might accept only a much weaker version of this view. We would then reject the argument that I have just given for doubting Scanlon's (G). But it is not worth claiming that *some* versions of Kantian Rule Consequentialism conflict with Scanlonian Contractualism. There are also conflicts between different versions of Rule Consequentialism, such as those versions which appeal to the principles whose being universally *accepted,* or universally *followed,* would make things go best. As I have said, what matters is whether plausible versions of Scanlonian Contractualism *necessarily* conflict with plausible versions of Kantian Rule Consequentialism. And the Telic Priority View can plausibly take a fairly strong form.

81 The Independence of Scanlon's Theory

Remember next that, on

> *the Contractualist Priority View*: People have stronger moral claims, and stronger grounds to reject some moral principle, the worse off these people are.

Scanlon might claim that, compared with the Telic Priority View, this Contractualist view can plausibly take an even *stronger* form. That might be enough to make (G) true.

Return for example to *Case One*, in which the possible outcomes are these:

A:	Blue will have 1,000 days of pain	Each of the Many will have no pain
B:	Blue will have no pain	Each of these people will have one brief period of pain

It is often assumed that, if all pain is bad, there must be some number of brief periods of pain that would make B worse than A. This assumption is, I have claimed, mistaken. We can coherently and plausibly believe that, if Blue had her 1,000 days of pain, that would be worse than if any number of other people had some brief period of pain, such as 1 minute, or 10 minutes. We might have stronger impartial reasons to want or hope that, in all such cases, it would be the single person who would be saved from her great ordeal.

In some other cases, however, we could not plausibly make such claims. It might be implausible to claim that, rather than Blue's having her 1,000 days of pain, it would be better if a million, or a billion, or a billion billion people each had 10 days of pain, or 50 days of pain. We may therefore have to agree that, in some such cases, the optimific principles would require us to save some great number of people from their days of pain. And Scanlon might be right to claim that, in some of these cases, Blue could

reasonably reject these optimific principles, and none of the Many could reasonably reject some principle that required us to save Blue from her 1,000 days of pain. If these claims were true, Scanlonian Contractualism would here conflict with Kantian Rule Consequentialism, since these views would require us to act in different ways.

This conflict would not, however, be deep. On both these views, we ought to give strong priority to saving Blue from her great ordeal. The difference would be only that, on Scanlonian Contractualism, this priority would be somewhat stronger.

There are other ways in which, in some cases, these two views might have different implications. We can now return to the Contractualist part of Kantian Rule Consequentialism. According to the Kantian Contractualist Formula, we ought to follow the principles whose universal acceptance everyone could rationally choose. Suppose that, in

> *Case Four*, we could easily save the lives of one of two relevantly similar people.

According to

> *the Principle of Equal Chances*: In such cases, we ought to save one of these people in some way that would give each person an equal chance of being saved.

This is the only principle, we might claim, that both these people could rationally choose. Though this claim is plausible, it is not obviously true. Perhaps these people could also rationally choose some principle that merely required us to save one of them, leaving it up to us how we choose whom we save. The Kantian Formula would not then support the Principle of Equal Chances. The Scanlonian Formula, in contrast, decisively supports this principle. Neither of these people could reasonably reject this principle, since neither person has any claim to be given *more* than an equal chance of being saved, nor is there any other reasonable ground for rejecting this principle. [657]

Suppose next that, in

> *Case Five*, some quantity of unowned resources can be shared between different people, none of whom has any special claim to these resources. However we distribute these resources, these people would together receive the same total sum of benefits.

When we apply the Kantian Formula, we could claim that

> (H) everyone could rationally choose some principle that requires us, in such cases, to give everyone equal shares,

and that

> (I) no one could rationally choose any principle that permits us, in such cases, to give them less than equal shares.

I believe that, since these claims are true, the Kantian Formula requires us to follow this *Principle of Equal Shares*. But Utilitarians might reject (I), claiming instead that

> (J) everyone could rationally choose some principle that permitted us to give them unequal shares, since the total sum of benefits would be the same.

Though I believe that this claim is false, (J) is not *obviously* false. The Scanlonian Formula, in contrast, decisively supports the Principle of Equal Shares. No one could reasonably reject this principle, since no one has any claim to be given *more* than an equal share, nor is there any other possible objection to this principle.

Four and *Five* are not cases in which Kantian and Scanlonian Contractualism conflict. The difference is only that, though the Kantian Formula gives some support to the Principles of Equal Shares and Equal Chances, the Scanlonian Formula supports these principles in a stronger and decisive way. But suppose next that,

> in *Case Six*, if some people were given unequal shares, the total sum of benefits would be much greater.

In such cases, there might be some people who could not rationally choose the Principle of Equal Shares, since an equal distribution would both be much worse for these people, and make things go worse. But it might still be true that no one could reasonably reject the Principle of Equal Shares. Kantian and Scanlonian Contractualism *would* then conflict.

We can next note what these examples have in common. When we apply the Kantian Formula, asking which are the principles whose universal acceptance everyone could rationally choose, we take into account facts about how it would be best for things to go, in the impartial-reason-implying sense. In assessing the goodness of outcomes, I have claimed, we can plausibly give weight to some distributive principles. We can believe that one of two outcomes would be better, despite giving people a smaller total sum of benefits, if these benefits were more equally shared, or if more of the benefits came to people who were worse off. We can also believe that it would be better if people were given equal chances to receive some benefit. But as some of my examples show, when we apply the Scanlonian Formula, these distributive considerations can plausibly be given greater weight. That is not surprising. When we ask which principles everyone could *rationally* choose, the answer depends on all of our non-deontic reasons for choosing different principles. These include, not only our impartial reasons to prefer better outcomes, but also various personal, non-moral reasons, such as our reasons to choose what would benefit ourselves. The Scanlonian Formula appeals instead to claims about what are *reasonable* grounds for rejecting moral principles, in a

partly moral sense of 'reasonable'. We would expect that, in answering this narrower question, distributive principles could plausibly be given greater weight. Though things might go somewhat better if people were given equal shares, or equal chances to receive some benefit, it is much clearer that no one could reasonably reject the Principles of Equal Shares and Equal Chances.

For an example of a different kind, suppose that in

> *Case Seven*, we could either save Green from some burden, or save Black from a much greater burden. Black has been negligent, and is responsible for the fact that Green and Black are threatened with these burdens.

When we ask which principle these people could rationally choose, the answer might be some principle that saved Black from her much greater burden. Green might have sufficient reason to choose this principle. But if we ask which principle no one could reasonably reject, we might conclude that Black could *not* reasonably reject a principle requiring her to bear this greater burden, given the fact that it was Black's negligence which caused both her and Green to be threatened with these burdens. Kantian and Scanlonian Contractualism would then conflict.

There may be other cases in which these two kinds of Contractualism conflict. [658] And Kantian Contractualism may sometimes conflict with Rule Consequentialism. I believe that, in all or nearly all important cases, everyone could rationally choose that everyone accepts some optimific principle. But there may be cases in which everyone could also rationally choose some significantly non-optimific principle. In such cases, Kantian Contractualism would differ from Rule Consequentialism, by permitting us to act on either of these principles. And there may be other ways in which the three parts of the Triple Theory sometimes conflict.

If there are such conflicts, that may seem to show that we should reject this Triple Theory. But that is not, I believe, true. All of our theories need to be developed further, and revised. If what seem the most plausible theories have very similar implications, this fact gives us reasons to believe that we are making progress, and that these are the theories that we should try to develop further, and revise. If these theories have some conflicting implications, that may help us to decide how these theories should 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 some questions, we can distinguish several views:

Are normative claims intended or believed to state truths?

Yes — No

Cognitivism — Non-Cognitivism

Are there any normative truths?

Yes — No

Are these truths irreducibly normative? — Nihilism

Yes — No

Are these truths about what exists in some non-spatio-temporal part of reality? — Are the concepts and claims with which we state such truths irreducibly normative?

Yes — No

Yes — No — Non-Analytical Naturalism — Analytical Naturalism

Platonism Non-Metaphysical
Non-Naturalist
Cognitivism

These distinctions are rough, and further distinctions could be drawn. We ought, I believe, to accept some form of Non-Metaphysical Non-Naturalist Cognitivism. I shall argue that we ought to reject Naturalism and Non-Cognitivism, and that we have reasons to reject Platonism and Nihilism.

A *concept* is what is meant or expressed by some word or phrase, and by other words or phrases with the same meaning. The words 'new' and 'nuevo', for example, both express the concept *new*. Of the concepts that we shall be considering, most refer to *properties*, such as the properties of being new, glittering, a poet, a convincing argument, the brightest star, the first man to walk on the Moon, and an act that is wrong. As these examples suggest, any true claim about something can be regarded or restated as a claim about this thing's properties, and most claims about properties could be restated as claims about facts. When we claim that some concept *refers* to some property, we are not thereby claiming that anything *has* this property. [659] No one is the first man to walk on the Sun; and Nihilists believe that no acts are wrong.

The same word can have different senses or meanings, thereby expressing different concepts. A genius and the brightest star are in different senses bright. We should also distinguish between some word's ordinary meaning and what some person uses this word to mean. These meanings differ when someone either misuses some word, or deliberately uses some word in something other than its ordinary sense. Some people, for example, misuse the word 'refute' to mean 'deny', and I deliberately use the word 'event' in a wide sense that covers acts and states of affairs. When enough people misuse some word, what these people use this word to mean becomes one of the ordinary meanings of this word.

Consider next these two lists of words:

A: wrong, right, ought, should, good, bad, excellent, mediocre.

B: kill, crimson, square, electric, cause, city, marble, alive, sister, tall, unexpected.

Though I have not said what the words in either of these lists have in common, most of us would guess correctly into which list most other words should go. We would guess, for example, that 'desirable', 'rational', 'duty', and 'blameworthy', should go in list A, and that 'desired', 'liquid', 'young', and 'sad' should go in list B.

Words in list A are *normative*, as are the concepts, claims, and facts that we can use these words to express or state. There are, as we shall see, several conceptions of normativity. Words in list B are *naturalistic*, and claims that use only such words,

when they are true, state natural facts. Some fact is *natural*, on the most common definition, if facts of this kind are investigated or discussed by people who are working in any of the natural or social sciences. I shall suggest later how we can make these definitions more precise.

There are also some words that are partly normative and partly naturalistic. Some examples are the word 'murder' when this is used to mean 'wrongly kill', and the words 'cruel', 'rude', 'unpatriotic', and 'dishonest'. I shall say little about such words, and what are called the *thick* normative concepts that these words express. Though such concepts can add subtlety and perceptiveness to our normative thinking, the deepest theoretical questions are about the relations between the concepts and properties that are expressed and referred to by the words in lists A and B.

These questions are answered differently by those who accept the kinds of theory shown in my diagram above. Non-Cognitivists believe that normative claims should not be regarded as intended to be true, except perhaps in some minimal sense. Such views I shall discuss in Parts Four and Five. Cognitivists believe that normative claims are intended to be true. Some of these people are Nihilists, or *Error Theorists*, who believe that all positive normative claims are false. Other Cognitivists believe that some of these claims are true, and state normative facts.

These other Cognitivist theories are of three main kinds. Normative facts, all *Naturalists* believe, are one kind of natural fact. According to *Analytical Naturalists*, normative words have meanings that can be fully analysed or defined by using naturalistic words. On this view, though there is no distinction between normative and naturalistic *claims*, we can distinguish between normative and naturalistic ways of stating the same claim. This view correctly describes some uses of normative words. For example, if I said

> My prediction was wrong, because my headache has got worse,

I might mean only

> My prediction was false, because my headache has become more painful.

These would then be different ways of stating the same claim, and the same natural fact. But Analytical Naturalism cannot be plausibly applied to many other uses of 'wrong' and 'worse', or to normative uses of some other words, such as 'irrational' and 'unjust'.

If some normative word, concept, claim, or fact cannot be defined or restated in non-normative terms, we can call it *irreducibly normative*. According to *Non-Naturalist Cognitivists*, when such normative claims are true, they state irreducibly normative facts. According to *Non-Analytical Naturalists*, such claims state natural facts.

As examples of such theories, we can take three versions of the *Utilitarian* view that

(1) some act is right

just when, and because,

(2) this act maximizes happiness. [660]

If Utilitarians were Analytical Naturalists, they would claim that, when we say that some act is right, we mean that this act maximizes happiness. On this implausible view, since these phrases mean the same, they refer to the same property. When some act maximizes happiness, that is the same as this act's being right, or is *what it is* for this act to be right. (1) and (2) are different ways of stating the same fact, which is both normative and natural.

According to those Utilitarians who are Non-Naturalist Cognitivists, the phrase 'is right' is irreducibly normative, as is the concept that this phrase expresses. On this view, when some act has the natural property of maximizing happiness, this fact makes this act have the different, irreducibly normative property of being right. (1) and (2) have different meanings, and state different facts. (2) states a natural fact, but (1) states a fact that is not natural but irreducibly normative.

According to those Utilitarians who are *Non*-Analytical Naturalists, though the phrase 'is right' is irreducibly normative, this phrase refers to the same property as the naturalistic phrase 'maximizes happiness'. Despite having different meanings, (1) and (2) state the same fact, which is both normative and natural.

Similar claims apply to other Cognitivist moral theories, and to Cognitivist theories about other normative concepts, claims, and facts. These theories can be either Analytically or Non-Analytically Naturalist, or Non-Naturalist.

Of those who are in these ways *Normative* Naturalists, most are also *Metaphysical* Naturalists, who believe that all properties and facts are natural. But some Metaphysical Naturalists reject Normative Naturalism, and are either Nihilists or Non-Cognitivists. Though I believe that Metaphysical Naturalism is false, I shall not try to show that here. So when I use the word 'Naturalism' on its own, I shall always be referring to *Normative* Naturalism.

Naturalism and Non-Cognitivism are both, I shall argue, close to Nihilism. Normativity is either an illusion, or involves irreducibly normative facts.

In considering these theories, our main question will be how we should understand normativity. I shall use 'normative' both in a wide sense, and in narrower senses

which express different conceptions of normativity.

On the *rule-involving* conception, normativity involves rules, or requirements, which distinguish between what is *allowed* and *disallowed*, or what is *correct* and *incorrect*. Some examples are laws, the requirements of some code of honour, the rules of etiquette, and rules about spelling, grammar, and the meanings of words. Such rules or requirements are often called *norms*, and claims that state or apply such norms we can call *normative* in the *rule-implying* sense.

On the *reason-involving* conception, normativity involves reasons or apparent reasons. When I call some claim

> *normative* in the *reason-implying sense,* I mean that this claim asserts or implies some claim about some reason or apparent reason.

This, I shall argue, is the best conception. To illustrate these conceptions, suppose that I say

> You shouldn't eat peas with a spoon,

and

> You shouldn't use 'refute' to mean 'deny'.

These claims are normative in the rule-implying sense. But I might add that, since these rules are now so often broken, you have no reason not to act in these ways. My claims would not then be normative in the reason-implying sense. [661]

On a third conception, normativity involves actual or possible motivation. Korsgaard, for example, writes that if some argument 'cannot motivate the reader to become a Utilitarian then how can it show that Utilitarianism is normative?' [662] Elizabeth Anderson similarly writes that 'any theory of the good must have normative force: we must be capable of being moved to action by the reasons it gives us.' [663] Many other people make such claims.

We ought, I shall argue, to reject this *motivational* conception. Normativity, we should agree, is closely related to motivation. If we are aware of certain reasons for acting, and we are fully rational, we would be motivated to act for these reasons. But that does not imply that normativity in part consists in actual or possible motivating force.

On a fourth conception, normativity involves certain kinds of attitudes to our own and other people's acts. Of those who defend this *attitudinal* conception, some are Naturalists, who believe that normative claims are about such attitudes. Others are Non-Cognitivists, who believe that these claims *express* such attitudes.

These conceptions can be combined. Some people, for example, give attitudinal accounts of morality, motivational accounts of reasons, and rule-involving accounts of some other normative facts.

When G. E. Moore started the long debate about Naturalism, he was discussing the concept *good* and the property of being good. [664] Many other writers discuss Naturalist theories about morality. But I shall first discuss non-moral practical reasons and reason-implying oughts. The questions raised by Naturalism here take simpler and clearer forms.

These are also the most important questions if, as I believe, normativity is best understood as involving reasons or apparent reasons. In the conflict between Naturalist and Non-Naturalist theories, reasons provide the decisive battlefield. If Naturalists can successfully defend some motivational account of reasons, they could claim to give a single, unified account of both reason-involving and motivational normativity. But if Naturalism fails as an account of reasons, it will also fail, I believe, elsewhere.

83 Analytical Subjectivism about Reasons

Of those who give Naturalist accounts of reasons, many are *Analytical Subjectivists*. On Williams's account, for example, when we say that

> (A) someone has a reason to act in a certain way,

we often mean something like

> (B) this act would fulfil one of this person's present fully informed telic or non-instrumental desires,

or

> (C) if this person knew the relevant facts, and deliberated rationally, this person would be motivated to act in this way. [665]

When people have reasons in what Williams calls this 'internal' sense, we can call these *internal reasons*. (B) and (C) state different claims, either of which might be true without the other's being true. But we can here combine these claims, and consider only cases in which they are both true.

Many other writers give such *Internalist* accounts of the concept of a reason. David Falk, for example, defines a reason as a fact belief in which would motivate us. [666] Williams, Falk, and others give similar accounts of the decisive-reason-implying senses

of 'should' and 'ought'. [667] According to this form of Analytical Subjectivism, which we can call

> *Analytical Internalism:* When we say that
>
> > (D) someone has *decisive reasons* to act in a certain way, or *should* or *ought* to act in this way,
>
> we often mean something like
>
> > (E) this act would best fulfil this person's present fully informed telic desires, or is what, after fully informed and procedurally rational deliberation, this person would be most strongly motivated to do, or would choose to do.[668]

This claim defines the *internal* senses of the words 'decisive reason', 'should', and 'ought'.

According to some other people, whom Williams calls

> *Externalists*: We often use words like 'reason', 'should', and 'ought' in irreducibly normative and indefinable senses.

These we can call the *external* senses of these words. [669]

To illustrate the difference between these senses, and these kinds of reason, we can suppose that, in

> *Early Death*, unless you take some medicine, you will later die much younger, losing many years of happy life. Though you know this fact, and you have deliberated in a procedurally rational way on this and all of the other relevant facts, you are not motivated to take this medicine.

When Williams discusses this example, he claims that you have no reason to take this medicine. [670] As he points out, you have no internal reason to act in this way. And Williams claims that there are no external reasons. I believe that there *are* such reasons. On my view, you have a decisive external reason to take this medicine, which is provided by the fact that this act would give you many more years of happy life.

This imagined case also illustrates the difference between the motivational and reason-involving conceptions of normativity. If we use the words 'reason', 'should', and 'ought' in their internal senses, these two conceptions can be combined, since the claim that we have some internal reason is a claim about our desires, or about how we might be motivated to act. If we use these words in their external senses, claims about reasons are not even in part claims about motivation, so these conceptions conflict. On

the Externalist version of the reason-involving conception, normativity is quite different from actual or hypothetical motivating force. In *Early Death,* for example, though you are not motivated to take your medicine even after informed and procedurally rational deliberation, this fact does not even slightly weaken your external reason to take your medicine, nor does it count against the claim that, in the external sense, this is what you ought to do.

In distinguishing between these views, I have assumed that we can use the phrase 'has a reason for acting' in at least two senses, which express different concepts, and refer to different kinds of reason. It might be objected that, when Internalists and Externalists discuss what we have reasons to do, these people must be using the same concept of a reason, and must be disagreeing only in their beliefs about which are the facts that give us reasons. But these people may, I believe, use different concepts. I understand the concept of an *internal reason* as described by Williams, Falk, and others. And I accept Williams's claim that, in *Early Death*, you have no internal reason to take your medicine. Our disagreement is only about external reasons.

When Williams argues that there are no such reasons, his main claim is that Externalists cannot explain what it could mean to say that we have some external reason. [671] I admit that, when I say that we have some reason, or that we should or ought to act in a certain way, what I mean cannot be helpfully explained in other terms. I could say that, when some fact gives us a reason for acting, this fact *counts in favour* of some act. But this claim adds little, since 'counts in favour of' means, roughly, 'gives a reason for'. Williams suggests that the phrase 'has a reason' does not have any such intelligible, irreducibly normative external sense. When he discusses statements about such external reasons, Williams calls these statements 'mysterious' and 'obscure', and suggests that they mean nothing. [672] Several other writers make such claims.

When I suggest that Williams and I use different concepts of a reason, I am assuming that each of us at least knows what he himself means. But that might not be true. People sometimes fail to understand, not only what other people mean, but even what they themselves mean.

It makes a difference here whether the phrase 'has a reason for acting' has only one ordinary sense or meaning, and, if so, what that sense is. Suppose first that this phrase has only one ordinary sense, which is the internal sense. That would give some support to the view that I misunderstand my own thoughts, since I am wrong to believe that I use the phrase 'has a reason' in an intelligible external sense. When I consider *Early Death*, I believe that you have a decisive external reason to take your medicine, though I know that you have no internal reason to act in this way. I cannot, I believe, be so deeply confused that I use the phrase 'external reason' to mean 'internal reason'. But I cannot exclude the possibility that, as Williams suggests, my use of the phrase 'has an external reason' does not really state some belief, which might be either true or false, but merely expresses some vague attitude.

Suppose next that the phrase 'has a reason' has only one ordinary sense, which is the external sense. That would give some support to the view that Williams misunderstands his own thoughts, since he does in fact use 'has a reason' in this external sense. Williams might mistakenly deny that he uses this external sense because he doubts whether we could understand and use any such irreducibly normative concept. Other people have rejected widely used concepts on similar grounds.

There are other possibilities. The phrase 'has a reason' might have two ordinary senses, which are the internal and external senses, or these senses might be used by different groups of people. Either fact would support the view that Williams and I each mean what we think we mean.

As well as distinguishing these senses, we can ask whether we have both kinds of reason. Since it is clear that we have some internal reasons, the most important possibilities can be shown as follows:

	The phrase 'has a reason for acting' has only one ordinary sense which is	
	the internal sense	the external sense
We have only internal reasons	(1)	(2)
We have both internal and external reasons	(3)	(4)

If (1) were true, Externalism would completely fail, since no one would ever have external reasons, nor would Externalists correctly describe the ordinary meaning of claims about reasons.

If (2) were true, Externalists would correctly describe the ordinary meaning of such claims, but these claims would all be false, since no one would ever have external reasons. Though Internalists would misdescribe the ordinary sense of the phrase 'has a reason', they could move to an error theory, claiming that most of us have false beliefs about reasons. Internalists could also claim that, since all reasons are internal, we should revise our normative thinking, by coming to use the phrase 'has a reason' in the internal sense.

Suppose next that (3) is true, because we have both internal and external reasons, but most people use the phrase 'has a reason' only in the internal sense. Internalists would then correctly describe what most of us mean, and make true claims about our internal reasons. Externalists could point out that, as well as having these internal

reasons, we also have external reasons. But Internalists might reply that, since most of us use the phrase 'has a reason' only in the internal sense, it is only Internalists who can help us to answer our questions about what we have reasons to do, and about what we should or ought to do in the reason-implying senses. Internalism is the more important view, these people might claim, because it is only Internalist theories that might tell us what we want to know.

This claim would not, I believe, be justified. What is most important is not whether Internalists discuss the questions that most of us ask, but whether we have external reasons. If most of us use the phrase 'has a reason' only in the internal sense, that might cast some doubt on the view that we have external reasons, since it might be unlikely that so many people have failed to recognize that they have such reasons. But if we do have external reasons, Externalism would not be a less important view if and because Externalists were not discussing the questions about reasons that most people ask. On the contrary, Externalism would then have *more* importance. Instead of merely describing the internal reasons that most of us already believe that we have, Externalists would be truly telling us that we have reasons of a kind that most of us overlook. Most of us would thereby learn new and important normative truths.

Suppose finally that (4) is true, because we have both internal and external reasons, but most of us use the phrase 'has a reason' only in the external sense. Externalists could again claim that theirs is the more important view.

Internalists might give a similar reply. These people might say that, if they are not discussing the questions that most of us ask, that would make *Internalism* the more important view. Instead of merely describing the external reasons that most us already believe that we have, Internalists would be truly telling us that we also have internal reasons, which would be reasons of a kind that most of us overlook. Most of us, Internalists might say, would thereby learn new and important normative truths.

This reply, I believe, fails. As I shall now argue, if we have both internal and external reasons, it is only external reasons that are important.

84 The Unimportance of Internal Reasons

I shall first repeat some definitions. Some normative claim is

> *conceptual* or *linguistic* when this claim is about some normative concept, or the meaning of some normative word or phrase.

One example is the claim that 'morally permitted' means 'not wrong'. Some normative claim is, in my sense,

substantive when this claim both

(a) states that something has some normative property,

and

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

One example is the claim that

(1) illegal acts are wrong.

Some other normative claims are *tautologies*, in the sense that these claims tell us only that something is what it is, or that, if something has some property this thing has this property. An *open* tautology uses the same words twice. One example is the claim that

(2) wrong acts are wrong.

Though (2) states that certain acts would have a certain normative property, (2) is not in my sense substantive. This claim is not significant, since we would all agree that, as we already knew, any acts that are wrong are wrong.

There are also *concealed* tautologies, which use different words or phrases with the same meaning. Such claims are deceptive, since they can seem to be substantive. If I said that

(3) illicit acts are wrong,

this might seem like the substantive claim that illegal acts are wrong. But if I were using 'illicit' to mean 'wrong', (3) would be a concealed tautology, one of whose open forms would be (2).

Let us next use 'desires' as short for 'fully informed telic desires' and 'ideal' as short for 'fully informed and procedurally rational'. We can also use the words 'should' and 'ought' in their decisive-reason-implying senses. We can then say that, according to

Subjectivism about Reasons: Some possible act is

(A) what we have most reason to do, and what we should or ought to do,

just when this act is

(B) what would best fulfil our present desires, or is what after ideal

> deliberation we would choose to do.

According to those Subjectivists who are *Analytical Internalists*, when we make claims like (A), we often mean something like (B).

If (A) and (B) meant the same, Subjectivism about Reasons would not be a substantive normative view. This view would be a concealed tautology, which merely said the same thing in two different ways. To make this clearer, we could restate this view as

> *SR2*: Some possible act is
>
> > (A) what we have most reason to do, and what we should and ought to do,
>
> in the sense that this act is
>
> > (B) what would best fulfil our present desires, or is what after ideal deliberation we would choose to do,
>
> just when this act is
>
> > (B) what would best fulfil our present desires, or is what after ideal deliberation we would choose to do.

SR2 adds nothing to Analytical Internalism. So Analytical Internalists should claim only that, when we say that

> some act is what we have most reason to do, or is what we should or ought to do,

we often mean that

> this act would best fulfil such desires, or is what after such deliberation we would choose to do.

It is not worth adding that some act is of this kind just when this act is of this kind. Everyone knows that something is true just when this thing is true.

Some Analytical Internalists have overlooked the argument that I have just given, since these people assume that they are defending or describing a substantive normative view. Darwall, for example, describes what he calls a 'system of rational norms', which includes the norm that

> (C) we ought rationally to do what we would be motivated to do if we were vividly aware of the relevant facts. [673]

On the view that Darwall describes, however, (C) could not state a substantive norm. According to this view, when we claim that

> we ought rationally to do something

we mean that

> we would be motivated to do this thing if we were vividly aware of the relevant facts. [674]

If these claims meant the same, (C) would be another concealed tautology, one of whose open forms would be

> (D) what we would be motivated to do if we were vividly aware of the relevant facts is what we would be motivated to do if we were vividly aware of these facts.

This is not a substantive claim.

I have just argued that

> (E) if we used these normative words only in the senses that Analytical Internalists describe, Subjectivism about Reasons would not be a substantive normative view.

Someone might add:

> (F) Subjectivism about Reasons *is* a substantive normative view.

Therefore

> We do not use these words only in these senses.

But this second argument would fail. For the reasons that I have just given, Analytical Internalists should reject (F). And some of these Internalists may have already seen that they should reject (F). Williams, for example, never claims that his Analytically Internalist version of Subjectivism is a substantive normative view.

Williams does, however, believe that

> (G) for Analytical Internalism to succeed, this view must explain how we can use the words 'reason', 'should', and 'ought' to make normative claims.

In Williams's words:

> It is essential to any adequate account of 'A has reason to do X' that it should be normative. . .

In defending his account, Williams writes:

> Unless a claim to the effect that an agent has a reason to do X can go beyond what that agent is already motivated to do. . . then certainly the term will have too narrow a definition. 'A has a reason to do X' means more than 'A is presently disposed to do X'. [675]

But this claim, Williams suggests, might mean that A *would be* disposed to do X if A knew some fact, or lost some false belief. In using this notion or concept of a reason, Williams writes, we would be 'adding to, or correcting,' this person's factual beliefs, 'and that is already enough for this notion to be normative'.

Williams here assumes that, when we tell people that they have a reason to do something, we often intend to be giving these people advice. It would seldom be advice to say 'You want to do X', since few people need to be told what they already want to do. But it might often be advice to say

> If you knew what I know, you *would* want to do X.

That is enough, Williams suggests, to make such claims normative.

It is not, I believe, enough. If I say 'Your wine is poisoned', or 'There's an angry bull in the next field', these claims may be intended as advice. But that would not make these claims, or the facts which they report, normative. For some claim to be normative, it must use at least one normative word, or concept.

Williams might reply that, though these claims are not *explicitly* normative, they would be warnings. Similarly, if I say 'This is the sharpest knife', or 'You would enjoy this book', these claims may be recommendations. Such claims, Williams might say, are *implicitly* normative. On Williams's Internalist account, when we say

> (G) You ought to do X,

we often mean something like

> (H) X is the act that would best fulfil your present desires, or is what after ideal deliberation you would choose to do.

Since it might help you to know that some act would best fulfil your desires, or is what after ideal deliberation you would choose to do, claims like (H) could be used to give you advice. That may seem enough to show that Williams's account sufficiently preserves the normativity of claims like (G). Williams might add that, since (H) uses

the normative words 'best' and 'ideal', this claim is explicitly normative.

Williams's account, I shall argue, does not succeed. We can first distinguish between facts that are *normative* and facts that have *normative importance* in the sense that these facts give us reasons. Two examples would be the facts that

(J) your wine is poisoned,

and that

(K) the fact stated by (J) gives you a reason not to drink your wine.

Of these facts, (J) is natural and (K) is normative. But it is (J), the natural fact, which has normative importance, in the sense of reason-giving force. Though (K) is a normative fact, this fact has no such importance. (K) is the second-order fact *that* the fact stated by (J) gives you a reason not to drink your wine. This second-order fact about this reason does not give you any *further* reason not to drink your wine. Similar claims apply to other cases. Whenever some natural fact gives us a reason, there is also the normative fact *that* this natural fact gives us this reason. [676]

It is easy to overlook such normative facts. This mistake is especially likely if, rather than saying that certain natural facts *give* us reasons, we say that these facts *are* reasons. These are merely different ways of saying the same things. But if we say that natural facts of certain kinds are reasons to act in certain ways, we may be led to assume that, to defend the view that there are normative reasons, it is enough to claim that there are natural facts of these kinds. That is not so. We must also claim that these natural facts have the normative property of *being reasons.* And this claim, property, and fact might all be irreducibly normative.

Such normative facts are, I have said, in one way unimportant, since these facts do not give us further reasons. But it has great importance whether we can recognize and think about such facts. Some of us often do things because we believe that we have sufficient or decisive reasons to do them, so it often matters whether these beliefs are true. And if we don't have such beliefs, or we don't think about our reasons, we may often fail to do what we have most reason to do.

Return next to Williams's suggestion that, when we say 'You ought to do X', we often mean something like

> X is the act that would best fulfil your present desires, or is what after ideal deliberation you would choose to do.

As I have said, such claims are in one way like 'Your wine is poisoned' or 'You would

enjoy this book'. These claims might tell you facts that would give you reasons for acting, so such claims could be used to give you advice. But to be able to use such claims in this way, we must have the concept of *advice*. [677] We must be able to understand the thought that

> (L) such facts might give us reasons for acting, and might make it true that we should or ought to act in some way.

We can now ask whether, if we used normative words only in the senses that Internalist like Williams, Falk, and others describe, we could understand such thoughts.

Remember first that, in

> *Early Death*, after ideal deliberation, you are not motivated to take the medicine that would give you many more years of happy life.

When Williams discusses this example, as I have said, he claims that you have no reason to take this medicine. Though Williams might hope that you will take this medicine, he could not honestly *advise* you to act in this way. We cannot claim to be advising people if we tell them to do what we believe that they have no reason to do.

Consider next

> *Revenge*: Someone insults you. After considering the relevant facts in a fully informed and procedurally rational way, you decide to avenge this insult by killing this person, whom you now regard as your enemy. This act would also best fulfil your present fully informed telic desires. You know, however, that if you kill your enemy, you would be caught, and punished with hard labour for the rest of your life.

As I argue in Chapter 3, if we appeal to claims about *procedural* rationality, we cannot claim that such cases are impossible, or can be ignored.

According to those Subjectivists who are Analytical Internalists, you have decisive internal reasons to kill your enemy, and this is what in the internal sense you ought to do. As before, though these Internalists might hope that you will not kill your enemy, they could not honestly advise you not to act in this way. We cannot claim to be advising people if we tell them *not* to do what we believe that they have decisive reasons to do.

It may seem that, in appealing to these imagined cases, I am trying to show that Analytical Internalism has implausible implications. But that is not my aim. As I have said, Williams is right to claim that, in *Early Death*, you have no internal reason to take your medicine. And these Internalists would be right to claim that, in *Revenge*,

> (M) you have decisive internal reasons to kill your enemy, and this is what in the internal sense you ought to do.

This claim is not implausible, but *true*. (M) means that, as we have supposed,

> (N) killing your enemy is what after ideal deliberation you have chosen to do.

In discussing these examples, my aim is to show that such Internalist claims, though true, have no importance.

Such claims can take two forms. According to *Analytical Naturalists*, normative words or claims can be defined or restated in non-normative and naturalistic terms. Though few people now defend such accounts of morality, many people either defend, or take for granted, Analytically Naturalist accounts of reasons. Some of these people are Analytical Subjectivists or Internalists, who assume that, when we claim that

> (A) we have decisive reasons to act in a certain way, or that we should or ought to act in this way,

we often mean something like

> (O) this act would do most to fulfil our present fully informed telic desires, or is what, after deliberating in certain naturalistically describable ways, we would choose to do.

This claim describes what we can call the *Naturalist internal senses* of the words 'reason', 'should', and 'ought'. According to these Analytical Naturalists, when we have decisive reasons to act in some way, and we ought to act in this way, these are natural facts of the kinds described by (O).

Several Internalists defend such a view. Falk, for example, writes:

> in what I have called the motivation sense, 'ought' statements would be about a certain kind of psychological fact. . . What are here called 'natural' obligations would in one sense be facts of nature in their ordinary empirical meaning. [678]

Darwall writes that, on this version of Internalism:

> the question of which. . . reasons there are for us to act, appears at this point to be unavoidably empirical. [679]

He also writes:

> the test of whether a fact is a reason for a person is for the person rationally to

> consider the fact for himself and to notice whether he is motivated to prefer the act. [680]

Describing one such test, he writes:

> When I consider the fact, the motivation lapses. What seemed a reason. . . turned out on further reflection not to be one at all. [681]

When Darwall discovers that he does not have this reason, what he discovers is the natural, empirical fact that he is not motivated to act in some way.

If we used normative words only in these Naturalist internal senses, we could not, I believe, have normative thoughts. To illustrate this point, suppose again that your hotel is on fire, and that you can save your life only by jumping into some canal. Suppose next that I am outside your hotel, which I know to be on fire, and I can see you at some window above the canal. According to these Internalists, if I think

> You ought to jump,

I would be thinking something like

> (P) Jumping would do most to fulfil your present fully informed telic desires,

or

> (Q) After deliberating in certain naturalistically describable ways, you would choose to jump.

But these would not be normative thoughts. (P) is merely a causal claim, and (Q) is merely a psychological prediction.

These Internalists might reply:

> Since the concepts *reason*, *should*, and *ought* are all normative, any account of these concepts, if it is true, preserves their normativity.
>
> Our view gives the true account of these concepts.
>
> Therefore
>
> Our view preserves their normativity.

To assess this reply, we can return to *Revenge*, and to the fact that

> (R) in the Naturalist internal sense, you ought to kill your enemy.

This fact, I have claimed, is not normative. If I agreed that, in this sense, you ought to

kill your enemy, I could honestly add that I was *not* advising you to act in this way. I would mean only that this act would do most to fulfil your present fully informed desires, and is what, after deliberating in certain ways, you have chosen to do.

These Internalists might now reply that, in making this causal and psychological claim, I *would* be advising you to kill your enemy. I would be telling you what you ought to do.

This reply assumes that

> (S) words like 'reason', 'should', and 'ought' have no external, irreducibly normative senses.

If (S) were true, and our normative claims were intended to state facts, these claims might have to state certain natural facts, such as facts about what would fulfil our desires, or about how we might be motivated to act. There might be nothing else for normativity to be. And we might have to admit, as part of the price of making such claims, that in *Early Death* you have no reason to take your life-saving medicine, and that in *Revenge* you ought to kill your enemy.

I believe that

> (T) these words can be intelligibly used in such external, normative senses,

and that

> (U) we can use these senses to make true claims.

Judged by this standard, psychological and causal claims are *not* normative. There is something else, and something better, for normativity to be. Even if I know that killing your enemy would do most to fulfil your desires, and is what after deliberating in certain ways you have chosen to do, I can still believe that you have no external reason to kill your enemy, and that you ought, and have decisive reasons, *not* to act in this way.

Some writers, such as Darwall, defend an importantly different form of Analytical Subjectivism or Internalism. These people claim that, when we say that

> (A) we have decisive reasons to act in a certain way, or that we should or ought to act in this way,

we often mean something like

> (V) this act would *best* fulfil our present fully informed telic desires, or is what, after fully informed and *procedurally rational* deliberation, we would choose to do.

According to these people, since (V) uses the normative words 'best' and 'rational', (V) cannot be restated in wholly naturalistic terms. [682] When these people claim that (A) often means something like (V), they are describing what we can call the *irreducibly normative internal senses* of the words 'reason', 'should', and 'ought'. This form of Analytical Internalism is *not* a form of Analytical *Naturalism*.

If we used these words in these normative internal senses, we could have normative beliefs. And these beliefs might seem to be about what we have reasons to do, and what we should or ought to do. But that would not really be true.

To illustrate this point, we can first consider a new, stipulated sense of the word 'ought'. We might tell other people that, when we claim that some act is

> what someone 'ought to do' in the *unjust-world* sense, we shall mean that this act is what, in an unjust world, this person has chosen to do.

Since the concept *unjust* is irreducibly normative, so is the complex concept that we could express with this new sense of 'ought'. But this new concept is only partly normative, and this concept's normative part is not about what people ought to do. Suppose again that, in *Revenge*, you have chosen to kill your enemy. Since I believe that the world is unjust, I would then agree that

> (W) in the unjust-world sense, you ought to kill your enemy.

Though this claim uses the word 'ought', (W) is merely another way of saying that, in an unjust world, you have chosen to kill your enemy. Such claims are irreducibly normative because they imply that the world is unjust. But these are not substantive normative claims about what people ought to do. Such claims add nothing to the claim that these people have made their choices in an unjust world.

Similar remarks apply to the normative internal sense of 'ought' that Darwall and others use. If we know that

> (X) you have chosen to kill your enemy after fully informed and procedurally rational deliberation,

we could truly claim that

> (Y) in this internal sense, you ought to kill your enemy.

But though (Y) is irreducibly normative, this claim adds nothing to (X). (Y) merely uses different words to restate the claim that you have chosen to kill your enemy after such ideal deliberation. Since (X) is not a claim about what you ought to do, and (Y) means the same as (X), (Y) cannot be a distinct substantive normative claim about what you ought to do.

If we used the word 'ought' only in this normative internal sense, we could ask

> Q1: Which ways of deliberating are procedurally rational, and in other ways ideal?

Our answers to this question would be normative. We could also ask

> Q2: After such a process of ideal deliberation, what would a certain person choose to do?

But we could not ask, as a further, independent question:

> Q3: What ought this person to do?

Given what we meant by 'ought', Q3 would be merely another way of asking Q2. We can therefore claim that

> (Z) if we used the word 'ought' only in this internal sense, we could have substantive normative beliefs about which ways of deliberating are ideal. We could also have beliefs about what, after such deliberation, we or other people would in fact choose to do. But we could not have any distinct substantive normative beliefs about what we or other people *ought* to choose, and *ought* to do.

I shall now summarize some of these claims. According to

> Subjectivism about Reasons: Some possible act is
>
> > (A) what we have most reason to do, and what we should and ought to do,
>
> just when this act is
>
> > (B) what would best fulfil our present fully informed telic desires, or is what after ideal deliberation we would choose to do.

According to Analytical Internalists, we often use (A) to mean something like (B). If

these claims meant the same, Subjectivism about Reasons would not be a substantive normative view, but a concealed tautology, which told us only that, if we acted in the ways described by (B), we would be acting in these ways.

Analytical Internalists might reply that, when they describe the internal senses of the words 'reason', 'should', and 'ought', they are not intending to state a substantive view. Their aim is only to give a true account of some of our normative concepts and claims. And these people might claim that their Internalist account has the feature that Williams calls 'essential', since this account explains how we can use these words and concepts to have and to state substantive normative beliefs about reasons, and about what we or others ought to do.

This claim, I have argued, is not true. According to some of these Internalists, we can restate claims like (A) and (B) in wholly naturalistic terms. If we used the words 'reason', 'should', and 'ought' in these *Naturalist* internal senses, the concepts that these words express would not even be normative. If our conceptual scheme took this impoverished form, we would not be able to give people advice. Nor could we think about what, in normative senses, we ourselves had reasons to do, and should or ought to do. We could still try to fulfil our desires. And there would still be facts that had normative importance, and reason-giving force. But that importance would be unknown to us---as it is unknown, for example, to some active, intelligent cat.

Some other Internalists claim that, since the phrases 'best fulfil' and 'ideal deliberation' are irreducibly normative, so are the internal senses of the words 'reason', 'should', and 'ought'. If we used only these Internalist normative concepts, we could have substantive normative beliefs about what would best fulfil our desires, and about which ways of deliberating are procedurally rational, and in other ways ideal. But we could not have distinct substantive normative beliefs about what we ought to choose, or to do.

85 Substantive Subjective Theories

Subjectivism about Reasons can take other, better forms. According to what we can call

> *the Externalist Subjective Theory*: Some possible act is
>
> > what we have decisive *external* reasons to do, and what we should or ought in the *external* senses to do,
>
> just when, and because,
>
> > this act would best fulfil our present fully informed telic desires, or is

what, after ideal deliberation, we would choose to do.

Unlike Analytical Subjectivism or Internalism, this form of Subjectivism is a substantive normative theory. Though I have objections to substantive subjective theories, which I present in Chapters 3 and 4, these objections are not relevant here.

Some people have assumed that, if we use the words 'reason', 'should', and 'ought' in their external senses, we cannot accept a desire-based or choice-based subjective theory. But as this Externalist Subjective Theory shows, that is not so. It is true that, if we use these words in their external senses, we can coherently deny that there any desire-based, aim-based, or choice-based reasons, thereby rejecting all subjective theories. But that is precisely why, if instead we use these external senses to state some subjective theory, we would be making substantive claims.

Subjective theories about reasons are often called 'Internalist'. But that label is misleading, since the subjective theory that I have just described makes claims about *external* reasons. Internalists might suggest that, according to this theory,

> we have an *external* reason to act in a certain way just when, and because,
>
> > (1) we also have an *internal* reason to act in this way,
>
> in the sense that
>
> > (2) this act would best fulfil our present informed desires, or is what after ideal deliberation we would choose to do.

But (1) would add nothing, and is less informative than (2). It would be clearer to drop the phrase 'an internal reason', which is now confusingly used in two different senses. When Analytical Internalists say that we have some internal reason to act in some way, these people use this phrase merely as an abbreviation of some claim like (2). Some other Subjectivists use the phrase 'an internal reason' to refer to the *external* reasons that these people believe to be provided by the truth of claims like (2). To avoid confusion, we should use the phrase 'a reason' only in its external, irreducibly normative sense. If we use this phrase to refer only to external reasons, we need not call such reasons 'external' or call this sense the 'external' sense. And when we are discussing substantive theories about reasons, some of which may be subjective theories, we need not call these theories 'Externalist'.

86 Normative Beliefs

To be able to understand and accept such substantive theories, we must use words like 'reason', 'should', and 'ought' in their indefinable, irreducibly normative senses. We

can now briefly reconsider whether these words have such senses.

Falk makes several relevant remarks. When we believe that we ought to do something, Falk claims, we may be believing that, if we reflected on the relevant facts, we would want to do this thing. This internal *motivational* sense of 'ought', Falk argues, is the best and most useful sense. It may be objected, Falk writes

> that 'I ought' is different from 'I would want if I first stopped to think'. The one has a normative and coercive connotation which the other has not. [683]

Falk replies that, when we use 'ought' in this sense, we may be talking, not only about what we *would* want, but also about what we would *have* to want. Such claims, Falk writes, meet Kant's criterion of normativity. According to Kant, when we say that we *ought* to do something, we mean that 'we have, contrary to our inclinations, not only a rational but a *rationally necessary* impulse or 'will"' to do this thing.

This reference to rational necessity looks promisingly normative. But this promise is not fulfilled. On Falk's account, an impulse is *rational* if it is one that 'a person would have if he both acquainted himself with the facts and tested his reactions to them'. Such an impulse is *necessary* if it would not be altered 'by any repetition of these mental operations.' Falk continues:

> And this is meant by a 'dictate of reason': an impulse or will to action evoked by 'reason' and. . . one which derives a special forcibleness from [the fact that] no further testing by 'reason' would change or dislodge it. . . A conclusive reason would be one [that is] unavoidably stronger than all opposing motives.

When we ask 'Must I do that?', Falk suggests, we are asking whether there are any facts belief in which would be 'sufficiently compelling to make' us do it. Some act is *rationally necessary* when knowledge of the facts would irresistibly move us to act in this way.

There is, I believe, no normativity here. An irresistible impulse is not a normative reason. Nor is an impulse made rational by its ability to survive reflection on the facts. Even after carefully considering the facts, we might find ourselves irresistibly impelled to act in crazy ways.

Falk himself asks whether, by expanding the motivational sense of 'ought', we could make this sense of 'ought' more obviously normative. Normativity, Falk assumes, belongs most clearly to imperatives or commands. A normative utterance, he writes, 'is one like "Keep off the grass!"' Falk therefore suggests that we might use 'ought' in a sense that combined a psychological prediction and an imperative. On this suggestion, when we say, 'You ought to do X', we might mean

> If you knew the facts, you would want to do X, so do it!

Such a claim might be both normative and true, since our imperative or command would make this claim normative, and our prediction might be true.

Though Falk calls this suggestion 'tempting', he points out that we cannot coherently combine commands with 'appeals to reason'. People could ask 'Are you *advising* me to do this, or are you merely *telling* me to do it?' [684] Some imperatives, Falk notes, do merely give advice, since we can say, 'My advice is: Do X!' But this use of imperatives, he writes, is too weak or 'anemic' to be normative.

Falk then suggests that, when we say 'You ought to do X', we are not merely claiming that you have reasons to do X, in the sense that there are facts belief in which would motivate you. We are also claiming that these facts are *good* or *valid* reasons for you to do X. [685] On Falk's account, however, this second claim would merely repeat our psychological prediction. We would mean only that, if you knew these facts, your belief in these facts really would motivate you. In Falk's words, we want 'the hearer to have the benefit of *experiencing* what we claim'. [686] If you find that you are *not* motivated by these beliefs, these facts would be shown *not* to be good or valid reasons for you. So this attempt to achieve normativity also fails.

It may seem surprising that, when Falk worries that his motivational sense of 'ought' is not normative, his first response is to expand this sense of 'ought' by making it including a command, or imperative. When Falk wrote, however, it was widely believed that normative claims must either be claims about natural facts, such as psychological predictions, or be commands, or expressive utterances such as 'Hurray!' or 'Boo!' Falk briefly mentions the view that we can use sentences containing 'ought' to state what we believed to be irreducibly normative truths. But this suggested sense of 'ought', Falk writes, is 'too nebulous . . . to be meaningful'. [687]

I believe that I use 'ought' in a meaningful, irreducibly normative sense. Suppose again that I am outside your burning hotel, and I believe that you ought to jump into the canal. My belief would not be about what would best fulfil your desires, or about what after ideal deliberation you would choose. Nor would I be merely thinking 'Jump!' I would believe that you have *decisive reasons* to jump, and that if you don't jump you would be making a *terrible mistake*. You *should* and *must* jump.

That, at least, is what I *believe* that I would believe. We have returned to the question whether I may misunderstand my own beliefs. If we claim to use some word that we cannot helpfully define, or explain by using other words, our hearers may doubt that this word means anything. But such doubts may be quite unjustified. Most words cannot be helpfully explained merely by using other words, since most definitions merely use other words or phrases which mean the same. We learn what most words mean, not by using dictionaries, but by living in complex ways on the complex surface

of our planet. That is why some dictionaries contain photographs or drawings of some of the other animals or inanimate objects with which we causally interact, and to which some of our words refer. But we are also intelligent and rational animals, who can think thoughts about what we cannot see, or hear, or touch. Some of these are abstract thoughts, such as thoughts about what it is for events to be in the past or the future, and thoughts about causation, possibility, necessity, or logic. The concepts that such thoughts involve cannot be helpfully explained either by using other words, or by pointing to something, or by using photographs or drawings. It would not be surprising if, as I believe, the same is true of our fundamental normative concepts, such as those expressed by the words 'a reason', 'should', and 'ought'. It is, I admit, unclear how we come to understand such words, and the concepts they express, and how we can recognize such irreducibly normative truths. I shall make one suggestion later. But these unclarities do not give us decisive reasons to conclude that we have no such concepts, or that there are no such truths.

We can make another, stronger claim. We could not have decisive reasons to believe that there are no such irreducibly normative truths, since the fact that we had these reasons would itself have to be one such truth. This point may not refute this kind of scepticism, since some sceptical arguments might succeed even if they undermined themselves. But this point shows how deep such scepticism goes, and how blank this sceptical state of mind would be.

I used to assume that most people have, or at least understand, such irreducibly normative beliefs about reasons and reason-implying *oughts*. As I have said, however, given what some people say and write, this assumption may be false. In arguing against Externalism, for example, Williams writes:

> Blame rests, in part, on a fiction: the idea that ethical reasons. . . must, really, be available to the blamed agent. *He ought to have done it*, as moral blame uses that phrase. . . hopes to say that he had a reason to do it. But this may well be untrue: it was not, in fact, a reason for him, or at least not enough of a reason. [688]

Given what Externalists mean by the claim that someone had a normative reason to act in a certain way, it is irrelevant to reply 'But this was *not in fact* a reason for him'. When Williams writes 'this may well be untrue', he assumes that he is denying what these Externalists are claiming. But that is not so. Williams's objection should instead be that, as he often says, he doesn't understand such claims, and he doubts whether they make sense. [689]

Darwall writes:

> The case for internalism is especially compelling when we apply it to reasons. . . Unless we suppose that a fact's being a reason has something to do with its

> capacity to motivate, perhaps under some kind of ideal consideration of it, there seems no alternative to supposing that it consists in some kind of non-natural property. And if we are willing to accept that, the resulting picture of rational motivation is an alien and unsatisfying one. It fails to make the desire to act for reasons intelligible as one that is central to us and not simply a superadded fascination with a non-natural metaphysical category. [690]

If Darwall had my concept of a reason, he would not make such claims. When I believe that I have a decisive reason to do something, and that I should and ought to do it, it is not *unintelligible* how these beliefs might arouse in me a desire to act for this reason.

CHAPTER 25 NON-ANALYTICAL NATURALISM

87 Moral Naturalism

There are, I have said, two kinds of Naturalism. According to Analytical Naturalists, though we can distinguish between normative and naturalistic words and sentences, this distinction is fairly superficial. All normative words can be defined by using purely naturalistic words, and normative and naturalistic sentences can state the same claims, which state the same facts. If we used normative words only in the senses that these Naturalists describe, we could not, I have argued, have any normative beliefs.

According to *Non*-Analytical Naturalists, we use some words and make some claims that are irreducibly normative, in the sense that these words or claims cannot be defined or restated in non-normative terms. When we turn to facts, however, there is no such deep distinction. All facts are natural, but some of these facts are also normative, since we can also state these facts by making irreducibly normative claims.

This kind of Naturalism may seem to be obviously mistaken, since it may seem impossible that irreducibly normative claims might state natural facts. As we shall see, however, some people defend Non-Analytical Naturalism in plausible and impressive ways.

Most of these people make claims that are not about reasons, but about morality. We can start by considering such claims. *Normative* Naturalism, as I have said, is often derived from *Metaphysical* Naturalism. Most Naturalists assume that, if there are any moral properties and facts, these would have to be natural properties and facts. Nicholas Sturgeon, for example, writes: 'I take natural facts to be the only facts there are. If I am prepared to recognize moral facts, therefore, I must take them, too, to be natural facts. ' Michael Smith writes that, since 'there are no non-natural properties. . . moral properties. . . must just be natural properties.' [691] Richard Boyd even writes that 'goodness is probably a physical property'. [692] [693]

Some of these writers argue that some form of Moral Naturalism must be true. Consider first those simple, *monistic* moral theories which make claims like

> (A) acts are morally right if and only if, or *just when*, these acts have a certain natural property.

If some such claim were true, the concept *right* and some other, naturalistic concept would be *necessarily co-extensive*, in the sense that these two concepts would necessarily apply to all and only the same acts. Some Naturalists claim that

(B) when two concepts are necessarily co-extensive, these concepts refer to the same property. [694]

When combined with (B), claims like (A) imply that moral rightness is the same as some natural property. For example, if it were true that

(C) acts are right just when they maximize happiness,

the concepts *right* and *maximizes happiness* would apply to all and only the same acts. (B) and (C) would together imply that being an act that maximizes happiness is the same as being right, or is *what it is* for an act to be right. Similar remarks apply to other, more complex, pluralistic moral theories. When combined with (B), these theories would imply that rightness is the same as, or consists in, some set of natural properties.

This argument does not, I believe, succeed. When we consider two concepts that both refer to natural properties, (B) is plausible, and might be true. But when applied to some other pairs of concepts, (B) is not, I believe, true. Consider first the arithmetical concepts expressed by these phrases:

the only even prime number,

the positive square root of 4.

These two concepts both refer to the number 2, which is---or has the properties of being---both the only even prime number and the positive square root of 4. Consider next the concepts expressed by these similar phrases:

being the only even prime number,

being the positive square root of 4.

These concepts refer, not to the number 2, but to these two properties of this number. These two concepts are necessarily co-extensive, since they refer to properties that are necessarily had only by the number 2. But in the sense of 'property' that is relevant here, these concepts refer to different properties. Being the only even prime number cannot be the *same* as being---or be *what it is* to be---the positive square root of 4. So, when applied to these concepts, (B) is false. [695] We can add that, if (B) were true when applied to such concepts, most of mathematics would be either impossible, or trivial.

Since (B) is false when applied to mathematics, it may also be false elsewhere. And (B) is false, I believe, when applied to pairs of concepts of which one is naturalistic but the other is normative. That is what we should expect, given the ways in which natural and normative properties are related. As I shall argue later, if (B) were true when applied to such concepts, normative theories would be either impossible or trivial. If

(B) is false when applied to such concepts, as I believe, we can reject this argument for Moral Naturalism. If acts were right just when they maximized happiness, this fact would not imply that being an act that maximizes happiness was the same as being right.

Other Naturalists give less ambitious arguments, which claim to show only that Moral Naturalism *might* be true, since moral rightness might be the same as some natural property, or set of properties. Some of these people argue that, given certain moral assumptions, Utilitarianism might defensibly take a Non-Analytically Naturalist form. It is worth asking whether that is true, since similar claims would apply to the more complicated moral theories, or sets of moral beliefs, that most of us find more plausible.

According to this form of Utilitarianism,

> (D) though the concept *right* is different from the concept *maximizes happiness*, these concepts both refer to the same property.

Such a claim may seem obviously mistaken. Given the difference between these concepts, we would expect them to refer to different properties.

These Naturalists would reply that, though different concepts usually refer to different properties, there are some important exceptions. Many of these people appeal to analogies drawn from the history of science. Two examples are the discoveries that water is H2O and that heat is molecular kinetic energy. These facts had to be discovered because they were not implied by the pre-scientific meanings of the words 'water' and 'heat'. We might similarly discover, these Naturalists argue, that rightness is some natural property or set of properties.

These arguments have one true premise. Naturalists can claim that

> (E) some irreducibly normative words and concepts might refer to natural properties.

To defend (E), moreover, there is no need to use analogies from the history of science. We can appeal directly to certain irreducibly normative words, and to the concepts that these words express. One example is the concept expressed by the phrase:

> *the natural property that makes acts right*.

Suppose that, as Utilitarians claim,

> (F) acts are right just when, and because, they maximize happiness.

It would then be true that

(G) being an act that maximizes happiness is the natural property that makes acts right.

This claim would use an irreducibly normative concept to refer to the natural property of maximizing happiness. [696]

If (G) were true, however, that would not support Moral Naturalism. Though (G) would use a normative concept that referred to a natural property, this claim would be merely another way of stating the normative claim that is stated by (F). And this claim might state an irreducibly normative fact.

In making these remarks, I have used a distinction that is both of great importance and surprisingly often overlooked. If we claim that

(H) some natural property is the property that *makes* acts right,

we are not claiming that

(I) this natural property is the property of *being* right.

In explaining this distinction, we can first note that, when some act has some natural property which makes it right, this act's having this property does not *cause* it to be right. Though there are several views about the nature of morality, no view claims that *making right* is a causal relation.

There are several ways in which, when something has some property, this fact may *non-causally* make this thing have some property. If I had a child, for example, that would make me a parent. But having a child would not cause me to be a parent. It could not do that, since causes must be different from their effects, and there are not *two* properties here. Having a child is the same as being a parent---or is *what it is* to be a parent. This truth is *analytic*, in the sense that it is directly implied by the meaning of the words 'child' and 'parent'. But some such truths are not analytic. One example is the truth that, when the molecules in some physical object move more energetically, that makes this thing hotter in the pre-scientific sense. Having such greater energy does not cause this thing to be hotter, but is the same as being hotter, or is *what it is* to be hotter. Heat *is* molecular kinetic energy.

There is another, similar pair of ways in which, when something has some property, this fact may non-causally make this thing have some property. Just as my having a child would make me a parent, so would my having a daughter. But unlike having a child, having a daughter is not the same as being a parent. These properties are different because, even if I didn't have a daughter, I could be a parent by having a son. As before, however, my having a daughter would not cause me to be a parent. The

truth is rather that, if I had a daughter, this would *constitute* my being a parent, and if my daughter was my only child, my having a daughter would be the property in which my being a parent would *consist*. While these truths are analytic, there are also non-analytic truths of this kind. Some of the properties of genes, for example, consist in some of the properties of DNA. And mental states, many people believe, consist in states of the brain. Though having a child is the *same* as being a parent, but having a daughter is merely one of the properties in which being a parent can *consist*, these relations are very similar. And there is little metaphysical difference between the claims that mental states *are* or *consist in* states of the brain.

Return now to *making right* and *being right*. According to some writers:

> If there is only a single natural property that makes acts right, we could claim that, when acts have this property, that is the same as being right, or is what it is for these acts to be right. If instead there are several properties that can make acts right, the rightness of acts would consist in their having one of these properties. Just as my being a parent might consist in my having either a daughter or a son, an act's rightness might consist in its being an act that either saves someone's life, or keeps some promise, or expresses gratitude, and so on.

These claims are, I believe, seriously mistaken. When having a child makes someone a parent, or having greater molecular kinetic energy makes something hotter, these relations hold between some property described in in one way and the *same* property described in another way. That is not true of the relation of *making right*. More exactly, there is a trivial sense in which rightness is the property that makes acts right. This is like the sense in which redness is the property that makes things red, and legality is the property that makes acts legal. It is in a different and highly important sense that, when some act has some other property---such as that of saving someone's life---this fact can make this act right. Being an act that saves someone's life couldn't be the same as being right. Nor, I believe, could it be one of the properties in which the rightness of acts consists. When some property of an act makes this act right, this relation holds between two quite different properties. That is why, if it were true that

> (G) being an act that maximizes happiness is the property that makes acts right.

this truth would not support Moral Naturalism. (G) does not imply that

> (J) being an act that maximizes happiness is the same as being right.

(G) implies that

> (K) when acts maximize happiness, that makes these acts right by giving them the different, normative property of being right. [697]

(K) can be used to state a *Non*-Naturalist form of Utilitarianism, of the kind that

Sidgwick defends.

These remarks do not refute Moral Naturalism. These Naturalists might still argue that moral rightness is, or consists in, one or more natural properties. But these Naturalists must defend such claims in a different way. They must argue that, like the concept expressed by the phrase *the properties that make acts right*, the concept *right* might refer to one or more natural properties.

This claim, however, is harder to defend. Return to the pre-scientific meaning of the word 'heat'. In the relevant sense, 'heat' means, roughly:

> the property, *whichever it is*, that can have certain effects, such as those of melting solids, turning liquids into gases, causing us to have certain kinds of sensation, etc.

This concept, we can say, has an *explicit gap* that is waiting to be *filled*, since this concept refers to some property without telling us what this property is. This concept refers to this property indirectly, *as* the property that can have certain effects, such as those of melting solids, etc. This feature of the concept of *heat* allowed scientists to fill this gap, by discovering that molecular kinetic energy is the property that can have these effects.

Similar claims apply to the concept expressed by the phrase:

> the properties, *whichever they are*, that make acts right.

This concept also has a gap that is waiting to be filled, since this concept refers to these properties in a similar, indirect way, *as* the properties that make acts right. That is how, though this concept is irreducibly normative, it might refer to one or more natural properties, such as the property of maximizing happiness.

No such claim applies, I believe, to the concept *right*, or the more fundamental concept *wrong*. We can use 'right' and 'wrong' in several definable moral senses, some of which I describe in Chapter 7. The concepts expressed by these senses do not, I believe, have similar explicit gaps that are waiting to be filled, in ways that would allow these concepts to refer to one or more natural properties.

One example is the concept expressed by the word 'blameworthy'. This concept does not refer to some property indirectly, without telling us what this property is. This concept refers directly to the property of being blameworthy. Rather than arguing that this concept might refer to some natural property, Naturalists would have to claim that blameworthiness is a natural property. And this claim would be harder to defend. Though social scientists can discover facts about which are the acts that various people judge to be blameworthy, these are not facts about the blameworthiness of these acts.

As I have also claimed, however, there are senses of 'right' and 'wrong' that cannot be helpfully defined in other terms. When some concept is indefinable, it does not, like the pre-scientific concept of *heat*, have an explicit gap that is waiting to be filled. But some Moral Naturalists put forward arguments of a similar though looser kind. According to these people, though we cannot define the concepts that are expressed by these senses of 'right' and 'wrong', we can describe the roles or functions that these concepts have in our moral thinking. By appealing to some such functionalist theory, these people argue, we may be able to show that these concepts refer to one or more natural properties. [698]

Though such arguments are ingenious and in some ways plausible, they could not, I believe, succeed.

Before defending this belief, I shall briefly describe why this disagreement matters. Sidgwick believed that rightness is an irreducibly normative property. So did some Non-Utilitarians, such as David Ross. Suppose that Sidgwick and Ross are talking to some Utilitarian Non-Analytical Naturalist. This person claims that, though the concept *right* is irreducibly normative, this concept refers to the natural property of maximizing happiness.

Sidgwick might say:

> If your view were true, Ross and I would have wasted much of our lives. We have spent many years trying to decide which acts are right. We both believe that, when acts maximize happiness, that might always make these acts have the different property of being right. I believe that it does, Ross believes that it doesn't. If there were no such different property, as your view implies, Ross and I would both be mistaken. Morality, as we understand it, would be an illusion.

This Naturalist might reply:

> That is not so. You and Ross both asked what it is for acts to be right, and which acts have this property. My view answers both your questions. Rightness is the property of maximizing happiness, and acts are right when they have this property.
>
> I do claim that, when acts maximize happiness, they cannot also have some *different* property of being right. But that does not imply that these acts are not right. Maximizing happiness is the same as being right. And since identity is a symmetrical relation, we can as truly claim that, when acts are right, they cannot also have some different property of maximizing happiness. As that shows, my view does not eliminate morality. On my view, there are certain natural properties and facts which are also *moral* properties and facts. That does not

> make morality an illusion.

Sidgwick might reply:

> You have not seen how deeply you and I disagree. Though you claim that you and I are both Utilitarians, and Ross rejects Utilitarianism, my view is much closer to Ross's view than it is to yours. Your view *does* eliminate morality, as Ross and I both think we understand it. Ross and I both know that some acts have the natural property of maximizing happiness. We believe that we can ask an important further question, which is whether all such acts have the *very* different, *irreducibly normative* property of being right. If your view were true, there would be no such different property, and no such further question. That would be how, in trying to decide which acts are right, Ross and I have would have wasted much of our lives.

As before, these remarks do not refute Moral Naturalism. Sidgwick, Ross, I, and others may have wasted much of our lives.

I have found, to my surprise, that this imagined dialogue baffles many Naturalists. These people repeat that, since Sidgwick wanted to know both what rightness is, and which acts are right, he should be glad to discover that rightness is the property of maximizing happiness. To explain why Sidgwick would not have been glad, I shall use a crude and only partial analogy. [699] Suppose that I believe in God, and I have spent many years trying to decide which religious texts and theologians give the truest accounts of God's nature and acts. You say that, like me, you believe in God. Love exists, you say, in the sense that some people love others. God exists, because God is love. I could reply that, if your view were true, I would have wasted much of my life. I believe that God is the omniscient, omnipotent, and wholly good Creator of the Universe. If God was merely the love that some people have for others, I would have made a huge mistake, and all my years studying religious texts would have taught me almost nothing.

88 Reductive Naturalism

We shall be asking whether, as Non-Analytical Naturalists believe, irreducibly normative claims might state natural facts. We can first try to make this question clearer.

Some fact is natural, on the most common definition, if facts of this kind are investigated or discussed by people working in any of the natural or social sciences. This definition is vague, since there is much disagreement about which kinds of theory or claim should be regarded as scientific. Rather than trying to resolve such disagreements, we can add another definition, which applies only to normative facts.

When we call some normative fact

> 'natural' in the *reductive* sense, we mean that this fact could be restated by making some non-normative and naturalistic claim.

Normative facts are *not* in this sense natural if they are irreducibly normative, in the sense that such facts could not possibly be restated in such ways. This definition is only partial, since it uses the word 'naturalistic'. But when we ask whether some normative fact is also, in this reductive sense, a natural fact, it is often enough to ask whether facts of this kind could be restated by making some non-normative claim. If the answer is No, this normative fact could not be natural in the reductive sense. We wouldn't need to ask whether this fact could be restated by making some naturalistic claim, so the vagueness of the word 'naturalistic' would not matter. And though the word 'normative' is also vague, it is both easier and more useful, I believe, to make this word, and the concept it expresses, more precise. We can do that by using 'normative' in its reason-implying sense, and making further claims about reasons.

We can now say that, according to

> Naturalism: Normative facts are all natural in the reductive sense,

and that, according to

> Non-Naturalist Cognitivism: There are some facts that are not in this sense natural, but irreducibly normative.

Some Naturalists make claims that may seem not to fit these definitions. Sturgeon, for example, defends what he calls 'a naturalistic but non-reductive view of ethics'. But Sturgeon means only that his view is not *analytically* reductive, since he believes that some normative *concepts* and *claims* may not be able to be defined or restated in non-normative terms. Sturgeon does not claim that normative *facts* could not possibly be restated in such terms. [700] On the contrary, he explicitly claims that normative facts might be able to be restated in non-normative terms. Sturgeon illustrates this claim in a familiar way. Though he is not a Utilitarian, Sturgeon claims that if one form of Utilitarianism turned out to be true, because acts are right just when they maximize pleasure, we could define the good as pleasure and the absence of pain, and define the right as what maximizes the good. On this form of Moral Naturalism, rightness would be the natural property of maximizing pleasure. [701]

Though Sturgeon does not reject my reductive definitions of 'natural' and 'Naturalism', other Naturalists might do that. According to what we can call

> *Wide Naturalism*: Normative facts would be natural facts even if such facts were irreducibly normative, because these facts could not possibly be restated in non-normative terms.

Normative *properties* would be natural properties, these Naturalists would similarly claim, even if it would impossible to refer to such properties by using non-normative concepts.

Since Wide Naturalists admit that certain properties and facts might be irreducibly normative, these people would need to explain in what sense these would also be *natural* properties and facts. These Naturalists might appeal again to the standard definition, claiming that such irreducibly normative facts would be facts of a kind that could be investigated or discussed by some natural or social scientist. But this definition would not be helpful here. We would have to ask whether, if someone made claims about certain irreducibly normative facts, such as facts about which acts are wrong, this person would be making these claims *as a natural or social scientist*. [702] It would be difficult for Wide Naturalists to defend the answer Yes except by claiming that these irreducibly normative facts would also be natural facts. So these remarks would not help to explain some sense in which these facts would be natural.

Sturgeon suggests another sense of 'natural' to which Wide Naturalists might appeal. According to

> *the Causal Criterion*: Some fact is natural if such facts help to provide good causal explanations of what are clearly natural facts. [703]

This criterion raises questions that I shall not try to answer here. [704] It will be enough to give some reasons why, as I believe, we need not ask whether normative facts are in this sense natural. First, this criterion is too narrow, since there are many kinds of natural fact that could not help to provide good causal explanations. Second, we cannot assume that, if certain normative facts did provide such causal explanations, that would make them, in some relevant sense, natural facts. If the Universe was created by God, for example, God would play an essential part in the best causal explanation of many natural facts. But this would not show that God is part of the natural world, nor would this be a naturalistic causal explanation. God is a *supernatural* entity, and this would be a supernatural explanation.

We can also understand, I believe, how irreducibly normative facts might be, or might have been, part of the best explanation of many natural facts. Given what we know about the lives of human beings and many other animals, it is hard to believe that the actual Universe, or world, is the best possible Universe, or world. But we can imagine how that might have been true. If the actual world had been the best possible world, in the sense that reality was as good as it could be, this fact might not have been a mere coincidence. Reality might have been this way *because* this way was the best. On the theistic version of this view, God would not merely happen to exist, since God would exist because God's existing is best. On this *Axiarchic View*, goodness would have a very fundamental explanatory role. But that would not make such goodness a natural property, nor would this be a naturalistic explanation. [705]

There is another, more straightforward reason why we need not discuss Wide Naturalism. We are asking whether we ought to accept some form of Non-Naturalist Cognitivism. When these Cognitivists claim that certain normative facts are not natural facts, they mean that these normative facts differ in several important ways from what are most clearly natural facts. The most fundamental normative facts are not, these people believe, contingent, empirically discoverable facts about the actual world. These facts are necessary truths, which would be true in all possible worlds. It could not have been true, for example, that undeserved suffering was not bad. And when these people claim that such facts are irreducibly normative, they mean that these facts are in a distinctive, autonomous category, which cannot be restated in other terms, or reduced to non-normative, empirically discoverable facts. Since Wide Naturalists would accept this claim, they would not reject Non-Naturalist Cognitivism. So, when considering the arguments for and against this form of Cognitivism, we need not ask whether, by appealing to the Causal Criterion or in some other way, these Naturalists could explain some wider sense in which irreducibly normative facts could be claimed to be natural facts.

It is, however, worth mentioning one such wider sense. Wide Naturalists might say that, when they claim that irreducibly normative facts would also be natural facts, they mean that such facts, or our beliefs that there are such facts, would be compatible with a scientific, naturalistic world view. Though I believe that we should reject other, narrower forms of Naturalism, I would be happy to accept this claim. There is nothing in science, I believe, that is incompatible with there being some irreducibly normative facts, such as facts about practical and epistemic reasons. Scientists make progress by responding to various epistemic reasons and appealing to such facts about these reasons.

89 Rules, Reasons, Concepts and Substantive Truths

. . . . [A section to be added here, some of whose claims will be: If we use 'normative' in the rule-involving sense, we can claim that certain facts are both normative and natural. We can give Naturalistic accounts, for example, of what it is for acts to be illegal, dishonourable, or bad etiquette, or for the uses of words to be incorrect. If we use 'normative' in the better, narrower, reason-implying sense, we cannot give such Naturalistic accounts of normative facts. There are no valid arguments with wholly naturalistic premises and normative conclusions. And, like truths about what exists, no substantive normative truths could follow from our concepts or the meanings of our words.] . . .

90 The Normativity Objection

According to

> *Non-Analytical Naturalists*: Though we make some irreducibly normative claims, there are no irreducibly normative facts. When such normative claims are true, these claims state facts that could also be stated by making other, non-normative and naturalistic claims. Such facts are both normative and natural.

Such views, I shall argue, cannot be true. I believe that

> (A) normative and natural facts are in two quite different, non-overlapping categories.

When people claim that there are two such non-overlapping categories, they are sometimes mistaken. According to Vitalists, for example, facts about living things are in a different category from merely physical facts. This claim, we have found, is false, since many mindless living things, such as amoebae or plants, can be entirely understood in physical terms. Other claims of this kind are more controversial. Thus some people claim, while others deny, that conscious experiences are the same as, or consist in, physical events in some brain. This disagreement is about whether such experiences have properties or features that could not possibly be had by physical events. In a similar disagreement, some people claim and others more plausibly deny that mental states are merely dispositions to behave in certain ways.

Some categorial differences are, on any defensible view, too great to be bridged. Rivers could not be sonnets, experiences could not be stones, and justice could not be---as some Pythagoreans were said to have believed---the number 4. [706] To give some less extreme examples, it could not be a physical or legal fact that 7 X 8 = 56, nor could it be a legal or arithmetical fact that galaxies rotate, nor could it be a physical or arithmetical fact that perjury is a crime. It is similarly true, I believe, that when we have decisive reasons to act in some way, and we should and ought to act in this way, this fact could not be the same as, or consist in, some natural fact, such as some psychological or causal fact.

In making that claim, I am appealing to what I mean by the words 'reason', 'should', and 'ought'. Some Naturalists would object that they are not discussing the meaning of our words. When these people claim that normative facts might be the same as, or consist in, natural facts, their claim is not intended to be analytic, or a claim whose truth is implied by what it means. These people might again cite the discoveries that water is H2O and that heat is molecular kinetic energy. When scientists made these discoveries, these Naturalists might say, they were not appealing to the pre-scientific meanings of the words 'water' and 'heat'.

These analogies, I shall argue later, do not support Naturalism. We can note here that, though these discoveries were not implied by the pre-scientific meanings of these

words, these scientists *did* appeal to these meanings. That is why these scientific discoveries were about *water* and *heat*. Of the reductive views that are both plausible and interesting, most are not analytical. But these views must still be constrained by the relevant concepts. These views are not analytical because the relevant concepts leave open various possibilities, between which we must decide on non-conceptual grounds. Many other possibilities are, however, conceptually excluded. Thus, on the pre-scientific concept of *heat*, it was conceptually possible that heat should turn out to be molecular kinetic energy, or should instead turn out to be a substance, as the *phlogiston theory* claimed. But heat could not have turned out to be a shade of blue, or a medieval king. And if we claimed that rivers were sonnets, or that experiences were stones, we could not defend these claims by saying that they were not intended to be analytic, or conceptual truths. Others could rightly reply that, given the meaning of these claims, they could not possibly be true. This, I believe, is the way in which, though *much* less obviously, Normative Naturalism could not be true.

It may next be objected that normative and natural facts cannot be in wholly different categories, since there is no sharp distinction between these two kinds of fact. It is often unclear whether some word is being used in a normative sense. And some words have complex senses that are partly normative and partly naturalistic. Some examples are 'dishonest', 'cruel', 'cowardly', and 'unpatriotic'.

For Naturalism to succeed, however, even the claims that are most purely normative must, if they are true, state natural facts. These claims, I believe, could not state such facts. And deep distinctions do not need sharp boundaries. Black is not white, and day is not night, though there is grey and twilight in between.

If, as I believe, normative facts are in a separate, distinctive category, there is no close analogy for their irreducibility to natural facts. Normative facts are in some ways like certain other kinds of necessary truths. One example are mathematical truths, such as the fact that 7 x 8 = 56. According to some empiricists, this fact is some natural fact, such as the fact that, when people multiply 7 by 8, the result of their calculation is nearly always 56. This view misunderstands arithmetic, and the way in which mathematical claims can be true. Nor could logical truths be natural facts about the ways in which people think. In the same way, I believe, normative and natural facts differ too deeply for any form of Normative Naturalism to succeed.

To give one example, we can remember that, in *Burning Hotel*, you will die unless you jump into the canal. Since your life is worth living, it is clear that

(B) You ought to jump.

This fact, some Naturalists claim, is the same as the fact that

(C) Jumping would do most to fulfil your present fully informed telic desires, or

> is what, if you deliberated in certain naturalistically describable ways, you would choose to do.

Given the difference between the meanings of claims like (B) and (C), such claims could not, I believe, state the same fact. Suppose that you are in the top storey of your hotel, and you are terrified of heights. You know that, unless you jump, you will soon be overcome by smoke. You might then believe, and tell yourself, that you have *decisive reasons* to jump, that you *should, ought to,* and *must* jump, and that if you don't jump you would be making a *terrible mistake*. If these normative beliefs were true, these truths could not possibly be the same as, or consist in, some merely natural fact or facts, such as the causal and psychological facts stated by (C). We can call this *the Normativity Objection*.

This objection, we can add, need not assume that there are some irreducibly normative *facts*. This objection could instead claim only that

> (D) natural facts could not be normative.

Of the people who are Metaphysical Naturalists, because they believe that all facts are natural facts, many would accept (D). Some of these people are Nihilists, or Error Theorists, who believe that normative claims are intended to state irreducibly normative facts, but that all such claims are false, since there are no such facts. There are also many Non-Cognitivists, who believe that normative claims should not be regarded as intended to state facts---except perhaps in some minimal sense. These people believe that, though there are no normative facts, we can justifiably make normative claims, since these claims do not state beliefs, but express certain kinds of attitude. Like Non-Naturalists, both Nihilists and Non-Cognitivists believe that normative claims are in a separate, distinctive category, so that natural facts could not be normative. These people would agree that when I say, with great passion, that you *should*, *ought to* and *must* jump, my claim could not state some natural fact, such as some causal fact or some psychological prediction. Though most Nihilists and Non-Cognitivists are Metaphysical Naturalists, these two groups of people would agree, though for different reasons, that *Normative* Naturalism could not be true.

Though this objection is, I believe, correct, it would persuade few Naturalists. These people would reply that, despite having quite different meanings, normative and naturalistic claims could state the same facts. But we have further arguments to consider.

CHAPTER 26 THE TRIVIALITY OBJECTION

91 Normative Concepts and Natural Properties

We can first look more closely at one of the ways in which many Naturalists defend their view. Allan Gibbard writes:

> normative concepts are distinct from naturalistic concepts: on this score, Moore was right. But normative and naturalistic concepts signify properties of the same kinds: indeed a normative and a naturalistic concept might signify the very same property. What's distinctly normative, then, are not properties but concepts. [707]

Several other people make such claims. These people argue:

> (A) Some irreducibly normative concepts refer to natural properties.
>
> (B) We can use these concepts to make irreducibly normative claims which are about these natural properties.

Therefore

> (C) When such claims are true, they would state facts that were both normative and natural.

(A) and (B), as we have seen, may be true, and the inference to (C) seems plausible. But this inference is not, I believe, justified. When we see *how* these words and concepts might refer to natural properties, we shall see that (A) and (B) do not imply or support (C). (The rest of this section is somewhat technical, however, and could be skipped.)

Consider first these phrases:

> (D) 'the largest planet',
>
> (E) 'being the largest planet'.

Despite their similarity, (D) refers to Jupiter, and (E) refers to something quite different, which is the property of being the largest planet.

The same distinction applies, though in a way that is easier to miss, when we turn from the properties that are had by *objects*, such as the planet Jupiter, to the *second-order* properties that are had by *properties*. As we have just seen,

> the largest planet

is different from

> the property of *being the largest planet*.

In the same way,

> the property that has some other property,

is different from

> the property of *being the property that has this other property*.

When stated so abstractly, this second distinction is slippery, and hard to grasp. But examples may make it clear. Return to the use of 'heat' which means

> the property, whichever it is, that can have certain effects, such as those of melting solids, turning liquids into gases, etc.

More fully stated, 'heat' means

> the property, whichever it is, that has the *different*, second-order property of *being the property that can have certain effects*, such as those of melting solids, turning liquids into gases, etc.

When scientists discovered that heat is molecular kinetic energy what they discovered was that molecular kinetic energy is the property that has this different, second-order property.

Consider next the claim that

> (F) maximizing happiness is the natural property that makes acts right.

If (F) were true, this claim would use an irreducibly normative concept to refer to the natural property of maximizing happiness. So (F) might seem to be the kind of claim for which Naturalists are looking: an irreducibly normative claim which, if true, would state a natural fact. But (F), I believe, is not such a claim. (F) could be more fully stated as

> (G) the property of maximizing happiness has the *different*, second-order property of *being the property that makes acts right*.

And this different property is normative. That is shown by the fact that both (F) and (G) are merely other ways of stating the normative claim that

(H) acts are right just when, and because, they maximize happiness.

So this example does not support Naturalism.

Naturalists might reply that, even if this example does not support their view, there may be other, better examples. There may be other ways in which, by using some normative word or concept which refers to some natural property, we might make a normative claim which states some natural fact.

In asking whether there could be such claims, we can first remember that, when we claim that some word or concept *refers* to some property, we are not thereby claiming that anything *has* this property. Moral Nihilists, for example, would agree that the concept *right* refers to the property of being right, though they believe that no acts are right.

We can next distinguish two ways in which words or phrases can refer to properties. The phrase 'the property of redness' refers *explicitly* to the property of redness, or of being red. The more common word 'red', when used in a claim like 'blood is red', refers to redness *implicitly*, since this claim describes blood as having this property. Return now to the phrase

(I) 'the natural property that makes acts right'.

If there is only one natural property that makes acts right, this phrase would refer explicitly to this natural property. As we have seen, however, (I) would refer to this property indirectly, as the natural property that has the different, second-order normative property of being the natural property that makes acts right. So (I) would *also refer implicitly* to this other, normative property. And (I) would refer to this natural property only *by* also referring to this normative property. Since all claims that use this phrase would refer to this normative property, such claims could not state facts that were natural in the reductive sense. [708]

Similar remarks apply, I believe, to all irreducibly normative words or uses of words, and to the concepts that such words express. No such normative concept could refer *only* to some natural property, or set of properties, since such concepts can refer to some natural property only by *also* referring to some other, normative property. Such concepts might refer to some natural property either as the natural property that *has* some normative property, or as the natural property that is related to some normative property in some other, less direct way. So we can claim that

(J) irreducibly normative concepts all refer, either explicitly or implicitly, to some normative property.

This is why, though it is true that

(K) irreducibly normative words and concepts might refer to natural properties,

this truth does not support Naturalism. As we have seen, Gibbard takes (K) to imply that it is only concepts, not properties, that are distinctly or irreducibly normative. That, I have argued, is not so. Since such normative words and concepts would refer to natural properties only by also referring to such normative properties, (K)'s truth does not help to show that there are no such normative properties. And we have no reason to expect that, as many Naturalists assume,

(L) we could use these words and concepts to make irreducibly normative claims which might state natural facts.

Since such claims would also refer to such normative properties, they would, if they were true, state what were partly normative facts. Such facts might be irreducibly normative. So this common argument for Naturalism fails.

92 The Fact Stating Argument

Naturalists might give other arguments. In considering other possibilities, we can distinguish three kinds of irreducibly normative concept. Such a concept might be

(M) definable in some way that shows how this concept might refer to some natural property,

(N) definable in some way that shows, or gives us reason to believe, that this concept could *not* refer to some natural property,

or

(O) indefinable.

We have just been discussing one concept of type (M): the concept of *the natural property that makes acts right*. As we have seen, though such concepts might refer to natural properties, they would do that only by also referring to some normative property, so these concepts do not provide an argument for Naturalism.

As an example of type (N), I gave the concept *blameworthy*. Other examples are the concepts expressed by these phrases:

being unjustifiable to others,

being disallowed by some principle that no one could reasonably reject,

being an act that gives the agent reasons to feel remorse and gives others reasons for

indignation.

It would be difficult for Naturalists to argue that these concepts refer to natural properties. These people would have to claim, for example, either that

> the concept *being unjustifiable to others* does not refer to the property of being unjustifiable to others,

or that

> though this concept is irreducibly normative, *being unjustifiable to others* is a natural property.

Such claims would be hard to defend. And even if some concepts of type (N) did refer to some natural property, Naturalists would have to argue that these concepts did not *also* refer to some normative property. Such claims would be harder to defend.

The most important normative concepts, however, are of type (O). These concepts are not complex and definable, but simple and not helpfully definable in other terms. Some examples are the concept of *a reason* and the concepts expressed by the indefinable decisive-reason-implying senses of 'should' and 'ought', and the indefinable moral sense of 'wrong'. When concepts are indefinable, that leaves it in one way more of an open question to which properties these concepts refer. And some Naturalists claim that, by appealing to the role or function that these concepts have in our thinking, we might be able to argue that these concepts refer only to certain natural properties. Such an argument might show that irreducibly normative claims, when they are true, state facts that are both normative and natural.

Though some of these Naturalists make interesting and important claims, I believe that no such argument could succeed. To see why, we can first distinguish two senses in which different claims may state the same fact. That is true

> in the *referential* sense when these claims refer to the same things and ascribe the same properties to these things,

and

> in the *informational* sense when these claims give us the same information.

Consider first these claims:

> (P) Shakespeare is Shakespeare.
>
> (Q) Shakespeare and the writer of *Hamlet* are one and the same person.

(R) Shakespeare wrote *Hamlet*.

In the referential sense, (P) and (Q) state the same fact, since both claims refer to Shakespeare and tell us that Shakespeare has the property of being numerically identical to himself. [709] In the informational sense, however, (P) and (Q) state different facts. Unlike (P), (Q) refers to Shakespeare in a way that also tells us that Shakespeare wrote Hamlet. In the informational sense, it is (Q) and (R) that state the same fact.

Consider next:

(S) water is water,

(T) water is H2O.

In the referential sense, these claims state the same fact, since both claims refer to water and tell us that water is identical to itself. If this is how we think of facts, we could not say that (T) states an important scientific discovery, since this fact would be the same as the trivial fact stated by (S). To explain how (T) was an important discovery, we must claim that (S) and (T) give us different information, thereby stating different facts. Unlike (S), (T) refers to water in a way that also tells us about the atoms of which water is composed. Similar remarks apply to

(U) heat is heat,

(V) heat is molecular kinetic energy.

(V) was an important discovery because, in the informational sense, these claims state different facts. By referring to heat in two different ways, (V) tells us about the relations between several different properties.

Many Naturalists claim that, just as we have discovered that water is H2O and heat is molecular kinetic energy, we might discover, or be able to show, that

(W) moral rightness is the same as some natural property.

In the referential sense, however, the fact that would be stated by (W) could also be stated by

(X) this natural property is the same as this natural property,

and this fact would be trivial. To defend their belief that (W) would state an important discovery, these Naturalists must similarly say that, since (W) and (X) would give us different information, these claims would state different facts. So when these people try to show that normative and naturalistic claims might state the same facts, they must use the phrase 'the same fact' in the informational sense. It would not be enough to argue that, since (X) states a natural fact, and (W) and (X) state the same fact, (W) states

a natural fact. If (W) were true, this claim would give us information that we are not given by (X). To defend their Naturalism, these people need to show that this different information is also natural fact. It would be irrelevant that (X) states a natural fact.

We can now argue:

(1) We make some irreducibly normative claims.

(2) According to Non-Analytical Naturalists, when such claims are true, they state facts that are both normative and natural.

(3) If such normative facts were also natural facts, any such fact could also be stated by some other non-normative, naturalistic claim.

Therefore

(4) Any such true normative claim would state the same fact as some other, non-normative claim.

(5) If these two claims stated the same fact, they would give us the same information.

(6) This non-normative claim could not state a normative fact.

Therefore

If these two claims stated the same fact, by giving us the same information, this normative claim could not state a normative fact.

Therefore

Such normative claims could not, as these Naturalists believe, state facts that are both normative and natural.

We can call this *the Fact Stating Argument*.

Premise (1), I have claimed, is true, and is accepted by Non-Analytical Naturalists. (2) describes this form of Naturalism. Since we are using 'natural' in the reductive sense, (3) is true by definition. These premises imply (4). Since these Naturalists must use 'same fact' in the informational sense, they must accept (5). So, if (6) is true, this argument succeeds.

To illustrate this argument, and help us to decide whether (6) is true, we can return to claims about practical reasons and decisive-reason-implying oughts. Most Naturalists accept some form of Subjectivism about Reasons. As before, it will be enough to

discuss the view that

> (A) we have decisive reasons to act in a certain way, and we should and ought to act in this way,

when

> (B) this act would best fulfil our present fully informed telic desires, or is what, after fully informed and procedurally rational deliberation, we would choose to do.

Of the people who accept this view, some believe that, when we make claims like (A), we often mean something like (B). Other Subjectivists defend this view in other ways.

According to some Non-Analytical Naturalists, though (A) and (B) are irreducibly normative claims, such claims, when they are true, state facts that are both normative and natural. [710] For these facts to be natural in the relevant reductive sense, they must be able to be restated by some other, non-normative, naturalistic claim. As before, we can sum up this claim as

> (C) this act would do most to fulfil our present fully informed telic desires, or is what, after some process of deliberation that had certain natural properties, we would choose to do.

These natural properties are the ones that would make this process of deliberation fully informed and procedurally rational. According to these Naturalists, the fact stated by (C) is normative, because this fact could also be stated by the normative claims (A) and (B).

This view, I believe, could not be true. Consider first these claims:

> (D) You drove at a speed of 100 miles an hour,
>
> (E) You drove at a speed of 100 miles an hour, thereby acting illegally.

If these claims gave us the same information, thereby stating the same fact, that would have to be because your act could not have the distinct property of being illegal. Only that would make it true that (E) would not give us any further information. If you *were* acting illegally, so that (E) *does* give us further information, (D) and (E) would not state the same fact in the relevant informational sense.

Similar remarks apply to (A), (B), and (C). If these claims stated the same fact, that would have to be because

> (F) no act could have the distinct normative properties of being what would *best* fulfil our desires, or being what we would choose to do after procedurally

rational deliberation, or being what we *ought* to do.

Only (F) would make it true that (A) and (B) would give us the same information as (C). If acts *could* have such distinct normative properties, these claims would give us different information. But if no act could have such normative properties, as this Naturalist view implies, we would have no reason to believe that claims like (A) and (B) would state normative facts. So this form of Naturalism would be true only if there were no such normative facts.

This objection, I conclude, succeeds. We can argue:

> If claims like (A), (B), and (C) stated the same fact, this fact could not be normative.

Therefore

> Such claims cannot, as these Naturalists believe, state facts that are both normative and natural.

Similar arguments apply to all other forms of Non-Analytical Naturalism.

93 The Triviality Objection

We can next give another, livelier argument. As before, it will be enough to discuss Utilitarianism, since our conclusions would apply to other moral views. All Utilitarians claim that

> (A) when some act would maximize happiness, this act is what we ought to do.

This view can take two forms. Non-Naturalists like Sidgwick claim that

> (B) when some act would maximize happiness, this fact would make this act have the different, irreducibly normative property of being what we ought to do.

Utilitarian Naturalists reject (B), claiming instead that

> (C) when some act would maximize happiness, that is the same as this act's being what we ought to do.

Suppose that you are a Utilitarian doctor. The Ethics Committee of your hospital asks you to imagine that, in

> *Transplant,* you know that, if you secretly killed one of your patients, this person's transplanted organs would be used to save the lives of five other young

people, who would then live long and happy lives.

You admit that, on your view,

(D) you ought to kill this patient, since this act would maximize happiness.

The Ethics Committee is horrified, and recommends that you be dismissed and debarred from any medical post. If you were a Naturalist, you could reply:

> When I claimed that I ought to kill this patient, I was claiming only that this act would maximize happiness. I was not claiming that this act would have the different property of being what I ought to do. On my view, there is no such different property. Being an act that would maximize happiness is the *same* as being what we ought to do. Since I was claiming only that killing this patient would maximize happiness, no one has any reason to reject my claim.

Though this reply might satisfy the Ethics Committee, it also gives us an objection to Naturalism. Normative claims are, in my sense,

> *substantive* when these claims are significant, because we might disagree with them, or they might tell us something that we didn't already know.

Such normative claims are

> *positive* when they state or imply that, when something has certain natural properties, this thing has some other, different, normative property.

When such claims are true, they state positive substantive normative facts.

Utilitarian Naturalists claim both

> (A) When some act would maximize happiness, this act would be what we ought to do,

and

> (C) When some act would maximize happiness, that is the same as this act's being what we ought to do.

We can argue:

> (1) (A) is a substantive claim, which might state a positive substantive normative fact.
>
> (2) If (C) were true, (A) could not state such a fact. (A) could not truly tell us that, when some act would maximize happiness, this act would have the

> different property of being what we ought to do. (A) would be only another way of stating the trivial fact that, when some act would maximize happiness, this act would maximize happiness.

Therefore

> This form of Naturalism is not true.

We can call this *the Triviality Objection*.

This objection might be misunderstood. We are not claiming that this form of *Naturalism* is trivial. (C) is a substantive claim. And (C) is, in one way, normative, since this claim is about the property of being what we ought to do. But (C) is a *negative* normative claim, since (C) implies that, when some act would have the natural property of being what would maximize happiness, this act could *not* have the *different*, normative property of being what we ought to do. If (C) were true, there would be no such different property. Though (C) is a substantive claim, we are arguing that, if (C) were true, *(A)* would be trivial. Since (A) is not trivial, (C) is not true.

In response to this argument, these Naturalists might first challenge premise (2). These people might say:

> (3) If (A) and (C) were true, these claims would not merely tell us that, when some act would maximize happiness, this act would maximize happiness. In telling us that we *ought* to act in this way, these claims would give us further information about such acts. So (A) and (C) are both positive substantive normative claims.

Any such further information must be statable, however, as the claim that such acts would have one or more other, different properties. And these Naturalists are trying to show that (A) and (C) are substantive *normative* claims. So, to defend (3), these people would have to defend the claim that

> (4) (A) and (C) would both state or imply that, when some act would maximize happiness, this act would have some other, different, normative property.

It is not obvious what this other property could be. When we ask which is the best candidate for the different, normative property which (A) might tell us that such acts would have, the obvious answer is: the property of being what we ought to do. By claiming (C), these Naturalists lose this obvious candidate, since (C) denies that being what we ought to do is a *different* property. To defend (4), these people would have to find some other normative property to play this role. We can call this the *Lost Property Problem*.

There is another problem. If these Naturalists could find some other normative

property to play this role, they would have to apply their Naturalism to this other property. These people would have to claim that, when some act would have certain natural properties, that would be the same as this act's having this other, normative property. These people would then have to defend another version of (4), which referred to some *other*, different, normative property. They would then have to apply their Naturalism to this other property, and so on for ever. That is clearly impossible.

These Naturalists might next give a simpler reply. According to

> (C) When some act would maximize happiness, that is the same as this act's being what we ought to do.

These people might say

> (5) (C) is a positive substantive normative claim because, if (C) were true, this claim would tell us what we ought to do.

In his defence of Naturalism, Gibbard defends (5). Gibbard argues that, if (C) were true, (C) would both tell us that we ought to maximize happiness, and explain why we ought to act in this way. Utilitarians would not need to claim that, when some act would maximize happiness, this fact would make this act have a *different*, normative property of being what we ought to do. In Gibbard's words:

> The properties are one and the same, and that explains, at base, why to do the things we ought to do. . . A further property of being what one ought to do would add nothing to the explanation. [711]

It is not clear what explanation Gibbard has in mind. If (C) explained why we ought to maximize happiness, what would this explanation be?

Utilitarian Naturalists might say

> We ought to maximize happiness because, when we use the phrase 'what we ought to do', we are referring to the property of being an act that would maximize happiness.

As Moore remarks, however, when we ask why we ought to do something, the answer cannot merely be that the word 'ought' 'is generally used to denote actions of this nature.' [712] We must know what it *adds* to claim that we ought to act in this way. So we can say:

> We already know that some acts would maximize happiness. What else does your view tell us about such acts? Which other property would such acts have?

Gibbard's Naturalists might reply:

> When we ask what we ought to do, why do we need to be asking about the relation between *different* properties? Why isn't it enough to learn that there are some things that we ought to do, because some acts would maximize happiness, and that is the *same* as being what we ought to do?

These properties could not, I believe, be the same. But we can try to suppose that these properties are the same, and ask what would then follow. If we learnt that there was only one property here, we would indeed be learning something. We would be learning that, when some act would maximize happiness, this act could not also have the different property of being what we ought to do. Since this information would be purely negative, however, it would not support Gibbard's view. If these Naturalists are not claiming that such acts have some other property, they are not giving us any positive information. And if this claim gives us no such information, it cannot be a positive substantive claim about what we ought to do. Utilitarian Naturalism could not, as Gibbard claims, tell us what we ought to do, nor could it explain why we ought to act in this way.

Though this objection is, I believe, decisive, it may fail to convince some Naturalists. We may also need to explain why this form of Naturalism can seem plausible. It may seem that, to learn what we ought to do, it *would* be enough to learn that some acts would maximize happiness, which is the same as being what we ought to do.

We are comparing two versions of Utilitarianism. Non-Naturalists like Sidgwick claim that

> (B) when some act would maximize happiness, this fact would make this act have the different, irreducibly normative property of being what we ought to do.

Utilitarian Naturalists claim instead that

> (C) when some act would maximize happiness, that is the same as this act's being what we ought to do.

When Gibbard defends his view, he compares (C) with the discovery that water is the same property as H2O. Other Naturalists appeal to the discovery that heat is the same property as molecular kinetic energy. Such analogies can make Naturalism seem plausible. But if we look more closely, I believe, we find that these analogies fail.

True claims about the *identity* of some property use two words or phrases that refer to the same property, and tell us that this property is the same as itself. When that is *all* that such claims tell us, these claims are not substantive, but trivial tautologies. We already know that every property---like everything else---is the same as itself. But some of these claims use certain concepts that enable them also to state important facts.

That is true of the claim that

> (E) molecular kinetic energy is the same as heat.

This claim gives us important information because the concept of *heat* is the concept of the property that is related in certain ways to certain other, different properties. (E) can be restated here as

> (F) having molecular kinetic energy is the property that can make an object have the different properties of being able to melt solids, turn liquids into gases, cause certain sensations, etc.

As a Non-Naturalist, Sidgwick could restate his view in the same way, by claiming

> (G) being an act that would maximize happiness is the property that makes an act have the different, irreducibly normative property of being what we ought to do.

If (G) were true, this claim would also tell us about the relation between different properties.

Return next to Gibbard's suggestion that Utilitarian Naturalism is like the claim that

> (H) the property of being H2O is the same as the property of being water. [713]

This claim, as Gibbard writes, has 'great explanatory power'. When scientists discovered that water is H2O, this discovery explained how some of the previously known properties of water are related to some of the properties of molecules of H2O. [714] Sidgwick's (G) could be similarly restated as

> (J) the property of being what would maximize happiness is the same as the property that makes an act what we ought to do.

Like the fuller, more explicit (G), (J) would also tell us about the relation between different properties. Utilitarian Naturalists claim instead that

> (C) the property of being what would maximize happiness is the same as the property of being what we ought to do.

Unlike Sidgwick's (G) and (J), however, (C) is *not* relevantly like the scientific claims about heat and water. (C) would not tell us about the relation between different properties, so (C) is not a positive substantive normative claim about what we ought to do. As these remarks imply, these scientific analogies do not support Naturalism. On the contrary, by reminding us that substantive claims like (H) and (J) tell us about the relations between different properties, which (C) does *not* do, these analogies count against Naturalism.

There are other ways in which Gibbard's view can seem plausible. (C) may *seem* to tell us what we ought to do. (C) may seem to be a longer statement of the claim that

> (K) maximizing happiness is what we ought to do.

If (K) were true, this claim would tell us what we ought to do. But (C) and (K) are quite different claims. Suppose that some rude person said

> Blowing your nose is what you ought to do.

This person would not mean

> The property of blowing your nose is the same as the property of being what you ought to do.

That claim would be absurd. This person would mean

> Blowing your nose is, or has the different property of being, what you ought to do.

In the same way, (K) means

> (L) maximizing happiness is, or has the different property of being, what we ought to do.

Since (C) implies that there is no such different property, (C) could not be a positive substantive claim about what we ought to do.

There is another, more insidious way in which we can be misled by the claims that some Naturalists make. I believe that, given the meaning of the phrases 'being an act that would maximize happiness' and 'being what we ought to do', it could not possibly be true that

> (C) being an act that would maximize happiness is the same as being what we ought to do.

These two phrases could not refer to the same property. But this very fact can make (C) seem informative. We may think that, if (C) *were* true, this claim *would* be informative, since (C) would tell us about the relation between two different properties. It may therefore seem that, as Gibbard claims, Utilitarian Naturalism might both tell us what we ought to do, and explain why we ought to act in this way.

To avoid being misled, it may help to compare (C) with some other, less plausible claim. Our example can be

> (M) being square is the same as being blue.

This could not be an informative claim. Nor is it worth saying that, *if (M) were true*, this claim *would* be informative. If we were only half awake, it might for a moment seem to us that (M) would be informative, because this claim would tell us about the relation between two different properties. But the fact that makes (M) seem informative also ensures that (M) is false. No claim could truly tell us that two different properties---such as being square and being blue---are one and the same property.

Similar remarks apply, though much less obviously, to (C). Utilitarian Naturalism may seem to be an important view, which might be informative. But what makes (C) seem informative also ensures that (C) is false.

As this comparison can also help to show, when some claim could not possibly be true, it can be misleading to *suppose* that this claim is true, and ask what would then follow. To defend his view, Gibbard might claim

> If being an act that would maximize happiness were the same as being what we ought to do, this fact would explain why we ought to maximize happiness, since maximizing happiness would be our only way of doing what we ought to do.

This claim may seem plausible. But we could similarly claim

> If being square were the same as being blue, this fact would explain why blue things were square, since being square would be the only way of being blue.

Such claims are not worth making.

I shall now summarize these remarks. Naturalists claim that certain normative properties are the same as certain natural properties. To explain and defend this claim, many Naturalists appeal to other claims about the identity of certain properties, such as the claim that heat is molecular kinetic energy or that water is H2O. Claims about the identity of some property are of two kinds. Some of these claims are trivial, telling us only that a certain property is the same as itself. Other such claims, if they are true, also give us important information, by telling us how some property is related to one or more other properties. Most of these Naturalists ignore this distinction. Gibbard recognizes this distinction, but explicitly denies its importance. As we have seen, Gibbard writes:

> The properties are one and the same, and that explains, at base, why to do the things we ought to do. . . A further property of being what one ought to do would add nothing to the explanation.

It is enough, Gibbard suggests, to make claims that are only about a *single* property. This view, I have argued, is seriously mistaken. For such claims to be informative, and worth making, they must tell us about the relation between two or more *different* properties. Only such claims could tell us what we ought to do.

This mistake is easy to make. When Utilitarian Naturalists claim that

> (C) maximizing happiness is the same as being what we ought to do,

they may seem to mean that

> maximizing happiness is what we ought to do,

and these may seem to be claims which, though telling us only about a single property, would thereby tell us what we ought to do. As I have said, however, for it to be true that

> maximizing happiness is what we ought to do,

it must be true that

> maximizing happiness is, or has the *different* property of being, what we ought to do.

Since (C) can easily *seem* informative, we might call this the *Single Property Illusion*, or, in homage to Moore, the *New Naturalistic Fallacy*.

There are some other ways in which Utilitarian Naturalists might defend their view. I have claimed that, since (C) does not tell us about the relation between different properties, (C) could not give us substantive information. These Naturalists might claim that (C) might *indirectly* give us such information.

These Naturalists might first point out that, if (C) were true, Sidgwick's view would be false. Sidgwick would be wrong to claim that, when some act would maximize happiness, this fact would make this act have the different property of being what we ought to do. There would be no such different property. But since this information would be negative, it is not enough to make Utilitarian Naturalism a positive substantive normative view.

These Naturalists might next claim that (C) would also give us positive information. Some of these people argue that, though the concept *ought morally* does not have an explicit gap that is waiting to be filled, we can give an account of the role or function that this concept plays in our moral thinking. By appealing to this account, these

Naturalists might say, we can show that, if (C) were true, this claim would indirectly give us important information. [715] For example, we might learn that

> (N) when some act would maximize happiness, or is what we ought to do, this fact is the same as this act's being justifiable to others, praiseworthy, and something that we have strong reasons to do.

As before, I believe, this claim could not possibly be true. When some act would maximize happiness, that could not be *what it is* for this act to be either what we ought to do, or justifiable to others, or praiseworthy, or something that we have strong reasons to do. But if we are able to conceive that these phrases all refer to the same property, we should conclude that (N) would not then state a substantive normative fact. If *impossibly* these phrases all referred to the same property, (N) would not tell us how this property was related to any other property. So (N) could not give us important positive information.

These Naturalists might instead suggest that, given the role of the concept *ought* in our moral thinking, (C) would indirectly tell us that

> (O) when some act would maximize happiness, this fact would make this act have certain different properties, such as being justifiable to others, praiseworthy, and something that we have strong reasons to do.

This claim *is* substantive, and *would* give us important information. But (O) would not support Naturalism. Utilitarian Non-Naturalists like Sidgwick could happily accept (O), claiming that these other, different properties are all irreducibly normative. To defend their Naturalism, these Naturalists would have to claim that these other normative properties were the same as certain natural properties. The same objections would then apply to these new claims.

Though these remarks have all been about Utilitarianism, similar remarks apply to other moral theories, and to most people's untheoretical moral beliefs. According to any

> *Standard Ought Claim*: When some possible act would have certain natural properties, this act would be what we ought to do.

As we have seen, there are two ways to understand such claims. According to Non-Naturalists, these are positive substantive normative claims, which could be stated more clearly as

> (P) when some act would have these natural properties, this fact would make this act have the different property of being what we ought to do.

Naturalists would reject (P), claiming instead that

> (Q) when some act would have these natural properties, this fact would be the same as, or would constitute, this act's being what we ought to do. [716]

All such views face the Triviality Objection. We can argue:

> (1) Since (Q) does not tell us how these natural properties are related to some other, different, normative property, (Q) is not a positive substantive normative claim.

Therefore

> (2) If Naturalism were true, Standard Ought Claims would be trivial, since these claims could not state positive substantive normative facts.
>
> (3) Such claims are not trivial, and might state such facts.

Therefore

> Naturalism cannot be true.

I have, I believe, sufficiently defended (1), here and in Section 3. (1) implies (2). And most Naturalists would accept (3). This argument, I believe, is sound, and shows that Naturalism is not true. [717]

94 Naturalism about Reasons

We can now apply this argument to claims about reasons. If normativity is best conceived as involving reasons or apparent reasons, as I believe, our main question is whether true claims about reasons might all state natural facts.

In defending his version of Subjectivism about Reasons, Mark Schroeder claims that

> (A) when some fact helps to explain why some act would fulfil one of our present desires, this fact is a reason for us to act in this way. [718]

Schroeder's view takes this form because, unlike many desire-based theorists, Schroeder distinguishes between the facts which are reasons for acting and the other facts about desire-fulfilment which make the first facts be reasons. On Schroeder's view, for example, if

> (B) you want to stay alive,

and

> (C) jumping into the canal would save your life,

the fact stated by (C) would be a reason for you to jump because this fact helps to explain why this act would fulfil your desire. When some fact helps to explain how some act would fulfil some present desire, we can describe this fact as having Schroeder's *explanatory property*.

Unlike many Subjectivists, Schroeder uses the phrase 'a reason' in its indefinable, irreducibly normative sense, which we can also express with the phrase 'counts in favour'. [719] So Schroeder's (A) can be restated as

> (D) When some fact has this explanatory property, this fact is a reason, in the sense of counting in favour of some act.

As a Naturalist, Schroeder also claims

> (E) When some fact has this explanatory property, that is the *same* as this fact's being a reason. Having this explanatory property is *what it is* for some fact to be a reason. [720]

We can argue:

> (1) (D) is a positive substantive normative claim.
>
> (2) For (D) to be such a claim, (D) must state or imply that, when some fact has this explanatory property, this fact also has some other, different, normative property.
>
> (3) If (E) were true, (D) would not be such a claim, since there would be no such different property. (D) would be a trivial claim, which could tell us only that, when some fact has this explanatory property, this fact has this property.

Therefore

> (E) is not true.

As before, this objection might be misunderstood. We are not claiming that Schroeder's view is trivial. Schroeder's (E) is a substantive claim, which many people would reject. We are claiming that, if (E) were true, *(D)* would be trivial.

In response to this argument, Schroeder might first challenge premise (3). He might say:

> (4) If (D) and (E) were true, these claims would not merely tell us that, when

> some fact has this explanatory property, this fact has this property. In telling us that this fact is a reason, these claims would give us further information about such facts. So (D) and (E) are both positive substantive claims.

Any such further information must be statable, however, as the claim that such facts would have one or more other, different properties. And Schroeder would be trying to show that (D) and (E) are positive substantive *normative* claims. So, to defend (4), Schroeder would have to defend the claim that

> (5) (D) and (E) would both state or imply that, when some fact has this explanatory property, this fact has some other, different, normative property.

When we ask which is the best candidate for the different, normative property which (D) might tell us that such facts would have, the obvious answer is: the property of being a reason. But Schroeder cannot give this answer, since (E) denies that being a reason is a *different* property. To defend (5), Schroeder would therefore have to claim that

> (6) when some fact has this explanatory property, this fact has some other, normative property, which is different from the property of being a reason.

Schroeder would then face the Lost Property Problem. It is hard to see what this other property could be. And if Schroeder could find some other property that could be the normative property to which (6) refers, he would have to apply his Naturalism to this other property. He would then have to defend another version of (5) by defending another claim like (6), and he would then have to find yet another normative property to which this version of (6) might refer, and so on for ever. That would be impossible.

Since Schroeder could not defend (4) and (5), he might deny (2). Schroeder might say

> (7) For (D) to be a positive substantive normative claim, this claim need not imply that, when some act would have certain natural properties, this act would also have some other, different, normative property. It is enough that, if (D) were true, this claim would tell us when some fact is a reason to act in some way.

To assess (7), we can turn to some imagined cases. According to Schroeder's

> (D) when some fact helps to explain why some act would fulfil one of our present desires, this fact is a reason for us to act in this way, in the sense of counting in favour of this act.

If (D) was Schroeder's only claim, his view would have some implausible normative implications. As Schroeder points out, (D) implies that we might have a reason to act in some crazy way, such as trying to eat our car, since we might have some present

desire that this act would fulfil. This imagined case, Schroeder assumes, casts doubt on his view, since it is hard to believe that we could have a reason to try to eat our car. Schroeder therefore tries to show that this desire-based reason would be 'of about as little weight as any reason could possibly be'. If this reason is extremely weak, Schroeder writes, that would reduce the 'unintuitiveness' of his view. [721]

Schroeder also claims, however, that

> (E) when some fact has this explanatory property, that is the same as this fact's being a reason.

If we accepted (E), we ought not to think it *unintuitive* or *implausible* to claim that we might have this desire-based reason to try to eat our car. Nor should we think it implausible to claim that we might have such desire-based reasons to act in other crazy ways, such as causing ourselves to be in agony for its own sake. If (E) were true, Schroeder could say:

> (8) When my view asserts that these facts would be reasons for us to act in these crazy ways, I am claiming only that these facts would help to explain how these acts would fulfil one of our present desires. I am not claiming that these facts would have the *different*, normative property of counting in favour of these acts. On my view, there is no such different property. When some fact has this explanatory property, that is the *same* as this fact's being a reason, in the sense of counting in favour of some act. Since these facts *would* have this explanatory property, and that is all that my view implies, these cases give us no reason to reject my view.

Schroeder's (E) is a substantive claim, which is in one way normative, since this claim is about the normative property of being a reason. But, as (8) illustrates, (E) is a *negative* normative claim. And if (E) were true, Schroeder's (D) would not be a positive substantive normative claim. (D) would be trivial, since (D) could tell us only that, when facts have this explanatory property, these facts have this property. When Schroeder claims that certain facts would be desire-based reasons for us to act in crazy ways, he would not be claiming that, as well as having his explanatory property, these facts would have some other, normative property. Since these claims could not conflict with anyone's normative intuitions, Schroeder would not need to argue that these desire-based reasons would be very weak. Rather than trying to show that, even in these cases, his view does not have intuitively unacceptable implications, Schroeder could point out that his view has *no* positive substantive normative implications.

Why does Schroeder believe that his view *does* have such implications? The answer may in part be that Schroeder uses the phrase 'is a reason' in the normative sense that

we can also express with the phrase 'counts in favour'. When we ask whether Schroeder's (D) is true, we can ask

> Q1: When some fact helps to explain how some act would fulfil one of our present desires, would this fact be a reason for us to act in this way, in the sense of counting in favour of this act?

This question, we might assume, could be restated as

> Q2: When some fact has the property of helping to explain how some act would fulfil one of our present desires, would this fact also have the different property of counting in favour of this act?

On Schroeder's view, we might assume, the answer is Yes. In certain cases, as we have seen, this answer may seem implausible. We may find it hard to believe that such explanatory facts could count in favour of our trying to eat our car, or causing ourselves to be in agony for its own sake. Schroeder himself seems to assume that, in asking Q1, we would be asking Q2. Schroeder is worried by such imagined cases. When he supposes that some fact would explain how some crazy act would fulfil one of our desires, Schroeder finds it implausible to claim that, as his view implies, this fact would count in favour of this crazy act.

On Schroeder's view, however, *there is only one property here*. According to Schroeder's (E), when some fact has this explanatory property, that is the *same* as this fact's counting in favour of some act. When we ask whether Schroeder's view is true, we should therefore state our question, not as Q2, but as

> Q3: When some fact helps to explain how some act would fulfil one of our present desires, would this explanatory property be the same as the normative property of counting in favour of this act?

The answer, I believe, is No. This explanatory property could not be the same as this normative property. As we have seen, Schroeder himself seems to believe that the answer to Q3 is No. When he worries about his imagined cases, Schroeder seems to be assuming that these explanatory and normative properties would not be the same. Though Schroeder's view implies that there is only one property here, Schroeder's worries seem to show that he does not really accept his own view.

Schroeder's view *is* intuitively implausible. But what is implausible is not, as Schroeder assumes, some of his view's positive normative implications. What is implausible are the features of Schroeder's view which prevent this view from having any such implications. As we have seen, Schroeder seems to believe that his view has such implications. That is why he argues at length, and with great ingenuity, that his view's normative implications are not as implausible as they may seem to be. As

Schroeder notes, however, his view 'analyzes reasons. . . in wholly non-normative terms'. [722] Since non-normative claims could not state substantive normative facts, Schroeder should admit that, on his view, claims about reasons could not state such facts.

Schroeder might accept this conclusion. He might turn to the view that, since there are no such normative facts, claims about reasons *are* trivial. As we shall see, some Naturalists believe that we need not make such claims.

Schroeder could defend more of his beliefs, however, if he revised his theory in a different way. He could claim instead that, when some fact explains how some act would fulfil one of our present desires, that makes this fact have the different, irreducibly normative property of counting in favour of this act. Though Schroeder would then cease to be a Naturalist, he could keep his belief that claims about reasons are not trivial. And since this version of Schroeder's view would make substantive normative claims, Schroeder could also keep the impressive arguments with which he defends the normative implications of his desire-based subjective theory.

Other Naturalist Subjectivists should, I believe, revise their views in similar ways. On Darwall's view, for example, when we say that

> (F) we ought to act in a certain way,

we often mean that

> (G) this act is what, after fully informed and procedurally rational deliberation, we would choose to do.

Darwall might now add that

> (H) since (F) and (G) are irreducibly normative claims, such claims, when they are true, state irreducibly normative and hence non-natural facts.

If Darwall gave up his Naturalism by accepting (H), this version of Darwall's view would avoid the Fact Stating Argument and the Triviality Objection. As I argued in Section 3, however, if we used (F) to mean (G), that would have one disadvantage. We could have substantive normative beliefs about which ways of deliberating are procedurally rational, but we could not also have distinct substantive beliefs about what we ought to do. If we claimed that we ought to do what, after such deliberation, we would choose to do, this claim would be a concealed tautology. For this reason, I believe, Darwall should also give up his Analytical Subjectivism, by starting to use 'ought' in the indefinable decisive-reason-implying sense. As a Non-Naturalist and Non-Analytical Subjectivist, Darwall could then have substantive beliefs about what we

ought to do, such as the belief that we ought to do whatever, after such fully informed and procedurally rational deliberation, we would choose to do.

If Schroeder and Darwall revised their views in these ways, their subjective theories would be of the kind that I discuss in Chapters 3 and 4. Since we would then be discussing the same questions, we could learn from each other. After further discussion, we might find that we have been climbing the same mountain on different sides.

In giving various arguments against Naturalist accounts of reasons, I have discussed only subjective theories. Some Naturalists might respond to these arguments in a different way. These people might reject Subjectivism about Reasons, and defend some Objectivist view of the kind that I describe in Chapter 2. For Naturalists, however, it is hard to defend such views. It is not surprising that most Naturalists are Subjectivists. These people are also Metaphysical Naturalists, who believe that all properties and facts are natural. Such people have strong apparent reasons to accept a motivational account of normativity. In Darwall's words, which are worth repeating here:

> For the philosophical naturalist, concerned to place normativity within the natural order, there is nothing plausible for normative force to be other than motivational force. . . [723]

By giving a motivational account of normative reasons, Naturalists may seem to explain the normativity of these reasons. If Naturalists appealed instead, not to subject-given desire-based reasons, but to object-given value-based reasons, they would find it harder to defend the claim that their account explains, in naturalistic terms, the normativity of these reasons.

There is another reason why most Naturalists are Subjectivists about Reasons. It is fairly easy to believe that, when we have decisive reasons to act in a certain way, and we should or ought to act in this way, this fact is the same as, or consists in, some fact about what would best fulfil our present informed desires, or about what, after some kind of ideal deliberation, we would be motivated to do, or would choose to do. But if we have reasons that are object-given and value-based, it is implausible to claim that the fact that we have such a reason is always the same as, or consists in, some natural fact.

This claim might seem least implausible if we accept some form of Rational Egoism. Naturalists might then claim that, whenever we have a reason to act in a certain way, this fact would be the same as the fact that this act would promote our own well-being, on some Naturalist account of well-being. But most of us believe that facts of other kinds can give us reasons for acting. And we cannot plausibly claim that, when any

of these other facts gives us some reason, the fact that we have this reason is the same as, or consists in, the natural fact that gives us this reason, or some other natural, non-motivational fact. Suppose for example that

> (I) if I acted in certain ways, I would relieve your pain, and keep some promise, and add to our knowledge of some significant historical event, and help to save Venice from being destroyed.

We may believe that

> (J) these facts would all give me reasons to act in these ways.

The normative facts that are stated by (J) cannot be plausibly claimed to be the same as, or to consist in, the natural facts that are stated by (I). Of the features of Subjectivism that make this view appealing, one is the way in which subjective theories can provide unified accounts of how a great variety of facts can give us reasons. On these theories, the facts stated by (I) might all give me reasons to act in these ways. These facts would give me such reasons if these acts would fulfil some of my present fully informed telic desires, or they are acts that, after some process of deliberation, I would be motivated to do, or would choose to do. If Naturalists are not Subjectivists, there is no similar way in which they could explain how such a great variety of facts can give us reasons. [724]

Suppose next that, despite these two objections, Naturalists seemed able to defend some form of Objectivism about Reasons. Such theories would be refuted, I believe, by the Fact Stating Argument and the Normativity and Triviality Objections. Irreducibly normative claims about reasons could not, if they were true, state natural facts. And such theories would claim that, when some act would have certain natural properties, this fact would be the *same* as the fact that we had reasons to act in this way, or that this act is what we ought to do. Like Naturalist Subjectivism, Naturalist Objectivism would thereby imply that there could not be any positive substantive normative facts about what we have reasons to do, or about what we should and ought to do. We can object that, since there are, or might be, such facts, these theories cannot be true.

95 Soft Naturalism

Though I believe that Naturalism could not be true, it is again worth supposing that I am mistaken, and asking what kind of Naturalism might be true.

Since it is clear that we make some irreducibly normative claims, it could only be Non-Analytical Naturalism that might be true. On this view, such normative claims might state natural facts. But this view can take two forms. According to what we can call

> *Hard Naturalism:* Since all facts are natural, we don't need to make such irreducibly normative claims. The facts that are stated by such claims could all be restated in non-normative and naturalistic terms.

Sturgeon for example, writes that, if some form of Moral Naturalism turned out to be true, we would 'be able to say, in entirely non-moral terms, exactly which natural properties moral terms refer to', and 'moral explanations would be in principle dispensable'. [725] Jackson similarly writes, that, when we have reported the facts in 'descriptive' terms,

> . . . there is nothing more 'there'. . . There is no 'extra' feature that the ethical terms are fastening onto, and we could in principle say it all in descriptive language. . [726]

According to another view, which we can call

> *Soft Naturalism*: Though all facts are natural, we need to make, or have strong reasons to make, some irreducibly normative claims.

Peter Railton, for example, writes that, in giving his Naturalist account of our moral thinking, he hopes to explain 'why morality matters as it does', and hopes to support our belief 'that ethics---real ethics---can be a force in the world'. [727] Darwall is another Soft Naturalist. On Darwall's view, claims about reasons and reason-implying oughts are irreducibly normative. We have strong reasons to make such claims, Darwall assumes, even though these claims, when they are true, state natural facts.

Soft Naturalism is, I believe, an incoherent view. Unlike Non-Cognitivists, Naturalists assume that normative claims are intended to state facts. On that assumption, if we have strong reasons to make irreducibly normative claims, these reasons would have to be provided by the fact that

> (A) there are some important irreducibly normative facts, which we could state only by making such irreducibly normative claims.

If (A) is true, however, Soft Naturalism would fail. In its relevant, reductive form Naturalism is the view that

> (B) all normative facts are also, in the reductive sense, natural facts.

Facts are in this sense natural if they could be restated by making non-normative and naturalistic claims. So (A)'s truth would make (B) false, thereby undermining Naturalism. If instead (B) is true, (A) would be false, and Soft Naturalism would again fail. If all normative facts were also, in the reductive sense, natural facts, Hard Naturalists would be right to say that we don't need to make irreducibly normative claims, since we could state all normative facts by making non-normative and purely

naturalistic claims. This objection we can call the *Soft Naturalist's Dilemma*.

This objection is, I believe, decisive. To illustrate this objection, we can discuss one way in which Soft Naturalists might defend their view.

If all normative facts were also natural facts, that would in part be true because the normative properties that such facts involve were also natural properties. Hard Naturalists would then claim that we don't need to use any normative concepts, since we could refer to these properties by using only non-normative, naturalistic concepts.

Soft Naturalists might reply that, in some other kinds of case, it is important whether we can refer to some property in two different ways, by using two different concepts. It was important to learn that

> (C) heat is the same as molecular kinetic energy,

because (C) tells us how such energy is related to various other properties. Return next to the claim that

> (D) being an act that would maximize happiness is the same as being what we ought to do.

Soft Naturalists might similarly say that, if (D) were true, this claim would not merely tell us that two concepts refer to the same property. Given the difference between these concepts, (D) would also give us further information. That is how (D) would differ from the trivial claim that

> (E) being an act that would maximize happiness is the same as being an act that would maximize happiness.

(D) would give us further information, Soft Naturalists might say, because, unlike (E), (D) uses the normative phrase 'what we ought to do'.

There are now two possibilities. It might be true that

> (F) the further information given by (D)'s use of 'ought' is irreducibly normative.

If (F) were true, Naturalism would be false, since (D) would state an irreducibly normative fact. It might instead be true that

> (G) this further information consists in one or more natural facts.

If (G) were true, Soft Naturalism would fail when applied to such claims. In stating this information, we would not need to use an irreducibly normative sense of 'ought'. As Hard Naturalists believe, the fact that (D) states could be restated by some non-normative and naturalistic claim. [728] So, on both alternatives, Soft Naturalism fails.

We can call this the *Further Information* version of the Soft Naturalist's Dilemma. [729]

This argument's conclusion should not be surprising. All Naturalists believe both that all facts are natural facts, and that normative claims are intended to state facts. We would expect that, on this view, we don't need to make irreducibly normative claims. If Naturalism were true, there would be no facts that only such claims could state.

If there were no such facts, and we didn't need to make such claims, Sidgwick, Ross, I, and others could truly say that we have wasted much of our lives. We have asked what matters, which acts are right or wrong, and what we have reasons to believe, and to want, and to do. If Naturalism were true, there would be no point in asking such questions, since they could not have substantive answers. Our consolation would be only that it wouldn't matter that we had wasted much of our lives, since we would have learnt that nothing matters.

These remarks do not imply that, if Naturalism is false, *Naturalists* have wasted much of their lives. When Naturalists develop theories about *what it is* for acts to be right or wrong, we can often revise these people's theories, so that they instead make claims about what *makes* acts right or wrong, in one or more irreducibly normative senses. When revised, so that they cease to be Naturalist, some of these theories would make plausible and important claims.

I have now defended two main conclusions. First, Naturalism could not be true. We make some irreducibly normative claims, and these claims could not state natural facts.

Second, even if Naturalism were true, *Soft* Naturalism could not be true. There could not be any natural facts that were also important normative facts. If all facts were natural, normative claims could not give us any further information.

Naturalists are not Nihilists, since Naturalists believe that there are some normative facts. But since Soft Naturalism is incoherent, and Hard Naturalism implies that normative facts have no importance, Naturalism is close to Nihilism. So we have reasons to be glad if, as I have argued, Naturalism is not true.

96 Hard Naturalism

Some Naturalists might agree that their view is close to Nihilism. According to these people, when we have reached the true moral theory, we wouldn't need normative concepts. As I have said, Sturgeon writes that, if some form of Moral Naturalism turned out to be true, we would 'be able to say, in entirely non-moral terms, exactly which natural properties moral terms refer to'. Jackson similarly writes 'we could in

principle say it all in descriptive language'. Given their assumptions, these Naturalists are right, I have claimed, to draw this conclusion.

Of those who deny that we need normative concepts, one of the most emphatic is Richard Brandt. Like many other people, Brandt believes that in giving someone advice we should appeal to facts about what this person would want after informed deliberation. Since our actual normative concepts do not explicitly refer to such facts, Brandt claims that we should redefine these concepts. As Brandt writes, 'the question for philosophers is not how normative words are used, for they are used confusedly, but how they are best to be used.'

We can best use these words, Brandt claims, in senses that are wholly naturalistic. When we call some desire 'rational', Brand proposes, we should mean 'fully informed', with 'no further meaning or connotation'. Our desires are in Brandt's sense rational if we would still have these desires even after full reflection on the relevant facts: or what Brandt calls *cognitive psychotherapy*. *We* are rational, Brandt claims, if our desires are in this sense rational, and the most rational thing for us to do is whatever would best fulfil our rational desires. Such an act, Brandt proposes, we can also call 'the best thing to do'.

Brandt compares his proposed senses of the words 'rational' and 'best' with what he calls their 'ordinary' senses. I shall take these other senses to be the ones that are used by those who accept some value-based objective theory about rationality and reasons, and who use 'good' and 'bad' in reason-involving senses. Though I shall use Brandt's word 'ordinary', it does not matter whether these *are* the *ordinary* senses of 'rational' and 'best'. Like Brandt, I am asking how these words can best be used. Value-based objective theories about reasons are the main rival to Brandt's Naturalist, subjective theory. In comparing these theories, we can ask whether, as Hard Naturalists claim, we would lose nothing if we replaced such normative beliefs with beliefs about certain natural facts.

To illustrate his proposals, Brandt first imagines someone with some 'compulsive ambition' that would be extinguished by cognitive psychotherapy. Brandt claims that, on his account, this man's ambition would be rightly called irrational. It is likely that, on plausible objective theories, this man's compulsive ambition would also be claimed to be irrational. To compare Brandt's proposed theory with these other theories, we should turn to cases in which these theories disagree.

As one example, we can suppose that some young woman is afflicted with *anorexia nervosa*. [730] Though this woman knows that she could live a long and rewarding life, her horror of gaining weight makes her prefer to starve herself to death. This preference, we can suppose, would be unaffected by cognitive psychotherapy. On

Brandt's proposals, this woman's preference would then be rational, and starving herself to death would be the best thing for her to do. That would be denied by any plausible objective theory.

After explaining his proposed new senses of the words 'rational' and 'best', Brandt imagines someone who questions these proposals. This sceptic asks

> Q1: Why ought I to want and to do what is in your sense rational?

Brandt claims that, if he cannot answer this question, this fact would not be damaging, since any view could be challenged in the same way. Brandt's imagined sceptic must admit, Brandt writes, that 'the same puzzle arises about knowledge that one "ought" to do something.' Brandt here compares Q1 with

> Q2: Why ought I to do what I know that I ought to do?

But these questions are very different. I might ask Q2 if I knew that I ought to do something, but I didn't know, or had forgotten, what made this true. Such cases raise no puzzle. Suppose next that, though I know both *that* and *why* I ought to do something, I ask why I ought to do this thing. The only puzzle here would be why I asked this question. When we know why something is true, we don't need to ask why this thing is true.

Q1 is a better question. We can ask, for example, why our anorexic woman ought to starve herself to death. Brandt might say 'Because this act is in my sense rational'. That would not be a good enough reply.

Brandt then imagines that his sceptic asks

> Q3: 'Why should I want only those things it is rational in your sense to want?'

Brandt comments:

> similar questions might be raised if we supposed it possible to know, in some other way than by determining what it is rational to want in my sense, which possible outcomes are good or worthwhile.

Brandt's 'similar' question would be

> Q4: Why should I want only those things that are good or worthwhile?

This would be a similarly difficult question, Brandt writes, because 'there is no definitional connection between something's being good. . . and desire.' But there *is* a definitional connection between something's being good in the reason-implying sense and this thing's being *desirable*. Such good things have features that might give us reasons to want them. So Q4 means

> Why should I want only those things whose features might give me reasons to want these things?

Since 'Why?' asks for a reason, this means

> What reasons have I to want only those things that I might have reasons to want?

This question is easy to answer. I couldn't have reasons to want what I couldn't have reasons to want.

Brandt makes other claims that are intended to support his proposed re-definition of the word 'rational'. For example, he writes

> (1) 'a distinctive feature of knowing that a choice would be rational in this sense is that there can be no further question whether it is reasonable to make that choice.'

If (1) uses 'reasonable' in its ordinary sense, this claim's truth *would* support Brandt's proposal. But to defend (1) Brandt writes

> if a man knows what he would choose if he had vividly in mind all the relevant facts. . . the question whether it is rational for him to do this, at least in my sense of rational, is devoid of all sense.

For this remark to be relevant, Brandt's (1) must use 'reasonable' to mean 'in my sense rational'. (1) then claims that, if some choice is in Brandt's sense rational, there can be no further question whether this choice is in Brandt's sense rational. That is true but trivial. Similar claims could be made about any proposed redefinition, however useless.

Brandt also writes

> the question of what I would desire intrinsically if my desires were rational in my sense is a more important question than the question of what is intrinsically desirable, in the ordinary sense, if the two questions really are different.

Since this is Brandt's main claim, his defence of this claim is worth quoting in full. Brandt writes:

> we have a choice as moral philosophers: Whether to recommend that a person make the best choice in the ordinary sense of 'best', or the rational choice in my sense of 'rational'. . .
>
> Consider an example. Suppose I prefer to hear one orchestra program rather than another, in the situation that I know whatever facts might affect my preferences; my preference is then rational in my sense. But suppose someone claims that the opposite preference would be better. Perhaps this could not be

> shown; but since it is an empirical question how 'better' is actually used as applied to such choices, it is logically possible that the opposite preference is the better one in the ordinary sense. The question then arises why one must recommend the preference that is 'better'. Is the fact that it is better a reason for adopting it? The fact that it would be better could not be a new empirical fact that would tend to move my preference in a certain way, for our definition of a 'rational' preference requires that it already have been formed in full view of *all* the relevant empirical facts, including whatever empirical fact is meant by 'the other being better'. One might of course say that some non-natural fact is in question; but, since it is not clear what kind of fact such a non-natural fact might be, I shall ignore this possibility. I concede that perhaps it is tautologously true that it would be better to follow the better preference rather than the rational one if there is a conflict; but this, if true, only re-raises the initial question, why one should take an interest in the better rather than the more rational. It is also true that the expression 'is the best thing' may have *de facto* authority over conduct in the sense that when we decide that something is 'best' in the ordinary sense, our conditioned responses to the phrasing may be such that we incline to do the thing that we have judged best. It may well be that our conditioned responses are firmer and more favourable to 'is the best thing' than to 'is the rational thing' especially when explicitly understood in my sense. But it would be absurd for a person to guide his conduct not by the facts but by the words which may properly be applied to it. My conclusion is that a more rational choice, in my sense, cannot in good reason take second place to a choice which is better in the ordinary sense, if there should be a conflict between the two. [731]

This paragraph illustrates, I believe, much of what went wrong in the moral philosophy of the mid 20th Century.

Brandt starts with an irrelevant example. He supposes that he has a well-informed preference to hear one of two musical programs. Brandt has this preference, we can assume, because he believes that he would enjoy this program more. Brandt then argues that, since this preference is in his sense rational, this would be more important than the fact, if it *were* a fact, that hearing the other program would be, in the ordinary sense, a better choice. If we wished to challenge Brandt's view, we would have to claim that, in the ordinary sense, it would be better and more rational for Brandt to choose the program that he believes he would enjoy less. But that would not be, in the ordinary sense, true. To compare Brandt's view with some value-based view, we must consider cases in which these views conflict. One example is the case of our anorexic woman, who prefers a miserable and early death to a long and rewarding life. Brandt's claim would here be that, since this woman's preference is in his sense rational, this fact is more important than the fact that, in the ordinary sense, it would be better for this woman to prefer a long and rewarding life. That is a much less plausible claim.

Brandt then writes that, if these senses did conflict, the question would arise

> why one must recommend the preference that is 'better.' Is the fact that it is better a reason for adopting it?

The answer to this second question is, strictly, No. If some other preference is better, this fact is not itself a *further* reason for having it. But this does not support Brandt's view. If some preference is better, this fact *is* the fact that we have more reason to have it. That is what this use of 'better' means. So Brandt's first question is easy to answer. We should recommend the preference that is better because this is the preference that we have more reason to have.

If this preference would be better, Brandt continues, this could not be a new empirical fact that would cause us to have this preference. That is true. On the value-based alternatives to Brandt's view, when we have more reason to have some preference, that is not an empirical fact that causes us to have this preference, but an irreducibly normative truth. Brandt mentions this other view, but merely writes that, since it is unclear what kind of fact such a truth might be, 'I shall ignore this possibility'. We cannot defend some view by ignoring the alternatives.

Brandt then writes:

> I concede that perhaps it is tautologously true that it would be better to follow the better preference rather than the rational one if there is a conflict; but this, if true, only re-raises the initial question, why one should take an interest in the better rather than the more rational.

Brandt is here comparing what is better in the ordinary sense with what is more rational in Brandt's sense. Some preference would be better to follow, in the ordinary sense, if we have more reason to follow this preference. So Brandt's sentence should be taken to mean:

> If we have more reason to follow one of two preferences, but the other preference is in my sense rational, it may be tautologously true that we have more reason to follow the preference that we have more reason to follow. But that only returns us to the question: Why should we follow the preference that we have more reason to follow, rather than the preference that is in my sense rational?

Since 'Why?' asks for a reason, this means 'Why do we have more reason to follow the preference that we have more reason to follow?' This question answers itself.

Brandt next suggests that, if we did what we judged to be best, such acts might be merely a 'conditioned response' to the ordinary sense of 'best'. He then writes

> it would be absurd for a person to guide his conduct not by the facts but by the words which may properly be applied to it.

As before, Brandt does not take seriously the value-based alternative to his view. Brandt is here supposing that the ordinary sense of 'best' would be '*properly* applied' to what we do. If that were true, and we did what was best because it was best in this reason-implying sense, our act would not be merely a conditioned response to the word 'best'. We would be guided, not absurdly by mere words, but by our awareness of the facts that gave us decisive reasons to act in this way.

Brandt ends:

> My conclusion is that a more rational choice, in my sense, cannot in good reason take second place to a choice which is better in the ordinary sense, if there should be a conflict between the two.

Choices are better in the ordinary sense if they are choices that we have more reason to make. Brandt is here supposing that one choice would be in this sense better, but that some other choice would be in Brandt's sense rational. So Brandt's conclusion is that, in such cases, the choice that we had less reason to make could not be the choice that we had less reason to make.

Since Brandt is an excellent philosopher, why does he make such claims? The answer seems to be that, even when Brandt says he is supposing that one of two choices would be, in the ordinary sense, better, he is not really doing that. Though Brandt mentions the view that there are irreducibly normative non-natural facts, he writes, 'I shall ignore this possibility'.

If we ignore this possibility, and we use naturalistic substitutes for normative concepts, we can be led to conclusions that seem absurd. As I have said, Brandt would have to claim that our anorexic woman ought rationally to starve herself to death, and that this would be the best thing for her to do.

As before, however, though these claims may seem absurd, this should not be our objection to Brandt's view. As Brandt could reply, his claims about this woman would *not* be absurd. Given his proposed definitions, when Brandt claims that this woman's act would be rational, and would be the best thing for her to do, he would mean only that, in starving herself to death, this woman would be doing what, even after cognitive psychotherapy, she would most want. That claim is true. This woman's act *would* be, in Brandt's sense, rational.

What makes this claim true, however, also makes it trivial. This should be our objection to Brandt's view. When Brandt claims that we ought rationally to do what would fulfil our fully informed desires, he means that, in doing what would fulfil these desires, we would be doing what would fulfil these desires. If we used such naturalistic substitutes for normative concepts, our claims would never be absurd

because they would not be substantive normative claims. We could not significantly claim, or think, that this anorexic woman should *not* starve herself to death.

Brandt's remarks illustrate another point. Though Hard Naturalists claim that we don't need normative concepts, they use such concepts. Brandt rightly claims, for example, that the philosopher's question is how normative words are *best* used. [732] He makes claims about what is *more important*. And in the passage just quoted, Brandt writes that choices that are more rational in his naturalistic sense 'cannot *in good reason* take second place' to choices that are better in the ordinary sense. These are not claims about what we would want after cognitive psychotherapy.

Jackson provides some other examples. We don't need normative concepts, Jackson claims, because there are no irreducibly normative properties or facts. In his words, there is nothing else 'there'. But Jackson also writes:

> . . . it is hard to see how [such] properties could be of ethical significance. Are we supposed to take seriously someone who says, 'I see that this action will kill many and save no one, but that is not enough to justify my not doing it; what really matters is that the action has an extra property that only ethical terms are suited to pick out'? In short, the extra properties would be ethical 'idlers'. [733]

Jackson seems to mean:

> Even if acts could have irreducibly normative properties, such as the property of being wrong, it is hard to see how such properties could have any ethical significance. If some act would kill many people and save no one, this fact is enough to justify our not acting in this way, and enough to give us a sufficient or even a decisive reason not to act in this way. Our reason not to kill these people would not have to be given by the fact that this act would have the extra property of being wrong.

These claims are irreducibly normative, and they would state irreducibly normative truths. On Jackson's view, there cannot be such truths. Jackson also writes that, if the best Naturalist theory turned out to be one form of Hedonism,

> we should identify rightness with maximizing expected hedonic value. . . [because this would be] what. . . we ought to aim at. [734]

If we didn't need normative concepts, as Jackson believes, we would be able to restate this claim without using the words 'should' and 'ought'. But that would be impossible. Jackson might write that, on these assumptions,

> it would maximize expected hedonic value to identify rightness with maximizing such value, because this would be what it would maximize such value to aim at.

But that is not what Jackson means, nor could it be what he ought to mean.

Though some Greek sceptics may have been able, for a while, to use no normative concepts, and to have no normative beliefs, few ordinary people can do that. And most Normative Naturalists make some irreducibly normative claims.

Normative Naturalism, I have argued, cannot be true, because such normative claims could not state natural facts. But there is another way in which normative claims have been held to be compatible with a wholly naturalistic view.

CHAPTER 27 NON-COGNITIVISM AND QUASI-REALISM

97 Non-Cognitivism

According to *Non-Cognitivists*, normative claims are not intended to state facts. When these people reject Naturalism, many of them say that, as I have argued, natural facts could not be normative. Some of these people add that, when Moore criticized what he called 'the Naturalistic Fallacy', he was only half right. Though Moore saw that normative claims could not be claims about natural properties and facts, he mistakenly assumed that such claims must be about *non-natural* properties and facts. That assumption, Non-Cognitivists believe, still underrates the distinctiveness of normative claims. According to these people, it is not merely *natural* facts that could not be normative. *No* facts could be normative, since no facts, or factual beliefs, could have the role in our lives of norms or values. These people distinguish between *facts* and *values*, assuming that there could not be evaluative or normative facts. When we claim that some act is rational or right, these people say, we are not claiming that this act has even a special, irreducibly normative non-natural property. Normativity is to be found, not in the properties of acts, but in our attitudes towards these acts. In Hume's words, we must 'look within'.

There is another, partly overlapping view. According to

> *Moral Sentimentalists*: Morality involves passion rather than reason, or the heart rather than the mind, since our moral convictions are best understood as consisting in certain kinds of desire, sentiment, or other *conative attitude*.

This view can take Cognitivist forms. According to those who are often called

> *Moral Subjectivists*: When we claim that some act is wrong, we mean that we have some disapproving attitude towards this act. [735]

As Sidgwick points out, this view is clearly false. If this view were true, we could not have moral disagreements. If I said 'Stealing is wrong', and you said 'No it isn't', these claims would not conflict, and they might both be true, since we might each be correctly describing our own attitude to stealing. When we make such claims, however, we *are* disagreeing.

According to some

> *Moral Intersubjectivists*: When we claim that some act is wrong, we mean that most people, at least under ideal conditions, would have some disapproving attitude towards such acts.

On this view, acts can be right or wrong in the kind of way in which apples can be red or green, jokes can be funny or feeble, and faces can be beautiful or ugly. Apples are red if they look red to normal observers in daylight, jokes are funny if they amuse most people, and acts are wrong if they would arouse a sentiment or attitude of disapproval in most well-informed and impartial observers. [736]

Though such an Intersubjectivist, *response-dependent* view is clearly correct when applied to colours, and plausible when applied to jokes and to beauty, there are strong objections to such accounts of morality. If I am colour-blind, for example, I might truly claim that two apples have different colours, because they look different to normal observers, though these apples look the same colour to me. According to these Moral Intersubjectivists, I might similarly truly claim that some act is wrong, because most people have a disapproving attitude toward such acts, though I myself approve these acts. That is not how we think about morality. If we approve some act, we cannot also believe that this act is wrong. In response to this objection, Intersubjectivists might say that, when we claim that some act is wrong, we mean that *everyone*, under ideal conditions, would have a disapproving attitude toward such acts. Though this view is more plausible, it also misdescribes how most of us think about morality. When we claim that some act is wrong, we might believe that everyone, under ideal conditions, would disapprove such acts. But that is not what we mean. And we would not believe such acts to be wrong because, under ideal conditions, we would all disapprove of them. We would all disapprove, we assume, because such acts are wrong. If we supposed that some people would not disapprove such acts, we would not take that to show that such acts are *not* wrong.

Sentimentalism can also take Non-Cognitivist forms. According to

> *Moral Expressivists*: When we claim that some act is wrong, we are not intending to say something true, but are expressing our disapproving attitude toward such acts.

On the earliest and simplest view of this kind, *Emotivism*, if we claim that lying is wrong or that we ought to keep our promises, we mean something like 'Lying: Boo!' or 'Keeping promises: Hurray!' Later Expressivists, as we shall see, make more plausible suggestions.

Such Non-Cognitivist views may seem obviously false. When you claim that lying is wrong, I might say 'That's true'. But this *minimal* use of 'true', these writers say, is merely another way of expressing the same attitude. For example, if you said 'Milk chocolate is disgusting', I might say 'That's true' merely as a way of expressing the same dislike.

There are three main arguments for Non-Cognitivism. According to what we can call *the Humean Argument*:

> (A) It is inconceivable that someone might be sincerely convinced that some act was their duty, but not be in the slightest motivated to act in this way.
>
> (B) If moral convictions were beliefs, such a case would be conceivable.

Therefore

> Moral convictions cannot be beliefs, and must be some kind of desire, conative attitude, or motivating state.

To defend (B), some Non-Cognitivists appeal to

> *the Humean Theory of Motivation*: No belief could motivate us unless this belief is combined with some desire.

These people claim that, if moral convictions were beliefs, it would make sense to suppose that we might believe some act to be our duty, without having the desire that would be needed to motivate us to act in this way. Since such a case is not conceivable, these Non-Cognitivists argue, moral convictions must themselves *be* desires. Only that could guarantee that, when we have moral convictions, we are motivated to act upon them. [737]

Some Humeans claim that, for some belief to motivate us, this belief must be combined with some *independent*, *pre-existing* desire. As Nagel argues, we can reject this claim. When we come to have some belief, such as the belief that we ought to act in a certain way, this belief might motivate us by causing us to have some new desire. Nor do we even need to have some new desire. Whenever we act in some voluntary way, Humeans say, we must have wanted to act as we did. But our having this desire, we can reply, might consist only in our being motivated by some belief. [738]

Humeans might accept this reply, and retreat to a weaker view. These people might claim

> (C) No belief could motivate us *all by itself*, since no belief could motivate us unless it is also true that we are *disposed* to be motivated by this belief.

Such dispositions, Humeans might say, are one of the kinds of mental state that they call desires.

In this form the Humean Theory is undeniable, but has less importance than it is often claimed to have. Consider, for example, Kant's anti-Humean claim that pure reason can by itself motivate us. Kant would not have minded claiming that, for pure reason

to be able to motivate us, we must be rational beings, who are disposed to be motivated by pure reason. It is no objection to Kant's view that pure reason could not motivate a snail, or a stone.

Even in this much weaker form, however, the Humean Theory may sufficiently support premise (B). We might have to admit that, if moral convictions are beliefs, it would be conceivable that someone might have some moral belief without being disposed to be motivated by this belief.

We can reject this argument, however, in a different way. Premise (A) is plausible, we can point out, because we would not call someone's moral belief 'sincere', or a 'moral conviction', if this person claimed to believe that some act would be wrong but was not in the slightest motivated to refrain from acting in this way. If we ask instead whether such a person might *know* that such acts are wrong, our answer would be Yes. And in knowing that such acts are wrong, this person must in one sense believe that such acts are wrong. If we revise premise (A) so that it refers to moral beliefs rather than what we call 'sincere moral convictions', (A) ceases to be true, so the Humean Argument fails.[739]

We have other normative beliefs, such as beliefs about what we should or ought to do in the decisive-reason-implying senses. When we consider such beliefs, there is no similarly plausible Humean Argument for Non-Cognitivism. If people are deeply depressed, for example, they may believe that they have decisive reasons to do something, such as acting in some way that would protect their future well-being, without being in the slightest motivated to act in this way. It would be implausible to claim that such people cannot *sincerely* believe that they have these decisive reasons to protect their future well-being. When people are deeply depressed, what they lose may only be their motivation, not their normative beliefs. Such examples support the claim that (A) seems plausible only because (A) uses the phrase 'be sincerely convinced' rather than the word 'believe'.

The second main argument for Non-Cognitivism starts as follows:

> (D) Moral claims cannot be explained or restated in non-normative and wholly naturalistic terms.

Therefore

> (E) If these claims were true, they would state facts that were not natural but irreducibly normative.
>
> (F) All facts are natural.

Therefore

(G) Moral claims could not state facts.

There are now two alternatives. Nihilists continue:

(H) Moral claims are intended to state facts.

Therefore

(I) These claims are all false.

Non-Cognitivists continue:

(J) We can justifiably make moral claims.

Therefore

(K) These claims are not intended to state facts.

Since premise (F) assumes Metaphysical Naturalism, we can call these *the Naturalist Arguments for Nihilism* or *Non-Cognitivism*. Though I believe that we can justifiably reject (F), thereby rejecting both these arguments, I shall say little to defend that belief here.

In its earliest, Emotivist form, Non-Cognitivism was close to Nihilism. I was present when the most notorious 'Boo-Hurray' Theorist, A. J. Ayer, heard John Mackie present his Nihilistic Error Theory. Ayer's first comment was: 'That's what I should have said'. Ayer happily gave up his Non-Cognitivism, turning instead to the view that most people misunderstand morality, since most people mistakenly believe that there are moral truths.

Several later Non-Cognitivists firmly reject any such Error Theory. According to these writers, most of us know, or would on reflection agree, that moral claims are intended, not to state facts, but to express certain attitudes.

Some of these writers, however, make a surprising further claim. According to these Non-Cognitivists, though we do not intend our moral claims to state facts, such claims can, in a way, state facts. Two such writers are Simon Blackburn and Allan Gibbard, who defend partly overlapping Expressivist theories. [740] By asking what these original and impressive theories achieve, we can reach some conclusions that apply to all forms of Non-Cognitivism.

98 Normative Disagreements

The 'key to meaning', Gibbard writes, lies 'in agreement and disagreement: we know what a thought is when we know what it would be to agree with it or disagree with it.' [741]

Moral Subjectivism fails, as we have seen, because this view falsely implies that we cannot have moral disagreements. On this view, apparently conflicting moral claims might all be true. Since Non-Cognitivism avoids this objection, this view is much more plausible. But Non-Cognitivists, I shall argue, also cannot explain what is involved in moral disagreements.

On Blackburn's theory, moral claims do not fundamentally state beliefs, but express certain kinds of desire, value, or other conative attitude. The essential phenomenon, Blackburn writes,

> is that of people valuing things. . . we recognize no interesting split between values and desires. . . we call 'values' just those desires and attitudes that stand fast when we contemplate others and try to alter them. [742]

Such attitudes conflict whenever one person is in favour of some act or policy, and someone else is against this act or policy. Such people disagree, Blackburn claims, in the sense that their desires or other conative attitudes cannot both be fulfilled. [743]

It is misleading, I believe, to describe such people as *disagreeing*. When two people have conflicting desires, they cannot both get what they want. These people may oppose each other, and they may even fight. But fights may not involve any disagreement. For people to disagree, they must have conflicting beliefs.

Gibbard similarly claims that we can disagree with people's preferences and acts. [744] This claim is also misleading. If I believe that one of your preferences or acts was irrational or wrong, you and I may disagree, since you may believe that your preference or act was rational or right. But I would then be disagreeing, not with your preference or act, but with your belief.

Though Gibbard discusses our moral beliefs, his main claims are about rationality, and about what we ought to do in a practical, non-moral sense. To explain 'what *ought* assertions mean', Gibbard writes, we can say:

> the concept of ought just *is* the concept of what to do. [745]

He also writes:

> The hypothesis of this book is easy to state: *Thinking what I ought to do is thinking what to do*. [746]

Gibbard's phrase 'thinking what to do' is ambiguous. If I said that I was trying to

decide what to do, I would often mean that I was trying to decide what I *ought* to do. But this is not what Gibbard means, since that would make his hypothesis trivial. Gibbard means:

> Thinking what I ought to do is thinking about what I *shall* do.

As Gibbard also writes:

> If we understand concluding what to do, then we understand concluding what a person ought to do. [747]
>
> When I speak of concluding 'what to do', understand this to mean coming to a choice. [748]

These claims may correctly describe how, in Gibbard's unusual phrase, he and some other people *conclude what to do*. That Gibbard thinks in this way is suggested by his use of this and similar phrases. Gibbard talks of our 'disagreeing what to do', he calls his book *Thinking How to Live*, he asks 'why to care?', and he writes 'what's obvious is to choose life over death'. [749] [750] To some of us, however, these phrases seem to have a normative word missing. Rather than asking *why to care* about something, we would ask why we *should* care about this thing, or what *reasons* we have to care. Rather than concluding what to do, we reach conclusions about what we should do. And we wouldn't think it obvious to choose life over death. What can be obvious, we believe, isn't to choose something, but only some truth or fact, such as the fact that we *should* or *ought* to choose something, or that something is *the thing to choose*.

If we use these normative words and concepts, Gibbard's suggested view does not, I believe, correctly describe our practical reasoning. When we conclude that we ought to do something, we are not deciding to do this thing, but coming to have a normative belief. Though our decisions to act are often based on such beliefs, these decisions are not the same as our coming to have these beliefs. We always have two questions:

> Q1: What ought I to do?
>
> Q2: What shall I do?

This distinction is clearest when we must make decisions that could not even be based on any normative belief. Such cases take their simplest form when we must choose between two qualitatively identical items. Buridan's imagined donkey, or ass, was given two identical bales of hay. Because this animal was too rigidly rational, being unable to make decisions for no reason, it could not decide which bale to eat, since it had no reason to prefer either bale to the other. So it starved to death.

Return next to the case in which, to escape from the fire in your burning hotel, you must jump into the canal. Suppose that your room has two windows. On Gibbard's

suggested view, if you decide to jump through one of these windows, you would be deciding that this is what you ought to do. That may not be true. You might know that jumping through the other window would be just as good. You wouldn't then believe that you ought to jump through one particular window. But you would still have to decide through which window you will jump.

In many cases, our decisions can be based on normative beliefs. But that does not show that, when we come to believe that we ought to do something, that is the same as our deciding to do it. We may decide *not* to do what we believe that we ought to do, or decide to *do* what we believe that we ought not to do. Gibbard might qualify his view, so that it does not apply to such cases. In response to a similar objection, Gibbard writes 'we'd best look first to thinkers who are consistent'. [751] But even when considering people who are always practically consistent, in the sense that these people always decide to do, and try to do, what they believe that they ought to do, we should distinguish between these people's decisions and their normative beliefs. If we ignore this distinction, we shall misunderstand these people's practical reasoning.

Gibbard claims the opposite. It is *by* ignoring this distinction, he believes, that we can best understand practical reasoning. Gibbard writes:

> I the chooser don't face two clear, distinct questions, the question what to do and the question what I ought to do. [752]

We can best explain the concept *ought*, Gibbard suggests, by describing what is involved in making plans, and in disagreeing with our own or other people's plans. In Gibbard's words:

> Disagreement in plan. . . is the key to explaining normative concepts.

We decide what we ought to do, on Gibbard's account, by choosing between possible plans, thereby deciding what to do. To explain our beliefs about what other people ought to do, Gibbard supposes that we choose between plans that would apply to some merely imaginary case. We decide what we would do if we were going to be in someone else's position, and we would be relevantly like this other person. Suppose you tell me that, if a certain person offered you a job, or proposed marriage, you would accept. I might decide that, if I were in your position and were in other ways like you, I would refuse these offers. On Gibbard's account, our plans would then disagree, and we would thereby disagree about what you ought to do.

It may be objected, Gibbard notes, that when two people make such different decisions, they may not be disagreeing. The truth may be only that these people have adopted different plans. [753] If such a difference between people's plans is not a disagreement, Gibbard could not explain our normative disagreements by appealing to such differences between people's plans.

In responding to this objection, Gibbard first claims that, when we change some plan without some change in our factual beliefs, we thereby disagree with one of our *own* earlier normative beliefs. In Gibbard's words:

> We must count a change of plan as not only a change like a shave or a haircut, but as coming to disagree with one's earlier planning. . . [or] with what one previously thought. [754]

This claim is not, I believe, true. As I have said, we must sometimes choose between plans that seem to us to be equally good. We may adopt one of these plans, and then later change to some other plan, without any change in our factual beliefs or any disagreement with our previous normative beliefs. This might be true in *Burning Hotel*, for example, if you changed your decision about through which window you will jump.

Responding to a similar objection, Gibbard qualifies his account. We disagree with some earlier normative belief, Gibbard suggests, whenever we change some plan because our preferences change. [755] But that is not so. Suppose that when I most enjoyed climbing I planned to buy some hut in the mountains, but now that I prefer sailing I plan to buy some hut near the sea. This change of plan may involve no disagreement with my earlier normative beliefs.

Gibbard also claims that our plans must act as 'judgments' or 'determinations' to which we are committed, and with which we might later disagree. To defend this claim, Gibbard appeals to the fact that, if we don't commit ourselves to our plans, we shall be less likely to achieve our aims. [756] But this fact does not support Gibbard's claim. We often act on some plan because we know that, if we don't, we shall not achieve some aim. In such cases, we don't need to believe that we shall be acting on some plan that we ought to follow. We may know that some other plan would be just as good. To be motivated and moved to jump through one of your two windows, you wouldn't need to believe that this is the window through which you ought to jump. [757]

When Gibbard returns to our beliefs about what other people ought to do, he concedes that different people can have different plans about how to act in some kind of case, without thereby disagreeing. Such people may merely have different plans. But it would be better for everyone, Gibbard claims, if we all regarded such cases as involving disagreements, since that would make it easier for different people to give each other advice. 'In thinking how to live', he writes, 'we need each other's help.' [758]

As before, this claim does not support Gibbard's account. Gibbard is trying to explain normative disagreements by appealing to the simpler idea of disagreements between such plans. On Gibbard's suggested explanation, people who have such different plans thereby disagree. Gibbard concedes that such people may not be disagreeing. We cannot believe that such people are disagreeing merely because, if we had this

belief, that would be better for us. Gibbard's claim could be only that it would be worth *pretending* that such people are disagreeing. But if we merely pretend that such cases involve disagreements, this could not help us to understand what is involved in real normative disagreements. [759]

99 Can Non-Cognitivists Explain Normative Mistakes?

Even if we understand normative disagreements, there is another, more important question. In Gibbard's words:

> Can I ever be mistaken in an *ought* judgment?. . . Do we discover how best to live, or is it a matter of arbitrary choice. . .? [760]

If such judgments cannot be either correct or mistaken, and merely involve arbitrary choices, there would be no point in trying to answer questions about what we ought to do, or how it would be better or worse to live, since we could not reach better conclusions. We might as well act on impulse, consult some astrologer, or toss coins.

Gibbard and Blackburn both believe that, though our normative judgments express desires, decisions, or other conative attitudes, these attitudes and judgments *can* be correct or mistaken. We can therefore claim, they say, that such judgments can be true or false. By making certain further claims, Blackburn suggests, Expressivist Non-Cognitivists can be *Quasi-Realists*, who can justifiably say all, or nearly all, that Cognitivists---whom he calls *Realists*---can say. As Blackburn writes:

> quasi-realism is trying to earn our right to talk of moral truth, while recognizing fully the subjective sources of our judgments inside our own attitudes, needs, desires, and natures. [761]

For Gibbard and Blackburn to defend these claims, they must explain what it would *be* for our conative attitudes and judgments to be true or false, correct or mistaken.

According to Cognitivists, normative judgments express beliefs. When two people's judgments conflict, at least one of these judgments must be false, since contradictory beliefs cannot both be true. Non-Cognitivists, as Gibbard concedes, cannot make such claims. [762] On Gibbard's account, our normative judgments conflict when we make different decisions about how we would act in some situation, thereby adopting different plans. As Gibbard points out, we cannot argue that this difference between our plans involves a contradiction, so that one of these decisions must be false. Gibbard suggests that, if we regard such different plans as being inconsistent, so that one of them must be mistaken, this would be better for us, since we shall then get 'the benefits of normative discussion'. But as before, this fact could only give us reasons to *pretend* that, when people have such different plans, one of these plans must be

mistaken. And this pretence would not help either to show that one of these plans must be mistaken, or to explain what it would be for some plan to be mistaken.

Blackburn appeals to a different kind of inconsistency. When he discusses practical conflicts, Blackburn writes:

> if our attitudes are inconsistent, in that what we recommend as policies or practices cannot all be implemented together, then something is wrong. [763]

But when our attitudes are in this sense inconsistent, something is wrong only in the sense that some of us will be disappointed, since some people's recommended policies will not be carried out. We cannot claim that, when two attitudes are in this sense inconsistent, one of these attitudes must be false or mistaken. Such attitudes, on Blackburn's view, fundamentally involve desires; and when two desires cannot both be fulfilled, that does not imply that one of these desires must be in some way mistaken. We have many rational desires that cannot all be fulfilled. As Blackburn himself writes, 'desires can be faultlessly inconsistent'. [764]

Since Blackburn claims that we should 'recognize no interesting split between values and desires', and he admits that desires can be faultlessly inconsistent, it is hard to see how Blackburn can hope to defend the Quasi-Realist view that, when two people make value judgments that express such inconsistent desires, at least one of these value judgments must be false or mistaken.

Blackburn, however, does ingeniously and resourcefully defend this view. He suggests several ways in which Non-Cognitivists might be able to explain what it would be for people's attitudes and moral judgments to be mistaken. Blackburn first remarks:

> Of course there is no problem in thinking that *other* people may be mistaken. [765]

There *is*, I believe, a problem here. To explain a sense in which other people may be mistaken, it is not enough to say that we may think that these people are mistaken, or that we may disagree with these people. On Blackburn's account, we disagree with other people when we and they have different desires or other conative attitudes that cannot both be fulfilled. We cannot say that, in such cases, 'mistaken' means 'different from mine'. Here is one way to illustrate this point. As Gibbard claims,

> You can't disagree with a headache. [766]

But suppose I reject this claim, since I believe that people's headaches can be true or false, correct or mistaken. If I was trying to explain this strange view, it would not be enough for me to say that other people's headaches are false, or mistaken, when their mental state differs from mine, because they have a headache and I don't.

Blackburn continues:

> The problem comes with thinking. . . that I may be mistaken. How can I make sense of fears of my own fallibility?

To explain such fears, Blackburn claims, he can appeal to the idea that he would cease to have some present attitude if he were in some improved state of mind. That might be true, for example, if he were better informed, or more impartial. Blackburn then writes:

> the quasi-realist can certainly possess the concept of an improved standpoint from which some attitude of his appears inept, and this I suggest is all that is needed to explain his adherence to the acceptance of the apparently realist claim 'I might be wrong'. [767]

This is not, I believe, all that is needed. For Blackburn to appeal to this idea, he must explain in what sense this possible standpoint would be *improved*.

When we are discussing beliefs, we can describe some standpoint as improved in the sense that, if we had this standpoint, our beliefs would be less likely to be mistaken, by being false. Juries, for example, are less likely to convict innocent people if they know more of the facts, and they are not swayed by prejudice. This use of 'improved' makes sense because we already know what it would be for some jury's verdict to be mistaken.

Blackburn, however, is trying to *explain* some sense in which some of his present desires or other conative attitudes might be false, or mistaken. That might be true, he suggests, in the sense that he would not have these attitudes if his standpoint were improved. To explain the sense in which this standpoint would be improved, Blackburn would have to claim that, if he had this standpoint, his attitudes would be less likely to be mistaken. And this claim would have to use the word 'mistaken' in the very sense that Blackburn is trying to explain. So this suggested explanation fails. I might similarly claim that my headache might be mistaken in the sense that I would not have this headache if I was in some improved state of mind in which my headaches would not be false, or mistaken. This claim would not explain what it would be for my headache to be false, or mistaken.

To explain the sense in which his conative attitudes might be mistaken, Blackburn elsewhere writes:

> there are a number of things I admire: for instance, information, sensitivity, maturity, imagination, coherence. I know that other people show defects in these respects, and that these defects lead to bad opinions. . . So I can think that perhaps some of my opinions are due to [such] defects.[768]

In claiming to know that other people have bad opinions, Blackburn again assumes what he needs to explain. In what sense are these opinions *bad*, rather than merely different from Blackburn's opinions?

We have other reasons to believe that Blackburn's appeal to an improved standpoint cannot explain any sense in which our conative attitudes might be mistaken. As Blackburn notes, what he would regard as an improved standpoint depends on his present attitudes. He imagines knowing that, if he were fully informed and impartial, he would lose all of his present attitudes. If he knew this fact, Blackburn remarks, he would claim that this possible standpoint, despite being fully informed and impartial, would *not* be improved. [769] On this version of Blackburn's view, some of our attitudes might be mistaken in the sense that we would not have these attitudes if we had less information, and we were less impartial. It would be harder to defend the claim that this more ignorant and biased standpoint would be, in some relevant sense, improved. And as these remarks imply, when we ask whether our own present attitudes might be in Blackburn's sense mistaken, it is our own present attitudes that provide the answer. These attitudes would be their own judge and jury. [770]

Blackburn might reply that, on any view, we cannot avoid giving priority to our own present point of view. As he writes,

> when I wonder how I might improve, I have to think about it deploying my current attitudes---there is no standing aside and apart from my present sensibility. [771]

But this reply would not succeed. It is true that, even on a Cognitivist view, we must give one kind of priority to our own present beliefs. Though we know that our present beliefs might be mistaken, we cannot base our decisions on the truth rather than on what we *now believe* to be the truth. But, despite this fact, we can explain what it would be for our present beliefs to be mistaken. These beliefs would be mistaken if they were false. As a Non-Cognitivist, Blackburn cannot give this explanation. He cannot claim that our present conative attitudes might be mistaken by being false, since these attitudes are fundamentally desires, and desires cannot be false. And as I have said, he cannot claim that our attitudes might be mistaken in the sense that we would not have these attitudes if we were in some state of mind in which our attitudes would not be mistaken. These objections to Blackburn's Quasi-Realism are, I believe, decisive.

Andrew Egan adds a more particular objection. [772] Of our present moral attitudes, some are *unstable*, in the sense that we would lose these attitudes if we had what Blackburn calls some improved standpoint. These are the attitudes that, on Blackburn's view, we can regard as possibly mistaken. Our other present attitudes are *stable*, in the sense that we would keep these attitudes in any such improved state of mind. These unchangeable attitudes are deeper, or more fundamental. On

Blackburn's view, we can understand what it would be for *other people's* stable attitudes to be mistaken. These other people might disagree with us, and they would then be making fundamental moral errors. But on this view, as Egan argues, we cannot intelligibly think that our *own* stable attitudes might be mistaken. So each of us can justifiably believe that we are the only person who has an *a priori* guarantee against fundamental moral error. This conclusion, Egan writes, would be at best 'very, very strange', and at worst 'incoherent'. It would, I believe, be incoherent. We could not each be entitled to be certain that we are the only person who could not make fundamental errors.

Blackburn might retreat to the view that *everyone* has a guarantee against fundamental moral error, since no one's stable moral attitudes could be mistaken. But this revision would abandon Quasi-Realism, since Blackburn would then be admitting that different people could have conflicting attitudes, and make conflicting moral judgments, none of which would be false or mistaken.

On Blackburn's view, as he might instead reply, each of us could still claim to know that our own judgments were true. We can talk of 'knowledge', Blackburn writes, if 'we rule out any possibility that an improvement might occur'. [773] But we cannot turn our beliefs into *knowledge* merely by excluding the possibility that we are mistaken. People with contradictory beliefs might all exclude the possibility that they are mistaken.

Blackburn gives another defence of his Expressivist Quasi-Realism. When we ask what may seem to be external, *meta-ethical* questions, Blackburn claims, these may really be internal *moral* questions.

This *internalist* response can be plausibly applied to some questions. As Blackburn says, we can use 'true' in a minimal sense, which is merely a way of repeating some claim. If you said that honey meringues were even more disgusting than milk chocolate, I would say 'That's true'. Suppose that someone asks Blackburn whether it is really true that, for example, cruelty is wrong. On Blackburn's Expressivist view, he could answer 'Yes', since this answer would express his disapproving attitude towards such acts. [774] Someone might next object that, on Blackburn's view, cruelty isn't really in itself wrong, since what makes cruelty wrong is only our attitude towards such acts. Blackburn could reply that, on his view, what makes cruelty wrong is not his disapproval of such acts, but the suffering that these acts cause. This reply would reflect the fact that Blackburn's attitude to cruelty is a response, not to his own attitude, but to this suffering. [775]

As Blackburn admits, however, there are some meta-ethical questions that cannot be regarded as internal moral questions. We are now discussing one such question. We

are not asking whether, on Blackburn's view, it is really true that cruelty is wrong. We are asking what it would *be*, on such Non-Cognitivist theories, for some moral judgment to be true or false, correct or mistaken. Since we are not asking whether some *particular* moral judgment is true, our question is morally neutral, and cannot be given an internal moral answer. And we may be right to conclude that Non-Cognitivists cannot answer this question, since there is no intelligible sense in which, on Non-Cognitivist theories, moral judgments might be true or false, correct or mistaken.

Blackburn tries to avoid this conclusion. Making an internalist move, he writes:

> To think that there are no moral truths is to think that nothing should be morally endorsed, that is, to endorse the endorsement of nothing, and this attitude of indifference is one that it would be wrong to recommend and silly to practise. [776]

But this claim is unjustified. When other Non-Cognitivists say that there are no moral truths, they are not making the moral claim that we ought not to make any moral claims. They are making the quite different meta-ethical claim that, even if moral claims can be said to be true in some minimal sense, such claims cannot be true or false in the strong sense to which Moral Cognitivists or Realists appeal. This, moreover, is Blackburn's own view. Blackburn writes:

> There is no problem of relativism because there is no problem of moral truth. . . moral opinion is not in the business of representing the world. . [777]

> . . . if realism were true. . . there would be a fact, a state of affairs (the wrongness of cruelty). . . But anti-realism acknowledges no such states of affairs. [778]

These, we can add, are not internal moral claims. When Non-Cognitivists claim that there is no property of wrongness that cruelty might have, and no such state of affairs or fact, they do not thereby reject the abhorrence of cruelty that humane Humeans like Blackburn eloquently express.

Blackburn elsewhere writes that, if some Non-Cognitivist adopts the Expressivist strategy, this person can tell us what is involved when someone believes that something is good. But if we

> go on to ask this strategist what it is for something to *be* good, the response is that this is not the subject of this theoretical concern---that is, not the subject of concern for those of us who, while naturalists, want a theory of ethics. Either the question illegitimately insists that trying to analyse the ethical proposition is the only possible strategy, which is not true. Or it must be heard in an ethical tone of voice. To answer it would then be to go inside the domain of ethics, and start expressing our standards. [779]

Blackburn here suggests that we cannot legitimately ask Expressivist Non-Cognitivists

what it would *be*, on their theories, for something to be good. But we *can* legitimitately ask *Cognitivists* this question, and these people can give us answers. Cognitivists might tell us, for example, what it would be for something to be good in the reason-involving sense. If we ought not to ask these Expressivists what it would be, on their theories, for something to be good, this would have to be because we already know that, according to these Expressivists, nothing *could* be good, so that it would be pointless, or discourteous, to ask these people to explain how, on their view, something might be good. [780]

Blackburn also claims that, as a Quasi-Realist Expressivist, he doesn't need to explain what it would be, on his view, for conative attitudes or judgments to be false or mistaken. He writes:

> If some theorist. . . asks me what my account of moral error *itself* is, then I am not very forthcoming. . . It is much more in the spirit of quasi-realism. . . to avoid such formulations. This is not an ad hoc move, but an integral part of the package. . . the quasi-realist. . . avoids saying *what it is* for a moral claim to be true, except in boring homophonic or deflationary terms. The only answer we should recognize to the question 'what is it for happiness to be good?' is happiness being good. [781]

But, as Blackburn earlier wrote,

> quasi-realism is trying to earn our right to talk of moral truth, while recognizing fully the subjective sources of our judgments. . .

As Blackburn rightly claimed, Quasi-Realists need to *earn* this right. On Blackburn's view, though our moral judgments fundamentally express certain kinds of desire or other conative attitude, such judgments can be true or false. That is a bold and surprising claim, which needs to be both explained and defended. When Blackburn applies his Quasi-Realism to some other areas of our thinking, such as our beliefs about probabilities and counterfactuals, he persuasively defends our right to call some of these beliefs true.

In the longer passage just quoted, however, Blackburn merely asserts that Quasi-Realists have this right. When we ask what it would be, on Blackburn's view, for us to judge truly that happiness is good, Blackburn thinks it enough to reply 'This judgment would be true if happiness is good'. We judge truly that some act is wrong, Blackburn would similarly say, if this act *is* wrong. Such claims cannot give Expressivists the right to talk of moral truth. We judge truly that some headache is mistaken, I might similarly claim, if this headache *is* mistaken. For Quasi-Realist Expressivists to *earn* their right to talk of moral truth, they must explain what it would be, on their view, for it to be true that some act is wrong. That is why I could not hope to defend Quasi-Realism about headache judgments. I could not earn a right to

call these judgments true, because I could not explain what it would *be* for it to be true that some headache is mistaken.

Return now to Blackburn's claim that, by appealing to the idea of an improved standpoint, Expressivists can explain a sense in which, like any Realist or Cognitivist, they can think 'I might be wrong'. In this way, Blackburn writes, Expressivists can both hold fast to emotivism and perfectly imitate, or 'mimic', this 'alleged realist thought'. [782] I have questioned these claims. But even if these claims were justified, they would not answer our questions. We are asking whether there are truths about decisive reasons, and about what we ought to do. These are not questions about what we can perfectly mimic, or pretend to think.

Quasi-Realism could not, I believe, succeed. Suppose that, as Moral Sentimentalists believe, morality essentially involves certain desires or other conative attitudes towards our own and other people's acts. There are then two possibilities. If these attitudes can be correct or mistaken, we ought, I believe, to be Realists or Cognitivists. On one such Realist view, our moral judgments are true when we claim that some correct attitude is correct, and our judgments are false when we claim that some correct attitude is mistaken. We can reject such forms of Realism only if these attitudes cannot be correct or mistaken. Only then should we believe that our moral claims cannot be true or false, and *merely express* such conative attitudes. Quasi-Realist Expressivists therefore face a dilemma. To defend their Non-Cognitivist Expressivism, these people 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 28 NORMATIVITY AND TRUTH

100 Expressivism

Gibbard and Blackburn might object that, in criticizing their views, I have failed to take seriously their *Expressivism*. When I ask what it would *be* for normative judgments to be correct or mistaken, I assume that we need to know what would make such judgments true or false. This 'truth conditions approach', Blackburn objects, is not 'the only possible strategy'. Expressivists explain such judgments in a very different way.

Gibbard gives an Expressivist account of rationality. His aim, he writes, is to explain 'what "rational" means'. But Gibbard never directly answers this question. There is, he claims, no such property as that of being rational. Since that is so, we cannot explain the word 'rational' by describing what it is for something to be rational. The best we can do is to describe

> what it is for someone to *judge* that something is rational. We explain the term. . . "rational", by saying what state of mind it expresses. [783]

Before considering Gibbard's account, we can start with some remarks about Expressivism. Some Non-Cognitivists claim that, in saying

> (A) X is good,

we *express our approval* of X. This claim may not help, since we may use the word 'approve' to mean 'believe to be good'. Someone might similarly claim that, in saying

> (B) The Earth is round,

we express our acceptance of the roundness of the Earth. That would not help to explain what (B) means. Such claims are unhelpful in two ways. We cannot explain some belief by appealing to the attitude of accepting this belief. And such accounts fail when they use the concept that we are trying to explain.

Consider next the utterance

> (C) Good-bye!

Here, in contrast, Expressivism helps. (C) once meant 'God be with you!' Since that utterance was not the statement of a belief, it needs to be explained as the expression of a wish, or prayer. And to explain what 'Good-bye!' means today, we can say that this phrase conventionally expresses, to those from whom we are about to be parted, an

attitude of goodwill.

To explain what 'rational' means, Gibbard claims that, in saying

(D) It is irrational to be angry with bringers of bad news,

we express our acceptance of a *norm* against such anger. Whether this account is helpful depends on what this norm is claimed to be. If this norm were

(E) There is no reason to be angry with such people,

this account would have both the flaws just mentioned. In expressing our acceptance of (E), we would be merely expressing our belief in (E). And since (E) uses the concept of a normative reason, an appeal to (E) could not explain what 'rational' means in non-normative terms.

Gibbard's account avoids both these flaws. Gibbard claims that, in saying (D), we express our acceptance of a norm like

(F) Do not be angry with bringers of bad news!

Like 'Good-bye!', (F) does not state some belief. And since (F) does not use any normative concept, Gibbard's claim might explain (D) in non-normative terms.

Gibbard uses the word 'norm' to 'mean simply a prescription or imperative'. [784] Imperatives are commands, like 'Do not be angry with such people!', 'Keep your promises!' or 'Never lie!' Such sentences cannot be either true or false. We *accept* some imperative, not by believing something, but by deciding to do what this imperative tells us to do. Imperatives are not in my sense normative, since they do not state or imply that we have some reason, or that we ought to act in a certain way. (It is no objection to this claim that, when some legitimate authority commands us to act in some way, we might be right to conclude that we ought to obey this command.) Since Gibbard's norms are, as he claims, merely imperatives, that is what I shall call them.

There may seem to be another way in which Gibbard's account is unhelpful. Gibbard claims that, when we try to decide whether some act is rational, we are trying to decide *whether to accept* some imperative. This claim may suggest that we are trying to decide whether we have sufficient or decisive reasons to accept this imperative, or whether we ought rationally to accept it. But this account would then be using the very concepts---*reason*, *ought*, and *rational*---which it claims to explain.

As before, Gibbard avoids this objection. On Gibbard's account, we do not try to decide which imperatives we *ought* to accept, or have *reasons* to accept. We merely decide which imperatives *to accept*. As Gibbard later claims, deciding what we *ought*

to do is choosing what to do. [785]

Gibbard makes some other suggestions, of a socio-biological kind, about what is involved when organisms like human beings accept such imperatives. An imperative, Gibbard writes,

> is a formulation of a pattern which, in effect, controls the organism's behavior. . . . If a norm is simply an imperative, the real psychological question is what it is to internalize it. A norm prescribes a pattern of behavior, and to internalize a norm. . . is to have a motivational tendency of a particular kind to act on that pattern.

We are not the only animals, Gibbard remarks, who are subject to 'normative governance'. The capacity to 'internalize norms' is 'one we share with other mammals', such as wild dogs. But though other animals *internalize* norms, only we, because we have language, can also *accept* norms. Gibbard writes:

> The capacity to accept norms I portray as a human biological adaptation; accepting norms figures in a peculiarly human system of motivation and control that depends on language.

To 'accept a norm', he continues, 'is to be prepared to avow it in normative discussion.' Or more exactly, 'accepting a norm is whatever psychic state, if any, gives rise to this syndrome of avowal of the norm and governance by it.' [786]

As these quotations show, Gibbard's account avoids circularity. If a norm is 'simply an imperative', if other animals can 'internalize' such imperatives, and if what we add to their 'system of motivation' is only the 'avowal' of these imperatives, Gibbard's account does not use materials which contain the very feature---normativity---that he is trying to explain.

Return now to Gibbard's main aim: to explain 'what 'rational' means'. If we can explain this idea, Gibbard writes, this would help us to decide 'how it is rational to conduct our lives. What are we asking? It seems the widest question in life: how to live.' [787]

When Gibbard rejects Naturalist accounts of words like 'rational', he rightly claims that these accounts make it impossible to ask such questions. Does Gibbard's account do better?

I believe not. If we apply Gibbard's account to these questions, we soon face a blank wall. Gibbard writes, for example:

> What is it, then, for an act or a way of feeling to be rational? In what way does a

> person who calls something rational endorse it? [788]

Our disappointment here is swift. Though Gibbard starts by asking *what it is* for an act or feeling to be rational, he turns at once to a different question. On Gibbard's view, since there isn't any property of being rational, there can't be anything *that it is* for an act or feeling to be rational. There are only endorsements of imperatives, such as, 'Act like that!' In asking how it is rational to live, we are choosing between such imperatives. Nor could we ask which imperatives it would be rational for us to choose, since no choice could be rational.

Gibbard would reply that, in making these claims, I am begging the question. I am assuming that, in believing that some act or choice is rational, we are believing it to be true that this act or choice has the property of being rational. On Gibbard's view, that is not so. To believe some act to be rational isn't really to have a belief, but to accept the imperative 'Act like that!' Gibbard would say that, if his account is correct, and we accept this imperative, we *can* claim that such acts are rational. And he writes, that, on his view, we can believe that various acts

> really are rational or irrational, right or wrong. [789]

This reply, I believe, fails. Like many great philosophers, Gibbard tries to have things both ways. On Gibbard's view, acts cannot really be rational. As he writes, 'to call a thing rational is not to state a matter of fact, either truly or falsely'. But Gibbard also claims that, even if we accept his view, we can go on believing that certain acts truly are rational. We can sometimes have things both ways. If I you said 'Milk chocolate is disgusting', I could both reply 'That's true' and deny that, on my view, milk chocolate truly has the property of being disgusting. But that is because, in saying 'That's true', I would be merely expressing the same dislike. When we believe that some act truly is rational, or that we really do have decisive reasons to act in some way, are we using *truly* or *really* in this minimal, expressivist sense?

I believe not. Like Naturalist accounts, Gibbard's account makes it impossible to ask certain important questions. If we interpret our questions in the way that Gibbard suggests, they cease to be the questions that we wanted to ask, or thought we were asking. For example, we can't really ask what it would be rational for us to do.

As before, Gibbard would reject this claim, since he often writes of what is 'rational in the expressivistic sense'. But this phrase is misleading. There is no expressivistic sense in which acts could *be* rational. Acts can merely have the property of conforming to the imperatives that I accept, or the imperatives that you accept, or the imperatives that other people accept. If some act conforms to one of these imperatives, that is not a way of *being expressivistically rational*. It would be empty for me to claim that an act is rational in the expressivistic sense if this act conforms to *my* imperatives. You could say the same about acts that conform to *your* imperatives.

The truth would be only that some acts conform to my imperatives, while others conform to yours.

Gibbard's account, he concedes, seems to leave something out. When a person calls something rational, Gibbard writes,

> he seems to be doing more than simply expressing his own acceptance of a system of norms. . . he claims to recognize and report something that is true independently of what he himself happens to accept or reject. Perhaps he is wrong. But that is the claim he is making. . . . If the person claims objective backing and the analysis misses the claim, then the analysis is defective. [790]

Some 'claims to objectivity', Gibbard then replies, 'are well explained by norm-expressivism'. When we accept some norm, we need not regard this norm as depending on our acceptance of it. In his words:

> If a person thinks something a matter of taste, then he does not think, 'This taste would be valid even if I lacked it'. In matters of rationality, in contrast, we do think, 'This norm would be valid even if I did not accept it'.

Expressivists, Gibbard says, can make such claims. If I say, for example, that slavery is wrong, my attitude is a response to certain features of slavery. Since my attitude is a response to these features, I would naturally extend my attitude to an imagined case in which, though I didn't have this attitude, slavery still had these features. I could say 'Don't enslave people, even if I cease to accept this imperative!'

It is true that, as Gibbard here claims, some of our attitudes are not conditional on our continuing to have these attitudes. If we want some enemy to suffer, for example, our desire may not be conditional on its own persistence. We may want our enemy to suffer whether or not we continue to have this desire. But as this example shows, this kind of non-conditionality doesn't amount, as Gibbard claims, to a kind of *objectivity*.

Gibbard then says that, when he expresses some norm, 'I demand acceptance of what I am saying'. 'This demand', he writes,

> is part of what has been missing in the analysis. Before, I said roughly that when a person calls something 'rational' he is expressing his acceptance of norms that permit it. . . Now I say he is doing more: he is making a conversational demand. He is demanding that the audience accept what he says, that it share the state of mind that he expresses. [791]

When we make such demands, as Gibbard notes, we are not merely issuing orders. We are making claims that we believe to have 'normative authority'. He then writes:

> To claim authority is to demand influence. . . . I say, implicitly 'Accept these

> norms!' and if you accept them because I have made the demand, I have influenced you. [792]

Most of us do not, I believe, claim *authority* for ourselves. We would at most claim authority for the principles to which we are appealing. And if we did claim authority, we would not be *demanding influence*. That would be to confuse authority with power. Suppose I claim that you ought not to accept two contradictory beliefs. We would misdescribe this use of 'ought' if we said that I am *demanding* that you accept my claim.

As before, Gibbard notes this point. He writes, 'I as a speaker do not simply demand; I claim to have a basis for my demands.' When I disagree with someone, I claim 'to be "seeing" something that she doesn't: that the fundamental norms she accepts just don't make sense.' [793] On Gibbard's account, however, there is nothing to see, since there are no truths about what 'makes sense'. And if we decide not to accept some imperative, that is not seeing that something does not make sense.

Gibbard similarly talks of our finding norms 'credible'. And he writes, 'The fact that I would enjoy something speaks in favor of doing it. I find that self-evident.' [794] But on Gibbard's view, norms are imperatives, and when we believe that some fact 'speaks in favor' of some act, we are merely accepting some imperative. Unlike beliefs or normative claims, imperatives cannot be either *credible* or *self-evident*.

Gibbard might reply that, as he and Blackburn claim,

> normative judgments mimic factual judgments. . . [or] the search for truth.

Though the relevant norm is, really, 'Act like that!', we express it in a form that mimics some factual belief, by saying 'Such acts are rational'. Our attitude to this imperative could then similarly *mimic* finding some belief to be credible, self-evident, or obviously true. Such mimicry may seem enough.

When Gibbard sums up his aims, he writes:

> Above all, I hope, the analysis will help us understand why it matters which acts and feelings are rational.

But as before, if Gibbard's view were true, there would be nothing to understand. Since there is no expressivistic sense in which anything could be rational, there would be no point in asking which acts and feelings are rational. Nor could anything matter. Just as our normative beliefs could only mimic the search for truth, things could only mimic mattering. Since a mimic is a fake, or sham, such mimicry is not enough.

Gibbard's analysis, he also claims,

> can transform our view of what we are doing when we ponder fundamental

> normative questions, and allow us to proceed more effectively in our normative thinking.' [795]

Gibbard's analysis would indeed transform our view. If we became convinced that there are no truths about what is rational, or about reasons, or about what we ought to do, we would cease to believe that normative questions could have answers. Our normative thinking would then be easier, since we would cease to worry that we might be getting things wrong. But that would not make our thinking *more effective*, since it would not help us to get things right. There would be nothing to get right.

After claiming that there are no truths about what is rational, or about reasons, Gibbard writes that this claim does not leave 'normative language defective, or second rate'. [796] That depends on whether, as Gibbard admits that our 'ordinary thought' assumes, there are truths about what is rational, and about reasons. If there are no such truths, our normative thinking *would* be defective, since we would be wrong to assume that our beliefs about rationality and reasons might be true. Accepting Gibbard's view would free us from that illusion. If instead there *are* such truths, accepting Gibbard's view would blind us to them.

Gibbard also hopes that, when we are trying to decide 'what really matters and why', his account of normativity can make some 'fruitful' answers 'seem evident and right'. If Gibbard's view were true, no answer could *be* right, And if we really accepted and understood this view, none could even *seem* to be evident or right. Phrases like 'what really matters' would be seen merely to mimic the search for truth.

As Gibbard writes, his main question is:

> Can I ever be mistaken in an *ought* judgment?. . . Do we discover how best to live, or is it a matter of arbitrary choice. . ? [797]

On Gibbard's view, I have argued, there would be nothing to discover. We could never be mistaken in our judgments about how it would be better or worse to live, since this *would* just be a matter of arbitrary choice.

Unlike many Non-Cognitivists, Gibbard realizes that his view cannot be restricted to practical reasons: reasons for caring and for acting. In his words, 'Norms are fundamental to thought. . . we cannot think at all without some implicit guidance by norms'. Just as 'what it is rational to do settles what to do. . . what it is rational to believe settles what to believe'. Remember finally that, on Gibbard's view, 'to call a thing rational is not to state a matter of fact, either truly or falsely'. If there could not be facts or truths about what it is rational to believe, as Gibbard's view implies, it could not be rational to believe anything, including Gibbard's view.

101 Hare on What Matters

A young Swiss guest of Richard Hare's, after reading a novel by Camus, concluded in despair that *nothing matters*. Hare suggested that his friend should ask 'what was the meaning or function of the word 'matters' in our language; what is it to be important?' His friend soon agreed, Hare writes,

> that when we say something matters or is important, what we are doing, in saying this, is to express our concern about that something. . . Having secured my friend's agreement on this point, I then pointed out to him something that followed immediately from it. This is that when somebody says that something matters or does not matter, we want to know *whose* concern is being expressed or otherwise referred to. If the function of the expression 'matters' is to express concern, and if concern is always *somebody's* concern, we can always ask, when it is said that something matters or does not matter, 'Whose concern?' [798]

As Hare pointed out, his friend *was* concerned about several things. So was everyone---except a few fictional characters in existentialist novels. People's values differ, and may change. But, since we all care about something, 'it is impossible to overthrow values as a whole.' Hare's treatment worked. 'My Swiss friend ate a hearty breakfast the next morning.'

If someone doubts whether anything matters, it may not help to ask 'Whose concern?' Hare managed to convince his friend

> that the expression 'Nothing matters' in his mouth could only be (if he understood it) a piece of play-acting. Of course he didn't actually understand it.

The word 'matters' has a meaning, I believe, which Hare did not understand. Things can matter in the sense that we can have reasons to care about these things.

When Hare writes that we use such words to *express* concern, he is not, he claims, using 'express' in an 'emotivist' sense. But Hare does here accept an Emotivist, Expressivist, or more broadly, Non-Cognitivist view. That is why, when Hare's friend concluded in despair that nothing mattered, Hare didn't remind his friend that some things, such as suffering, really do matter. As Hare writes:

> My friend. . . had thought mattering was something (some activity or process) that things did. . . If one thinks that, one may begin to wonder what this activity is, called mattering; and one may begin to observe the world closely. . . to see if one can catch anything doing something that could be called 'mattering'; and when we can observe nothing going on which seems to correspond to this name, it is easy for the novelist to persuade us that after all *nothing matters*. To which the answer is, '"Matters" isn't that sort of word; it isn't intended to *describe* something. . .'

On Hare's view, it makes no sense to describe something as mattering. The truth is only that we care about some things. In saying that these things matter, we are not *claiming* that they matter, but are merely expressing our concern.

Hare assumes that, in making these claims, he is not denying anything that other people might believe. There is nothing to deny, he claims, since no other view makes sense. Hare imagines an objector saying:

> All you have done is to show that people are *in fact* concerned about things. But this established only the existence of values in a *subjective* sense.

This objector, Hare supposes, claims that there are *objective* values. Hare then writes:

> I do not understand what is *meant* by the 'objectivity of values', and have not met anybody who does. . . suppose we ask 'What is the difference between values being objective, and values not being objective?' Can anybody point to any difference? In order to see clearly that there is *no* difference, it is only necessary to consider statements of their position by subjectivists and objectivists, and observe that they are saying the same thing in different words. . . An objectivist . . says, 'When I say that a certain act is wrong, I am stating the *fact* that the act has a certain non-empirical *quality* called 'wrongness'. . . A subjectivist says, 'When *I* say that a certain act is wrong I am expressing towards it an attitude of disapproval which I have.' [799]

When Hare claims that there is no disagreement here, he assumes that objectivists cannot mean what they say. There *is* a disagreement here. As Hare writes, some objectivists believe that there are facts about which acts have the non-empirical 'quality' or property of being wrong. Hare's 'subjectivists'---by whom he means Expressivists---believe that no act could have such a property.

Hare continues:

> We all know how to recognize the activity which I have been calling 'saying, thinking it to be so, that some act is wrong'. And it is obvious that it is to this activity that the subjectivist and the objectivist are both alluding. This activity. . . is called by the objectivist 'a moral intuition'. By the subjectivist it is called 'an attitude of disapproval'. But in so far as we can identify anything in our *experience* to which these two people could be alluding by these expressions, it is the same thing---namely the experience which we all have when we think that something is wrong.

When objectivists claim that certain acts really are wrong, they are not referring or alluding to the experiences that we have when we believe some act to be wrong. Their claim is about *what* we believe. More exactly, it is about what some of us believe. They might concede that some people---such as some Expressivists or sceptics---do not

have such beliefs.

Hare might reply that *he* has such beliefs. He is discussing the activity of 'saying, *thinking it to be so*, that some act is wrong.' Like Gibbard, Hare claims that such beliefs are not like ordinary, *descriptive* beliefs. In thinking something to be wrong, we are not believing something to be true, but accepting the universal imperative 'No one ever act like that!' If Hare gave this reply, however, he would be conceding that there is a disagreement here. According to objectivists, these beliefs *are* descriptive, since they are about normative truths.

Hare then considers another way in which some objectivists explain their view. These people claim that, when two moral judgments conflict, at least one of these judgments must be mistaken, since such conflicting judgments could not both be true. Subjectivists, these people argue, cannot make this claim. Hare replies that, though this claim can explain objectivity in some other areas, it does not, when applied to morality, draw any 'real distinction'. In his words:

> Behind this argument lies, I think, the idea that if it is possible to say that it is *right* or *wrong* to say a certain thing, an affinity of some important kind is established between that sort of thing, and other things of which we can also say this. So, for example, if we can say of the answer to a mathematical problem that it is right, and can say *the same thing* of a moral judgment, this is held to show that a moral judgment is in some way *like* the answer to a mathematical problem, and therefore cannot be 'subjective' (whatever that means).

That is what it means. Like answers to mathematical problems, moral judgments can be objective in the sense that they can be right or wrong, by being true or false.

102 Normative Questions

Hare might give a different reply. He might concede that, when objectivists claim that some moral judgment is wrong, or false, they mean something different from what subjectivists mean. Hare believes that, if objectivism is put forward as a *moral* view, it is self-defeating. As he writes elsewhere:

> moral judgments cannot be merely statements of fact, and. . . if they were, they would not do the jobs that they do do, or have the logical characteristics that they do have. In other words, moral philosophers cannot have it both ways; either they must recognize the irreducibly prescriptive element in moral judgments, or else they must allow that moral judgments, as interpreted by them, do not guide actions in the way that, as ordinarily understood, they obviously do. [800]

As this passage shows, Hare ignores the possibility that there might be normative truths.

Hare assumes that, if moral judgments were capable of being true, or stating facts, they could not guide actions. But if we judged that we ought to do something, that judgment could guide our acts. So Hare assumes that judgments like 'I ought to do that' could not conceivably be true.

Many other writers make such claims. There is a reason, Blackburn writes, why Expressivist Non-Cognitivism 'has to be correct'. If our normative judgments were beliefs, such as beliefs about what we have reasons to do or what we ought to do, these beliefs could not answer practical questions. For any such normative fact, 'there is a question of what to do about it'. [801] To provide answers to practical questions, normative judgments cannot be beliefs about some normative fact, but must be some kind of desire or other conative attitude.

Gibbard also claims that, when applied to the judgments with which we make decisions, 'expressivism has to be right'. [802] According to *Non-Naturalists* like Sidgwick, asking what we *ought* to do is not the same as asking *what to do*. Gibbard claims that, if these were different questions, asking what we ought to do could not help us to decide what to do. Non-Naturalists, Gibbard writes:

> just change the subject. We ask what to do, and they hand us analyses of a different question. [803]

Like Blackburn, Gibbard here claims that normative facts could not answer practical questions.

Gibbard's claim is surprising. Suppose that, in *Burning Hotel*, you decide that you ought to jump into the canal, because that is your only way to save your life. On Gibbard's view, if it was merely a normative fact that you ought to jump, and your belief that you ought to jump was not a decision to jump, your belief could not help you to decide whether to jump. That is clearly false.

Gibbard makes another, more cautious claim. He supposes that, as Non-Naturalists believe, possible acts can have the non-natural property of being what we ought to do, and that when some act has this property that would 'settle' the question of what to do. Even on these assumptions, Gibbard writes, we would never need to ask what we ought to do. It would always be enough to consider the natural facts about our different possible acts, and then decide to act in one of these ways. [804] Nothing would be gained by our having true beliefs about what we ought to do.

Patrick Nowell-Smith similarly writes: 'Moral philosophy is a practical science; its aim is to answer questions of the form 'What shall I do?' But he then warns that 'no general answer can be given to this type of question'. [805] That is an understatement. As Nowell-Smith notes, the word 'shall' is ambiguous. In asking 'What shall I feel?', for example, we are trying to make some prediction, which other people might correctly

give. But in asking 'What shall I do?', we are not trying to predict our acts. We are trying to make a decision. If moral philosophy had the aim of answering such questions, it could not possibly succeed. Moral philosophy cannot make our decisions.

Nor can other people. When we ask 'What shall I do?', that is not a question to which even the wisest adviser could give an answer. If I say, 'That's what I shall do', others might say, 'No you shan't', or 'No you won't.' But these claims would not make my decision. They would be either a prediction, or the expression of a contrary decision---as when a parent says to a child 'You *will do* what I tell you to.'

As these remarks imply, the question 'What shall I do?' is not normative. Nor can this question be, as Nowell-Smith claims, 'the fundamental question of ethics'. The fundamental question is: 'What *should* I do?', or 'What *ought* I to do?' Moral philosophy, or other people, might help us to answer *this* question. There might be truths about what we should or ought do.

Nowell-Smith considers this objection, and replies:

> My reason for treating the 'shall' question as fundamental is that moral discourse is practical. The language of 'ought' is intelligible only in the context of practical questions, and we have not answered a practical question until we have reached a decision.

Though moral discourse is practical, that does not imply that its fundamental question is about what we *shall* do, rather than what we *should* or *ought* to do. We may have already decided that we shall do, or shall try to do, whatever we conclude that we should or ought to do. In answering moral questions, we would then be answering Nowell-Smith's question, by deciding what to do.

Like the other people I have quoted, Nowell-Smith might now reply that, when we are trying to decide what to do, it would not help to form beliefs about what we ought to do, since no such true *belief* could answer our question.

Such claims provide the third main argument for Non-Cognitivism. It will help to compare this argument with two other claims. According to the Naturalist Argument for Non-Cognitivism:

> (A) Since all facts are natural, there could not be any normative truths.

Some Non-Cognitivists add

> (B) Even if there were such truths, they would not really be normative. Truths cannot answer normative questions.

Remember next that, in arguing against Naturalism, I claimed:

(C) Natural facts could not be normative.

Non-Cognitivists, as I have said, accept (C). Some of them add

(D) Even irreducibly normative facts would not really be normative.

(B) and (D) provide what we can call *the Normativity Argument for Non-Cognitivism*.

This argument is often stated in surprisingly self-undermining ways. When discussing Moore's alleged normative truths, for example, Nowell-Smith writes:

> No doubt it is all very interesting. If I happen to have a thirst for knowledge, I shall read on. . . Learning about 'values' or 'duties' might well be as exciting as learning about spiral nebulae or waterspouts. But what if I am not interested? Why should I *do* anything about these newly-revealed objects? Some things, I have now learnt, are right and others wrong; but why should I do what is right, and eschew what is wrong? [806]

When words are 'used in the ordinary way', Nowell-Smith goes on to say, such questions are absurd. But they 'would not be absurd if moral words were used in the way that intuitionists suppose'. In 'ordinary life there is no gap between "this is the right thing for me to do" and "I ought to do this"'. Nowell-Smith then objects that, if 'X is right' were taken to mean that X 'had the property' of being right, we *could* sensibly deny that we ought to do what is right.

There is an obvious reply. As well as asking which act would be right, we can ask what we ought to do. And when we claim that we ought to do something, we may mean that this act has the property of being what we ought to do. According to Nowell-Smith's objection, if this is what we mean, we could sensibly deny that we ought to do what we ought to do. That is not so.

Williams similarly writes that, if the claim that we ought to do something

> just tells one a fact about the Universe, one needs some further explanation of why [we] should take any notice of that particular fact. [807]

Suppose that we knew another such fact, since we also knew why we should take notice of this fact about what we ought to do. On Williams's objection, we could still sensibly ask why we should take notice of this fact. That is not so.

Hare similarly writes that, if it is merely a fact that some possible act has 'the moral property of wrongness', why should we be troubled by that?' [808] But suppose we knew why we should be troubled by this act's wrongness. On Hare's objection, this would merely be another fact. Though we knew why we should be troubled, we could still sensibly ask why we should be troubled. That is not so.

Korsgaard similarly writes:

> If it is just a *fact* that a certain action would be good, a fact that you might or might not apply to deliberation, then it seems to be an open question whether you *should* apply it. [809]

But suppose that you *should* apply this fact to your deliberation. On Korsgaard's objection, since this would just be another fact, it would still be an open question whether you should apply this fact to your deliberation. That is not so. If you should do something, it is not an open question whether you should do it.

According to the writers that I have been discussing, normativity has nothing to do with truth. We can next consider some of Korsgaard's arguments for this view.

There are, I have claimed, some irreducibly normative truths. Korsgaard calls this view *normative realism*. [810] Realists, Korsgaard argues, cannot help us to decide 'what, if anything, we really ought to do', nor can they *justify* the claim that morality makes on us. Suppose, she writes:

> you are being asked to face death rather than do a certain action. You ask the normative question: you want to know whether this terrible claim on you is justified. Is it really true that this is what you *must* do? The realist's answer to this question is simply 'Yes'. That is, *all* he can say is that it is *true* that this is what you ought to do. [811]

Practical reasoning, Korsgaard also claims, is not about what we should *believe*, but about what we should *do*. Realists misunderstand this difference. These people mistakenly assume that, when we ask 'practical normative questions. . . there is something. . . that we are trying to find out. ' [812] On their view, 'our relation to reasons is one of seeing that they are there or knowing truths about them.' [813] Realism fails, Korsgaard claims, because no knowledge of truths about reasons could answer normative questions.

Korsgaard's objections to normative realism seem to be these:

> Realists discuss the wrong question.
>
> Realists may not be able to convince us that some answer to our question is really true.
>
> Even if our question had some true answer, that would not solve our problem. Ours is not a question to which some truth could be the answer.

These objections do not, I believe, succeed. If Korsgaard's question could not be

answered by some truth, this question could not be normative. When there are answers to normative questions, these answers must be normative truths. And if we cannot convince some people that there are such truths, that is no objection to realism.

Return to Korsgaard's imagined doubter who, in some crisis, asks

> Q1: Is it really true that this is what I must do?

Korsgaard discusses several ways of understanding this question, of which I shall here discuss only one. [814] Korsgaard's doubter might be asking:

> Q2: Do I have decisive reasons to act in this way?

Realists might answer 'Yes'. And they might convince this person that their claim is true, since this person really does have decisive reasons to act in this way. But Korsgaard's doubter might then ask

> Q3: Why should I do what I have decisive reasons to do?

To this question, Korsgaard claims, realists would have no answer. Decisive reasons, if understood in a realist way, would not have normative force. Realists 'cannot provide a coherent account of rationality'. According to these people, Korsgaard writes:

> rationality is a matter of conforming the will to standards of reason that exist independently of the will, as a set of truths about what there is reason to do. . . The difficulty with this account. . . exists right on its surface, for the account invites the question why it is rational to conform to those reasons, and seems to leave us in need of a reason to be rational. [815]

Like the other writers quoted above, Korsgaard presents this objection in a surprisingly self-undermining way. According to what Korsgaard calls normative realism, when we know the relevant facts, we are rational if we want, and do, what we have decisive reasons to want, and do. So Korsgaard seems here to suggest that, if realism were true, we might need a reason to want, and do, what we knew that we had decisive reasons to want, and do. That is clearly false.

This may not, however, be what Korsgaard means. She continues:

> To put the point less tendentiously, we must still explain why the person finds it *necessary* to act on these normative facts, or what it is about her that makes them normative *for her*.

Suppose that, as this person believes, there is something that she must do, in the decisive-reason-implying sense. Realists must still explain, Korsgaard writes, why this person *finds it necessary* to act on this normative fact, by doing what she believes that she

must do. Korsgaard might be asking why this person believes it to be *normatively* necessary to do what she believes that she must do. But realists might be able to answer that question. In believing that she must do something in the decisive-reason-implying sense, this person would be believing that this act is normatively necessary; and realists might be able to explain this person's reasons for having this belief.

Korsgaard may instead mean that realists must still explain why this person finds it *psychologically* necessary to do what she believes that she must do. When this person acts on these normative facts, Korsgaard writes, we must explain what makes these facts 'normative *for her*.' Korsgaard seems to be asking here what makes this person's normative belief *motivate* her. As Korsgaard goes on to write

> We must explain how these reasons *get a grip* on the agent. [816]

If Korsgaard is using 'normative for her' to mean in part 'motivates her', she would be giving an account of decisive reasons, and of practical necessity, of the kind that Falk and Williams give. On this account, some act is practically necessary, or is what we must do, when there are facts belief in which would irresistibly move us to act in this way. Korsgaard would add that such practical necessity involves, or is created by, our will.

We have returned to our central question: how we should understand normativity. Korsgaard would be right to claim that, when realists appeal to facts about what is normatively necessary, or about what we must do in the decisive-reason-implying sense, these people do not thereby explain how we are *motivated* to act in these ways. That is an objection to normative realism if, like many Naturalists and Non-Cognitivists, we assume that normativity is, or consists in, some kind of actual or hypothetical motivating force. But realists reject that assumption. When realists claim that we have decisive reasons to act in certain ways, they are not making claims about how, even under ideal conditions, we would be motivated or moved to act. On this view, as I have said, normativity is wholly different from, and does not include, motivating force.

There is a powerful objection, Korsgaard also claims, to any realist view. Realists face an infinite regress from which they cannot escape. When Korsgaard presents this objection, however, she ignores the replies that normative realists would make. She writes, for example:

> I ask you why you are doing some ordinary thing, and you give me your proximate reason, your immediate end. I then ask why you want that, and most likely you mention some larger end or project. I can press on, demanding your reason at every step, until we reach the moment when you are out of answers.

But Korsgaard then writes

> You have shown that your action is calculated to assist you in achieving what you think is desirable on the whole, what you have determined that you want most. [817]

Korsgaard here assumes that, in judging something to be desirable, we are judging that this thing is what we want most. If that were what we meant by 'desirable', Korsgaard would be right to claim that we would soon run out of answers. We would soon reach some desire for which we could give no further desire-based justification. But Korsgaard's realists are Objectivists about Reasons. Our aims are desirable, these realists believe, when these aims have features that give us reasons to have these aims, and to try to achieve them. If we have decisive or sufficient reasons to have our aims, we would not, as Korsgaard claims, run out of answers. We would answer by appealing to these reasons.

Korsgaard then supposes that we have adopted the maxim:

> 'I will do this action, in order to get what I desire'.

She comments:

> According to Kant, this maxim only determines your will if you have adopted another maxim that makes it your end to get what you desire. This maxim is:
>
> > 'I will make it my end to have the things that I desire'.
>
> Now suppose that I want to know why you have adopted this maxim. Why should you try to satisfy your desires?

This is a good question, which rightly challenges subjective desire-based theories about reasons. But if we accept some objective theory, we do not appeal to our desires. We appeal to the facts that give us reasons to have these desires. Our maxim might be:

> I will make it my end to achieve what I have most reason to try to achieve, because these are the ends that are most worth achieving.

Korsgaard's question would then become:

> Why should you try to achieve what you have most reason to try to achieve?

Since 'Why?' asks for a reason, this would mean

> What reasons do you have to try to achieve what you have most reason to try to achieve?

This question answers itself.

Korsgaard also writes:

> We are here confronted with a deep problem of a familiar kind. If you can give a reason, you have derived it from some more fundamental maxim, and I can ask

> why you have adopted that one. If you cannot, it looks as if your principle was randomly selected. Obviously, to put an end to a regress like this, we need a principle about which it is impossible, unnecessary, or incoherent to ask why a free person would have chosen it.

As before, Korsgaard ignores the realist's view. Any reason, she assumes, must be derived from some maxim, or principle, which we have *adopted*. To solve Korsgaard's problem, we must find some principle about which we cannot or need not ask why we have *chosen* it. According to realists, we can appeal instead to truths about what we have reason to want, and do. If there are such truths, these are not principles that we *adopt* or *choose*. We *believe* truths. And if we both believe such truths, and know why we ought to believe them, that would end Korsgaard's justificatory regress. Though it would not be impossible or incoherent to ask why we ought to believe these truths, this question would be unnecessary, since we would know the answer.

In trying to answer the normative question, Korsgaard adds, we are engaged in what Kant called 'the search for the unconditioned'. We are looking

> for something which will bring the reiteration of 'but why must I do that?' to an end. . . The realist move is to bring this regress to an end by *fiat*: he declares that some things are intrinsically normative. . . .

It isn't *realists* who end this regress by *fiat*. A *fiat* is an imperative, or command like 'Do that!' or 'Let that be done!' Unlike Korsgaard, realists do not believe that we can make something normative by commanding or willing that to be so.

Nor do realists merely *declare* that some truths are normative. Realists believe that, as Korsgaard writes, when we ask normative questions 'there is something. . . that we are trying to find out.' On their view, such questions can have true answers.

On Korsgaard's view, even if there were such truths, they could not answer normative questions. To end the justificatory regress, we must appeal to motivational necessity, and to our own will. That, I have argued, is not so. Motivational necessities are not reasons, nor are they normative. And Korsgaard's regress could not be ended except in the way that she rejects. If we knew both *that* and *why* we must do something, we could not then sensibly ask 'But why must we do it?' [818]

There is something right in Korsgaard's view. Our practical reasoning should not end with such normative beliefs. To be fully practically rational, we must respond to practical reasons or apparent reasons with our desires and acts. But only normative truths can answer practical questions. Normativity is not created by our will. What is normative are certain truths about what we have *reasons* to want, or will, or do.

CHAPTER 29 NON-NATURALIST METAPHYSICS AND EPISTEMOLOGY

103 Metaphysical Objections

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 most of the people whose views I have been discussing. These people are Metaphysical Naturalists, who believe that all properties and facts must be natural properties and facts. Irreducibly normative truths, these people assume, would involve 'curious metaphysical objects like Plato's forms', which are 'entities of a very strange sort, utterly different from anything else in the Universe'. Since we could not have any way of knowing about these strange entities, belief in them cannot be part of any scientific world-view. Though we cannot prove that there are no objective values, Gauthier writes, we also cannot prove 'that there are no fairies at the bottom of the garden. We are content to put objective value on a par with the fairies'.

These metaphysical and epistemological objections raise some deep and difficult questions. Some people believe that these objections succeed, thereby answering these questions. Blackburn, for example, writes: 'there is precious little surprising left about morality: its meta-theory seems to me pretty well exhaustively understood.' This meta-theory seems to me very far from being understood. When we consider both morality and practical and epistemic reasons, there are, I believe, several relevant and fundamental questions that we haven't answered. Some of these questions are about normativity. Others are wider questions about the nature and status of necessary truths, whether and in what ways such truths must have truth-makers, and how we can understand and recognize such truths. There are also, I assume, some relevant and important questions that we haven't even asked. Before we understand these questions better, we cannot claim to know whether there might be, or could not be, some irreducibly normative, reason-involving truths.

If there are no such truths, nothing matters. Since I want things to matter, I cannot pretend to be an impartial judge. But some desires are fulfilled. Though we are in the dark, I believe that we can dimly see how there might be such truths, and how we might be able to recognize them.

It may help to look first at the disagreement between *platonists* and *nominalists* about whether numbers and other abstract entities exist. Though numbers are unlike normative reasons, their existence has been denied on similar grounds.

Platonists and nominalists both believe that, if there are numbers and other purely abstract entities, these entities do not exist in space or time. We cannot see or touch numbers, or detect them with our scientific instruments. But platonists claim that such entities exist in some other way, or in some other part of reality. Nominalists reject this claim. Quine for example, writes:

> We do not believe in abstract entities. No one supposes that abstract entities. . . exist in spacetime; but we mean more than this. We renounce them altogether.

We can ask: '*What* more does Quine mean? What is he renouncing or denying?'

The answer, Quine suggests, could not be simpler. We all understand the question 'What is there?' Quine means: 'There are no abstract entities'. On what we can call this

> *single sense view*: When we claim that certain things exist, or that there *are* such things, we always use the words 'exist' and 'are' in the same, familiar sense. We know what it is for rocks or stars to exist. If numbers exist, though they are not in space or time, they exist in the very same sense.

When nominalists accept this view, they can be led to extreme conclusions. Hartry Field, for example, claims that there is no sense in which numbers or other abstract entities exist. If we said that

> (A) there are prime numbers that are greater than 100,

this claim would be about nothing. Field therefore concludes that such claims cannot be true.

Such versions of nominalism are hard to defend. Field also writes:

> the nominalistic objection to using real numbers was not on the grounds of their uncountability. . . the objection was to their abstractness: even postulating *one* real number would have been a violation of nominalism as I'm conceiving it.

These claims are about *uncountability*, *abstractness*, and *nominalism.* Since these would also be abstract entities, Field's view implies that these claims are also about nothing, and cannot be true. Field might accept this conclusion, since he argues that arithmetic, though not true, is a useful fiction. He might say the same about his nominalist view.

Most nominalists defend less extreme views, which they claim to be true. These people admit that the words 'are' and 'exist' can be used in different senses. Quine, for example, qualifies the single sense view. He concedes that in a 'popular' but 'misleading manner of speaking' we can allow ourselves to say that there are numbers, such as prime numbers greater than 100. But such 'casual remarks', he writes, 'would want dusting up when our thoughts turn seriously ontological'. When we speak seriously, we should claim that, in the 'literal and basic' sense, numbers do not *really* exist. Many other writers make such claims. Cian Dorr, for example, distinguishes between the *superficial* and *fundamental* senses of the phrases 'there are' and 'there exist'.

Given this distinction, we can redescribe this disagreement. We can ask

> Q1: Do numbers really exist in the fundamental, ontological sense, though they do not exist in space or time?

Platonists answer Yes. Nominalists answer No.

These are not the only possible views. According to a third view, Q1 is too unclear to have an answer. We can call this *the No Clear Question View*.

Of those who claim to be nominalists, some may really accept this third view. These people may believe that there is no clear ontological sense in which it could be claimed that numbers really exist, though they do not exist in space or time. This view is *not*, however, a form of nominalism. If Q1 is too unclear to have an answer, we should not claim this answer to be No. We are considering the platonist belief that

> (B) numbers really exist in the fundamental ontological sense, though not in space or time.

Nominalists believe that (B) is false. On the No Clear Question View, (B) is not clear enough either to be true, or to be false. Since nominalists believe that (B) *is* clear enough, these are different and conflicting views.

In explaining this third view, we need a distinction that is often overlooked. Consider the claim that

> (C) there are some colourless green ideas which sleep furiously.

Though (C) is in one sense meaningless, or nonsensical, there is another sense in which (C)'s meaning is clear enough. We can reply that

> (D) there could not possibly be ideas which are green, or sleep furiously.

Since (D) makes sense, and is true, (C) makes sense, and is false. Another such claim

is

> (E) there are some headaches which are correct, and others which are mistaken.

I claimed earlier that we could not defend (E), because we could not explain what it would *be* for some headache to be correct, or mistaken. In denying (E), however, I was not claiming that (E) makes no sense. (E) makes sense, and is clearly false. No headache could be correct or mistaken.

Some people assume that, to know what some claim or assertion means, we must know this claim's *truth condition*, or how things would be if this claim were true. But that is not so. We know the meaning of the claim that

> (F) 2 + 2 = 5,

though we cannot conceive how things would be if (F) were true.

It might be objected that, compared with (F) or the claim that

> (C) there are some colourless green ideas which sleep furiously,

it makes *more* sense to claim that

> (B) numbers really exist in the fundamental ontological sense, though not in space or time.

But this objection, I believe, misdescribes the difference between these claims. Rather than claiming that, compared with (F) and (C), (B) makes more sense, we should say that (B) is closer to being a claim that might be true. (B) is closer to being such a claim because it is *less clear* what (B) means. We cannot imagine coming to understand how it might be true either that 2 + 2 = 5, or that there are colourless green ideas which sleep furiously. When platonists assert (B), we can more easily imagine that these people might be able to explain or restate their view so that we can understand how this view might be true. (B), however, needs to be further explained. It is not clear enough in what sense it might be true, or might be false, that numbers really exist.

Several people have tried to explain (B). Dorr, for example, writes:

> There are no numbers. There are no properties. When I utter these sentences, I mean to be using them in the fundamental way. I mean, if you like, that numbers and properties are not part of the ultimate furniture of reality. . . there are, in the final analysis, no such things.

When Dorr uses the word 'reality' he cannot mean 'what exists in space or time'. If that were what Dorr meant, he would not be rejecting platonism, since platonists agree

that numbers do not exist in space or time. Since Dorr is not using 'reality' to mean 'the spatio-temporal world', it is not enough for him to say that, according to platonists, numbers *really* are in a *fundamental* sense part of the *ultimate furniture of reality*. These words do not sufficiently explain what it is that platonists assert, and nominalists like Dorr deny.

Dorr is aware of this objection. There are some people, he writes, who

> see *no alternative* to the superficial way of using sentences like 'there are numbers'. They simply have no idea what the allegedly distinct 'fundamental' uses of these sentences are supposed to be.

To explain these uses of the words 'there are' and 'there exist', Dorr appeals to the widely held view that it cannot follow from the meaning of some word that something exists. We cannot, for example, prove that God exists by appealing to the meaning of the word 'God'. On this suggestion, platonists could explain their view by claiming that

> (G) numbers exist in the sense in which God exists, or might exist.

As we shall later see, this suggestion helps. But some platonists might deny that they could restate their view as (G). Most platonists believe that numbers *necessarily* exist, and do not exist in space or time. Many of these people believe either that God does not exist, or that God could not necessarily exist, and that God either does exist, or would exist, might space and time. Given these differences, some of these platonists would claim that numbers exist in a different sense from that in which God exists, or might exist.

Dorr also writes that, when we make true claims about what exists in the superficial sense, these claims could all be paraphrased or restated as claims about what really exists in the fundamental sense. We can thereby make true claims that seem to be about abstract objects, without committing ourselves to the belief that there really are such objects. In all such cases, Dorr suggests, our paraphrase could start: 'If there were abstract objects, it would be true that . . .' If we find it hard to suppose that there are abstract objects, we could substitute 'According to the fiction that. . . '

These suggestions, I believe, do not help. As Dorr rightly claims, we can understand some impossible fictions. When we give some proof which is a *reductio ad absurdum*, we may start by supposing that something impossible is true, and then show that this assumption leads to a contradiction. Since we can understand the claim that 2 + 2 = 5, we might pretend that this claim is true. But Dorr is trying to explain the sense in which platonists assert and nominalists deny that

> (B) numbers really exist in the fundamental ontological sense, though not in space or time.

If we don't understand (B), we cannot try to suppose or pretend that (B) is true. We don't know what we should try to suppose, or pretend.

When Bob Hale discusses the No Clear Question View, he responds in a different way. Unlike Dorr, who believes that (B) could not possibly be true, Hale is a platonist who believes that (B) could not possibly be false. Critics of platonism sometimes ask, Hale writes, what difference it would make if (B) were false. What would the world be like if numbers and other abstract objects didn't exist? Hale replies that, since he believes that these objects *necessarily* exist, he cannot be reasonably expected to explain what it would be for these objects not to exist.

This reply is in one way justified. We could similarly claim that, since it is a necessary truth that 2 + 2 = 4, we cannot be expected to explain what the world would be like if 2 + 2 did *not* = 4. But Hale's reply does not help to explain what platonists believe. If we believe that (B) is too unclear either to be true, or to be false, it does not help to be told that (B) is necessarily true. As before, we don't know *what* is being claimed to be a necessary truth.

If we accept the No Clear Question View, we may not be rejecting (B) as nonsense, or gibberish. When someone says that someone else's theory 'isn't even false', or 'isn't even wrong', this person may be expressing contempt. We may have no such attitude to platonism and nominalism. Dorr talks of sceptics who are 'determined not to understand' what is meant by claims like (B). But we may be eager to understand this claim, and regret that we have so far failed. Since we have failed, however, we are still inclined to believe that

> (H) numbers are not a kind of entity about which it is a good question whether, in some ontological sense, they exist, or are real.

We can next consider another, more positive view. We are *Cognitivists* about some kind of claim if we believe that such claims can be, in a strong sense, true. Many such claims have metaphysical or ontological implications. In trying to decide whether these claims are true, we must answer questions or make assumptions about what exists. That is true, for example, of claims about rocks, stars, philosophers, and bluebell woods. And it may be true of all claims about features of the spatio-temporal world. When we believe that claims of these kinds can be in a strong sense true, we are *Metaphysical Cognitivists* about such truths.

When we consider certain other kinds of claim, we may accept a different view, which we can call *Non-Metaphysical Cognitivism*. On this view, these other claims can be in a strong sense true; but in trying to decide whether these claims are true, we don't need to answer questions about what exists, in any metaphysical or ontological sense.

Some examples are claims about numbers, sets, and mathematical proofs. On this view, for such claims to be true, there must be a sense in which these entities exist. But this sense of 'exist' is *not* ontological. We can reject the nominalist's claim that such entities exist only in a *superficial* sense. There is nothing superficial in the question whether there are numbers that have certain properties, or whether there exists some undiscovered proof of some theorem. And there can be several grounds for denying that certain abstract entities exist. When we try to refer to such entities, for example, we may fail, because our concepts are too unclear, or indeterminate. Our description of such entities may be in some way inconsistent, or lead to some contradiction. In such cases, we should deny that such entities exist. There is, for example, no set that contains all and only those sets which do not contain themselves, since the claim that there is such a set would involve a contradiction. And when we refer to such abstract entities as numbers or sets, and claim that they have certain properties, these claims may be false. But if our claims avoid these and similar objections, these claims cannot be false in some metaphysical or ontological way. One example is the claim that

(A) there are prime numbers greater than 100.

If our way of referring to numbers is sufficiently clear, and avoids contradictions, and our calculation shows that some number greater than 100 is a prime, that is enough for (A) to be, in a strong sense, true. We need not fear that (A) might be false because numbers don't really exist. According to this form of Non-Metaphysical Cognitivism,

(I) numbers exist in a fundamental but non-ontological sense.

Remember next that, on

> the single sense view: When we claim that certain things exist, or that *there are* such things, or that such things are real, the words 'exist', 'are', and 'real' always have the same, familiar sense.

If we accept some form of Non-Metaphysical Cognitivism, we must reject this view. We believe that these words can be used in at least two fundamental senses, one of which is not ontological.

There is, I believe, no decisive argument for the single sense view. Many other words have different though related senses. Some examples may be the words 'true', 'meaning', 'abstract', 'possible', 'necessary', 'reason', 'ought', 'property', 'entity', and 'thing'. If we are not extreme nominalists, we shall believe that there are different kinds of thing which exist in different ways. Two examples are commercial firms and the buildings they occupy. These two kinds of entity do not exist in the same way. Firms are created by a legal process, and they can cease to exist by going bankrupt and

being abolished. Buildings can be built by brick-layers, and they can cease to exist by burning down. Some philosophers claim that there are really no such things as commercial firms. Others claim that there are not even such things as buildings, or people, since there are only the fundamental entities studied by physics, which might be sub-atomic particles, or quarks. We can reply that, though there may be one sense in what exists might only be particles or quarks, there are other senses in which *we* exist, as do such things or entities as buildings, commercial firms, armies, nations, the books we write, and piano sonatas. There are also senses in which there are events, such as performances of these sonatas, economic depressions, famines, wars, and suffering. Nor do such things exist or occur only in a superficial sense. There is nothing superficial in our own existence, or the existence or occurrence of books, nations, armies, wars, famines, and suffering.

It might next be claimed that

> (J) though the words 'are', 'exist', and 'real' can have such different senses, these senses are all ontological.

On this view, if we claim that

> (A) there are prime numbers greater than 100,

we must mean that

> (K) prime numbers greater than 100 really exist in some ontological sense.

On one version of this view, though numbers are not physical entities, like particles or quarks, there are numbers in the same sense in which there are such entities as nations, laws, or piano sonatas. Just as composers do not *discover* but *create* piano sonatas, mathematicians create prime numbers and proofs. Numbers and proofs are part of reality in the sense that they are created or invented by mathematicians.

For reasons I shall soon give, we ought, I believe, to reject (J). We can understand the claim that, in a fundamental but non-ontological sense, there are prime numbers greater than 100, and there are some undiscovered mathematical proofs. Platonists and nominalists sometimes suggest that we should take claims like (A) at their *face value*, and accept that mathematicians *mean* what they *say*. But when mathematicians claim that there are numbers with certain properties, or predict that there may be some undiscovered proof, many of these people do not mean that these numbers and proofs exist, or are real, in some ontological sense. That is why these people never worry that arithmetic might all be false, because numbers don't really exist.

Of the four views that I have now described, some cannot be coherently combined. If we are platonists or nominalists, as I have said, we cannot coherently accept the No Clear Question View. Return to the claim that

> (B) numbers really exist in the fundamental ontological sense, though not in space or time.

If we believe that (B) is too unclear either to be true, or to be false, we cannot also be platonists, who believe that (B) is true, or be nominalists, who believe that (B) is false.

Nominalists might, however, be Non-Metaphysical Cognitivists. On this view, though it is false that numbers really exist in the fundamental ontological sense, numbers do exist in a fundamental but non-ontological sense. This form of nominalism is very different from the simpler form of nominalism described by Quine, Field, and Dorr. If nominalists are also Non-Metaphysical Cognitivists, they would reject the view that mathematical claims cannot be true, or are only superficially true, or are useful fictions. These nominalists would believe that some mathematical claims can be in the strongest sense true. Unlike claims about objects in the spatio-temporal world, these mathematical claims could not conceivably have failed to be true.

Platonists might also be Non-Metaphysical Cognitivists. On this view, numbers really exist in the fundamental ontological sense, but numbers also exist in a fundamental non-ontological sense. It may be somewhat harder, though, to combine these two views, since we may doubt that anything could exist in both of two such different senses. And if we are Non-Metaphysical Cognitivists, we might claim that platonism is too close to simple nominalism. Both views, we might say, mistakenly assume that claims cannot be in a strong sense true unless they are about what exists in some part of reality.

Of those who claim to be platonists, some may really be Non-Metaphysical Cognitivists. These people may be trying to show that our ways of referring to numbers are sufficiently clear, and involve no inconsistency, and they may assume that, if these things are true, numbers exist in a fundamental sense, which is the only relevant sense. This, for example, may be the view that some people call *neo-Fregean platonism*.

If nominalists or platonists are also Non-Metaphysical Cognitivists, the disagreement between these people would be less deep. These people would all believe that numbers exist in a fundamental non-ontological sense. If that is true, it would be a less important question whether numbers also exist in some ontological sense. Numbers would not need to exist in that other sense for claims about them to be, in a strong sense, true.

Return now to the No Clear Question View. If we accept this view, we might reject Non-Metaphysical Cognitivism. We might believe that there is no clear *non*-ontological sense in which numbers exist, or believe that numbers exist only in a superficial sense. We might also believe that arithmetical claims are really about mathematical symbols, or about mathematical procedures or constructions, or something of the kind. And we might believe that claims about numbers cannot be in any strong sense true.

The No Clear Question View might instead be combined with Non-Metaphysical Cognitivism. On this view, though there is no clear ontological sense in which numbers exist, claims about numbers can be in a strong sense true, since numbers exist in a fundamental non-ontological sense.

When we consider arithmetic, we ought, I suggest, to be Non-Metaphysical Cognitivists. We might justifiably combine this view with one of the other three views.

We can now apply these distinctions to some other questions. I believe that

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

I discuss this possibility in Appendix A. Here is one quick argument against (L). Someone might argue:

> (M) It could not have been true that nothing ever existed. If that had been true, there would have been the truth that nothing existed. So your alleged possibility is self-contradictory.

This objection, I believe, fails. We can claim that, like numbers,

> (N) truths are not a kind of entity about which it is a good question whether, in some ontological sense, they exist, or are real.

The relevant question is only whether something is, or would have been, *true*. Most truths are true only because things of some other kind exist, in an ontological sense. But truths do not have to exist, in some part of reality. Truths need only be true. We should admit that, in one sense, truths exist. We should not claim that

> (O) if it had been true that nothing ever existed, there would not even have been the truth that nothing existed.

We should admit that, if this had been true, there *would have been* this truth. But this claim would use the phrase 'would have been' in a non-ontological sense. In making these claims, we would be combining the No Clear Question View with another form of Non-Metaphysical Cognitivism.

Similar remarks may apply to possibilities, and to possible objects or events. We might claim that, like numbers and truths,

> (P) possibilities are not a kind of entity about which it is a good question whether, in some ontological sense, they really exist.

It might have been true, I have claimed, that nothing ever existed. If that is true, this would be one possibility. It is more obviously true that we ourselves might never have existed, because our parents might never have met, or never had children. It may also be true that we could later have children, or could act in several other ways. When we ask whether these things are true, it does not help to ask whether there really exist these possibilities, or these possible children, or possible future acts. On this view, possibilities or possible entities need not exist in some part of reality. They merely need to be possible. It can be of great importance whether there *is*, or *was*, some possibility. And we might also say that there is a *real* possibility that a certain thing will happen. But in making these claims we would be using the words 'is', 'was', and 'real' in non-ontological senses.

We can next compare these claims:

> (Q) Numbers exist, though not in space or time.
>
> (R) God exists, though not in space or time.

Non-Metaphysical Cognitivism can be plausibly applied to (Q), since we can plausibly believe that numbers exist in a deep and fundamental but non-ontological sense. But this view applies less plausibly to (R). On most people's views, and as I shall here assume, God could not be an abstract entity. Partly for that reason, it must be a metaphysical question whether (R) is true. In our beliefs about God, we cannot be Non-Metaphysical Cognitivists. If God exists, that would have to be true in an ontological sense.

When we ask whether (R) is true, we can plausibly accept the No Clear Question View. We can believe that, though it may make sense to claim that God exists *in* space and time, it is not clear enough what it would be for God to exist in a fundamental ontological sense, though God does not exist in space and time. For (R) to be true, God would have to exist in some non-spatio-temporal part of reality, like the Platonic realm in which, on Plato's view, there exists the Form of the Good. We may find it hard to understand such claims.

As before, however, there is a difference between (Q) and (R). If we are Non-Metaphysical Cognitivists about numbers, we can plausibly reject arithmetical platonism by appealing to a version of the No Clear Question View. We can claim that, since numbers are abstract entities, there is no clear ontological sense in which it might be true, or be false, that numbers exist in some non-spatio-temporal part of reality. Since God could *not* be an abstract entity, we should make a different and weaker claim about whether God might exist, though not in space or time. We can vaguely understand the possibility that space and time are not metaphysically fundamental. It makes sense to suppose that there is some entity that is more fundamental, and that both space and time metaphysically depend on this other entity. When they discuss the Big Bang, some physicists suggest hypotheses of this kind, and many people make such claims about God. We should admit that (R) might be made clearer, and that this clearer claim might be true or false. I discuss such views further in Appendix A.

We can now return to one of Dorr's remarks about the fundamental sense in which, as platonists claim and nominalists deny, numbers exist. Dorr writes:

> it is not an analytic truth that there are numbers, since it is not an analytic truth that there is anything at all. As Hume and Kant maintained in criticizing the standard a priori arguments for God's existence, denials of existence---when taken in the fundamental sense---cannot be self-contradictory.

We should agree, I believe, that *God's* existence cannot be proved merely by appealing to our concepts. According to one such argument, the concept of God is the concept of the most perfect being, and since such a being would be more perfect if this being exists, God must exist. That argument is unsound, as is suggested by the atheist's reply: 'God is so perfect that He doesn't even need to exist'.

Such claims do not, I believe, apply to numbers, or to some other kinds of abstract entity. It cannot follow merely from the concept of a number that numbers exist. But we may be able to prove that numbers exist by appealing to our concepts, and giving one or more a priori arguments. Such a proof may be possible *because* it is not true that numbers exist in some ontological sense. We can accept Dorr's claim that there could not be a priori arguments which showed that something exists in the ontological sense. But this sense of 'exists' is not the only important sense. If nothing had ever existed in the ontological sense, there would not have been any stars or atoms, nor would there have been space, or time, or God. But it would have been true that nothing ever existed. In other words, there *would have been* the truth that nothing existed in this sense. This truth would have *existed* in the *non*-ontological sense. It would also have been true that there are prime numbers greater than 100. There would have been such numbers, in this other sense. And there would also have been various other possibilities. According to Kronecker, God made the natural

numbers, and men made all the others. Numbers, we can defensibly believe, did not need to be made, even by God. Nor did possibilities.

We can now return from numbers to irreducibly normative truths, such as truths about practical and epistemic reasons. When people deny that there could be such truths, they often give metaphysical objections. Irreducibly normative reasons, these people claim, are entities which are 'too queer' to be part of the 'fabric of the Universe'.

There may seem to be a simple way to avoid such objections. Rather than saying that certain natural facts *give* us reasons to have certain beliefs or desires, we can say that these facts *are* reasons to have these beliefs or desires. On this account, normative reasons are not strange entities, since such reasons are natural facts. Or we might say that certain natural facts *count in favour* of our having certain beliefs or desires. But these are merely different ways of saying the same thing. Such facts, we believe, have the irreducibly normative property of *being a reason* or *counting in favour*. Metaphysical Naturalists would deny that any facts could have such a property. These people believe that there cannot be any irreducibly normative, reason-involving truths. It makes no relevant difference in which of these ways we state these truths.

These normative truths, I have claimed, are not like physical or psychological facts. Nor are they like facts that are normative in the rule-involving sense. We can explain how, by acting in certain ways, people can make it true that some act is illegal, or bad etiquette, or involves an incorrect spelling, or the misuse of some word. That is why, when it is true that some act is illegal, or is incorrect in such other ways, these are natural, empirically discoverable facts. No such claims apply to truths that are normative in the reason-involving sense. We cannot make it true that certain facts give us reasons, or that certain acts are wrong. Such normative truths are not natural, empirically discoverable facts.

According to the metaphysical objections, since such alleged normative truths could not be natural facts, any belief in such truths can be quickly dismissed. Gibbard, for example, writes:

> If this is what anyone seriously believes, then I simply want to debunk it. Nothing in a plausible, naturalistic picture of our place in the universe requires these non-natural facts.

To non-philosophers, Gibbard adds, such claims 'sound fantastic'. In several other recent books, such views are rejected in a paragraph or two. Jackson thinks it worth explaining why he even bothers to discuss such views. When Field considers the view that there are some non-natural normative properties, he calls this view 'crazy'.

To answer such objections, we might first appeal to the No Clear Question View. We

might claim that

> (S) normative reasons and reason-involving properties are not kinds of entity about which it is a good question whether, in some ontological sense, they exist, or are real.

We should also appeal, I believe, to another form of Non-Metaphysical Cognitivism. On this view:

> (T) There are some claims that are irreducibly normative in the reason-involving sense, and are in a strong sense true. But in trying to decide whether such claims are true, we need not answer ontological questions. For such claims to be true, it need not be true that reason-involving properties exist either in the spatio-temporal world, or in some non-spatio-temporal part of reality.

Though many such normative claims are about what is merely possible, some of these claims are about entities or events that exist or occur in the spatio-temporal world. One example would be my claim that, in killing your enemy in revenge, you acted wrongly, and did what you had decisive reasons not to do. But in claiming that your act had these normative properties, I would not be describing further features of the spatio-temporal world. That is like the way in which the symbols written one some page may be a valid proof of some theorem. Though these symbols exist in the spatio-temporal world, their property of being a valid proof is not a further feature of this world. As Nagel writes, such normative claims 'need not (and in my view should not) have any metaphysical content whatever.' We can call this view *Non-Metaphysical Non-Naturalist Cognitivism*.

Non-Naturalists, Korsgaard suggests, claim to have 'spotted some normative entities, as it were wafting by'. That, we can claim, is *not* what we believe. Our view cannot be metaphysically implausible, since it is not a metaphysical view. More exactly, ours is a *negative* metaphysical view, since we believe that, though there are some irreducibly normative truths, such truths do not involve the real existence, in some ontological sense, of any strange entities or properties.

This view, however, faces other, epistemological objections. If these normative truths are not about features of the spatio-temporal world, how we are able to understand and recognize such truths? If we have not spotted some normative reasons wafting by, and we could not be causally affected by normative properties, how could we know anything about them?

104 Epistemological Objections

When Metaphysical Naturalists give these objections, they often appeal to some *causal* theory of knowledge. On such theories, for us to have reasons to believe that certain entities exist, or that certain facts obtain, we must have some evidence for these beliefs. When some entity or property cannot be directly observed, we would have such evidence if the existence of this entity or property would provide the best explanation of some of what we can observe. That is how physicists justifiably believe in the existence of some sub-atomic particles, since such particles leave tracks when they pass through cloud-chambers. In explaining what we can observe, these Naturalists then argue, we never need to appeal to irreducibly normative entities or properties. In explaining people's acts, we sometimes need to appeal to these people's normative beliefs, such as their belief that certain acts are wrong. But in explaining why these people have these beliefs, we never need to claim that certain acts *are* wrong. It is enough to appeal to how these people were brought up, and to various other psychological and historical facts. These Naturalists now apply Ockham's razor. Simpler theories, they claim, are more likely to be true. Since nothing needs to be explained by appealing to irreducibly normative properties, we should deny that there are such properties.

As well as claiming that people's normative beliefs need not be explained by appealing to such normative properties, many Naturalists claim that we could not possibly explain these beliefs in this way. To know about such properties, we would need to have some mysterious quasi-sensory faculty. Gibbard, for example, writes that, according to Non-Naturalists like Sidgwick and Moore,

> among the facts of the world are facts of what is rational and what is not. A person of normal mental powers can discern these facts. Judgments of rationality are thus straightforward apprehensions of fact, not through sense perception but through a mental faculty analogous to sense perception.

This description is misleading. When Gibbard writes that, on such views, facts about what is rational 'are among the facts of the world', that phrase suggests that, according to these Non-Naturalists, the rationality of some desire or act is a feature of the spatio-temporal world. And when Gibbard talks of a faculty analogous to sense perception, that phrase suggests that, according to these Non-Naturalists, we detect the presence of rationality by being causally affected by this property, as when we feel the heat of the Sun or see the craters on the Moon.

Most Non-Naturalists do not hold such views. Sidgwick and others claim that we can understand and recognize some irreducibly normative truths in something like the intuitive way in which we can understand and recognize some logical or mathematical truths. These abilities overlap, since in thinking logically or mathematically we respond to epistemic reasons, and follow rules of inference. When Sidgwick calls our knowledge of such truths *intuitive*, he is not appealing to any special faculty that is like sense perception. He explicitly rejects this analogy. Sidgwick means only that there

are certain beliefs that we are justified in having, not because they are implied by other beliefs, nor because we have evidence for them, but because of their content, or *what* we are believing. That is also what Sidgwick means when he calls such beliefs 'self-evident'. Sidgwick does not use 'self-evident' to mean 'obviously true'. Apparently self-evident beliefs, he claims, may be false. Such claims, we can also say, are *intrinsically* plausible.

In suggesting how we can have such justified true beliefs, we can first explain why we do not need to have some quasi-perceptual faculty. When we have beliefs about features of the spatio-temporal world, these beliefs are about facts that are *contingent*, in the sense that the world might not have had these features. We could not hope to form such true beliefs unless we or others have causally interacted with these features, in ways that explain and justify these beliefs.

No such claims apply to *necessary* truths, such as mathematical truths. It is a not a contingent feature of our world that 7 is a prime number, or that 43 + 9 = 52. Such claims would be true in every possible world. That is why, though it is natural to say that we can see that such claims are true, this use of 'see' is a mere metaphor. When some truth is not contingent, we have no reason to assume that we could know about this truth only through some form of causal interaction.

Similar remarks apply to normative properties and truths. We don't have to find out that ours is a world in which we have decisive reasons to believe that 7 is a prime number. Nor do we have to find out that ours is a world in which people have reasons to want certain things for their own sake. Mackie writes that, according to Non-Naturalists, we must

> ascertain which of various possible worlds . . . is the actual one---for example, whether the actual world is one in which pain is prima facie to be relieved, or one in which, other things being equal, pain is to be perpetuated. . . . [The] moral thinker has, as it were, to respond to a value-laden atmosphere that surrounds him in the actual world.

We should reject these claims. Fundamental normative truths are not about how the world happens to be. In any possible world, pain would be in itself bad, or prima facie to be relieved rather than perpetuated. Similarly, even if the laws of nature had been different, rational beings would have had reasons to do what would be needed to achieve their rational ends or aims.

Though we can explain why, to have knowledge of such properties and necessary truths, we don't need to be causally affected by these properties, that is a merely negative claim. To defend our view, we must make some positive suggestions about how we might have such knowledge.

Logical and mathematical truths are often claimed to be analytic, or more broadly conceptual, in the sense that these truths are implied by, or in some way grounded on, the meaning of these claims. Such views might explain how we can recognize these truths. But these views cannot fully explain our knowledge of normative truths, since there are at least some fundamental normative truths which are not analytic, or conceptual. One example is the truth that it is bad to be in pain. I believe that, like truths about what exists in an ontological sense, *no* substantive normative truths could be analytic, or be in some broader sense conceptual claims. But I shall not defend that stronger belief here.

Though there are great differences between mathematical and normative truths, our claims to know about these truths can be challenged in similar ways. Platonists believe that, though numbers are abstract entities, we have many true arithmetical beliefs. Field claims that, since we could not be causally affected by such abstract entities, platonists cannot explain how mathematicians can have so many true beliefs about these entities. This correlation, platonists must admit, would involve some

> massive coincidence. . . We should view with suspicion any claim to know facts about a certain domain if we believe it impossible to explain the reliability of our beliefs about that domain.

Other writers make such claims about our alleged ability to recognize non-natural normative truths. Sharon Street, for example, writes that, though it is '*possible* that as a mere chance' many of our normative beliefs are true, by fitting these alleged normative facts, this would require a 'fluke of luck' that is 'extremely unlikely.' We can call this *the Massive Coincidence Argument*.

This argument does not, I believe, succeed. We can now design computers whose internal circuitry enables them to operate in ways that correspond to logical and mathematical reasoning. Though these computers do not have beliefs, or any other mental states, that is irrelevant here. If there are truths about numbers and their mathematical properties, these computers could reliably produce and print out true answers to mathematical questions. They could produce these answers without having any causal contact with numbers or their mathematical properties.

Similar claims apply to us. It might be true, for example, that God designed our brains so that we can reason in ways that lead us to reach true answers to mathematical questions. We might have similar God-given abilities to understand and recognize some irreducibly normative truths. These abilities are in part the same, since we form true mathematical beliefs by responding to epistemic reasons. God might have designed our brains so that we can respond to these and other

normative reasons. This ability need not involve some mysterious faculty of quasi-sensory intuition. When some fact has the property of *being* or *giving us* a reason, this property is not a feature of the natural world, so we cannot causally interact with such properties. We respond to reasons when we are aware of facts that give us such reasons, and this awareness leads us to believe, or want, or do what we have these reasons to believe, or want, or do. As the facts about computers show, we might be able to respond to such reasons without being causally affected by the normative properties of the reason-giving facts. When these computers produce true answers to mathematical questions, they are not responding to reasons, or following rules of inference, because computers have no mental states. But this analogy shows how, without any such causal contact, we might be able to respond to reasons, and form true beliefs about them.

Though Field admits that God might have given us such mathematical abilities, he assumes that we can exclude this possibility, since our brains were not intentionally designed by God. We can now ask whether, on that assumption, we can reject the Massive Coincidence Argument. Can we suggest some other way in which we might have have developed such abilities?

Field assumes the answer to be No. We should reject platonism, Field writes, because

> mathematical entities as the platonist conceives them exist outside of space and time and bear no causal relations to us or anything we can observe; and there just don't seem to be any mechanisms that could explain how the existence of and properties of such entities could be known.

There is, however, another possibility, which Field surprisingly ignores. Our brains may have been *unintentionally* designed by evolution. It may be true that, just as cheetahs were selected for their speed, and giraffes were selected for their long necks, we were selected for our rationality. Evolution may explain how human beings became able to respond to reasons.

Of the ways in which humans differ from other animals, the most fundamental are that we use language, and respond to reasons. Of these two, I suggest, it is our rationality that we began to acquire earlier, and that is more fundamental. We are members of the species *homo sapiens*, the animal that is *clever* or intelligent in its thoughts and acts, though it may not be wise. Though our use of language has immense importance, it is not what enabled us to think rationally. Language enabled us to have much greater numbers of more precise thoughts, and to share these thoughts. Many animals can form true beliefs about what is happening, or will soon happen, in their immediate environment. Because we can respond to reasons, we are able to form many other kinds of true belief, especially beliefs about the further future, and the long term effects of different possible acts. The ability of early humans to form such true beliefs had great evolutionary advantages, by

helping them to survive and reproduce. Natural selection slowly but steadily gave later humans greater cognitive abilities. Just as the faster cheetahs and taller giraffes tended to survive longer and have more offspring, who inherited similar qualities, so did the humans who were slightly better at responding to reasons.

When Nagel discusses the view that evolution explains our rationality, he calls this view 'laughably inadequate'. But Nagel is rejecting more ambitious, reductive views. As Nagel rightly claims, evolutionary theories cannot explain normativity itself, or what *it is* to have reasons, some of them decisive, to have certain beliefs or desires, or to act in certain ways. Nor can such theories explain why we can justifiably rely on our ability to respond to reasons. As Nagel writes:

> Whatever justification reason provides must come from the reasons it discovers, themselves. They cannot get their authority from natural selection. . . I follow the rules of logic because they are correct---not merely because I am biologically programmed to do so.

Nor can evolution explain how it is possible for there to be animals that respond to reasons. Evolution does not create any of the possible forms that living beings can take, but merely selects between these possibilities. As Nagel claims, evolution

> may explain why creatures with vision or reason will survive, but it does not explain how vision or reason are possible.

Nagel also suggests, however, that evolution cannot fully explain how we became able to respond to reasons. We ought, I believe, to reject this suggestion. If there can be animals who can see, or think rationally, evolution can do more than explain why such animals will survive. Evolution can explain how these animals became able to see or to think rationally. When Nagel discusses 'the advanced intellectual capacities of human beings', he calls these 'extremely poor candidates for evolutionary explanation'. This claim underestimates, I believe, what natural selection can achieve.

How most animals developed vision is now fairly well understood. Random mutation gave a few early animals slight sensitivity to light. These animals had a slight advantage over their contemporaries with no such sensitivity, and were therefore slightly more likely to have surviving offspring, whose genes would give them the same sensitivity. After very many more such randomly produced but advantageous mutations, some of the results were the superbly effective eyes of animals like eagles, hawks, and young human beings.

Consider next

> *Three Roads*: While using its sense of smell to track its quarry, some dog reaches a place from which there are only three exits, or roads. This dog goes down

> the first road, sniffs, and comes back. It then goes down the second road, sniffs, and comes back. It then runs down the third road *without sniffing*.

When a Greek Stoic learnt of this event, he conceded that man was not the only animal who can go through a process of reasoning. This dog may have realized, even if unconsciously, that its quarry must have gone down the third road. Early humans reasoned in similar ways, thereby forming many advantageous beliefs, such as beliefs about how they could trap some mammoth, or use the properties of fire. They also acquired some useful mathematical abilities. It would have helped to realize that, if five lions entered some cave, and only four have left, one lion must still be there. Though there are vast differences between such abilities and the genius of Euclid, Newton, or Godel, these are differences of degree, not of kind. These much greater abilities, we can plausibly believe, could have been produced by the natural selection, over thousand or millions of years, of those humans whose rational abilities were slightly greater.

Nagel also suggests that, if evolution fully explained our ability to respond to reasons, this fact would cast doubt on this ability. In his words:

> Without something more, the idea that our rational capacity was the product of natural selection would render reasoning far less trustworthy. . . There would be no reason to trust its results in mathematics and science for example.

When Nagel writes 'without something more', he may mean 'without the justification that reason itself provides'. We should accept this claim, so understood. But this justification would not be undermined or weakened, as Nagel may assume, if our rational abilities were the product of natural selection. Since evolution merely selects between the different possibilities, it does not undermine, or spoil, what it selects.

My claims so far have been these. When some fact gives us a reason to have some belief, this normative property of being reason-giving is not an empirically discoverable feature of the natural world. We could not be causally affected by such normative properties. But evolution might explain how, without any such causal contact, our ancestors became able to respond to such reasons, thereby forming true beliefs about the natural world, in ways that helped them to survive and reproduce. Since evolution can explain how we acquired this ability to respond to reasons, it is not a coincidence that our ancestors could form so many true beliefs. Nor it is surprising that, with their steadily improving rational abilities, and their curiosity, later humans became able to form many other true beliefs, such as beliefs about prime numbers greater than 100, or about black holes and neutron stars, or decisive proofs and epistemic reasons. When applied to our ability to form such true beliefs, this

evolutionary account provides, I believe, a good enough answer to the Massive Coincidence Argument.

We might next make a positive claim, running this argument the other way. Of our reasons to believe in mathematics, it is often said, some are provided by the ways in which physicists use mathematics to predict and understand many features of the natural world. Though these reasons for belief in mathematics are not, I believe, needed, they are strong. We might now give a similar defence of our belief that we have epistemic reasons. Unlike other animals, we form many true beliefs about what we cannot see, hear, touch or smell, such as beliefs about the further future. If we ask how we can form so many such true beliefs, the best answer seems to be: because we are responding, in this non-causal way, to epistemic reasons. If there were no such epistemic reasons, our ability to form these many true beliefs would involve a massive coincidence.

CHAPTER 30 IRREDUCIBLY NORMATIVE TRUTHS

105 Modal and Normative Epistemic Reasons

There are, however, various objections to these claims. I have assumed one view about epistemic reasons. On what we can call this

> *Normative Account*: Some fact gives us an epistemic reason to have some belief when
>
> > (1) this fact makes this belief certain or likely to be true,
>
> so that
>
> > (2) this fact has the irreducibly normative property of counting in favour of our having this belief.

In a fuller statement of this view, we would need to add various qualifications and further claims. According to some people, though, we can drop claim (2). When some fact makes some belief certain or likely to be true, that is enough to give us a reason to have this belief. This fact need not be claimed to have the normative property of *counting in favour* of our having this belief. The concepts *certain* and *likely* are not normative but *modal*, as are the related concepts *necessary*, *probable*, *possible*, and *impossible*. So we can call this *the Modal Account*, which describes *modal* epistemic reasons. We can also call (1) a *modal* belief.

I have suggested that, by appealing to evolution, we can explain how early humans became able to respond to normative epistemic reasons. Metaphysical Naturalists would reject this explanation. What evolution explains, these people might say, is how early humans became able to respond to *modal* epistemic reasons. These humans became able to form true beliefs when, and because, they were aware of facts that made these beliefs certain or likely to be true. That explanation is complete. To be able to form these true beliefs, these humans would not have needed the further ability to respond to *normative* epistemic reasons. Since that further ability would add nothing, evolution cannot explain how our ancestors became able to respond to these alleged reasons. This fact, these Naturalists might claim, undermines my answer to the Massive Coincidence Objection. And since we don't need to appeal to such irreducibly normative reasons, we should use Ockham's razor. We should not believe there *are* any such normative reasons.

Some Naturalists might go further, by denying that we even have *modal* epistemic reasons. Like normative concepts, modal concepts form a separate and somewhat

puzzling category. We can explain some modal concepts by appealing to others, but we cannot explain these concepts in non-modal terms. And modal properties and truths are, in one way, like normative properties and truths. Modal properties are not empirically discoverable features of the spatio-temporal world, and the most fundamental modal truths are not contingent facts. Some Naturalists believe that there could not be such non-empirical properties and truths.

These Naturalists might give a third account of epistemic reasons. To explain how we can form true beliefs, these people might say, we need not claim that we are responding to facts that make these beliefs *certain* or *likely* to be true. It is enough to claim that these facts are in certain ways causally related to the truth of these beliefs. When we see some flash of lightning, for example, this event both causes us to believe that we shall soon hear thunder, and makes our belief true, by causing this thunder. On this *Causal Account*, some fact gives us an epistemic reason to have some belief when such facts are causally related in certain ways to the truth of such beliefs. These Naturalists might then claim that, since we do not need to appeal to modal truths, we should not believe that there are such truths.

This argument, I believe, fails. We can admit that, in explaining how we form many true beliefs, it is enough to claim that we are aware of facts that have certain causal relations to the truth of these beliefs. In my example, it may be enough to claim that the flash of lightning causes both the thunder and our true belief. Many animals form their true beliefs only in such simple ways. But we can also form true beliefs in subtler ways. When there are such causal correlations, there will also be related modal truths. The flash of lightning makes it certain or very likely that we shall soon hear thunder. By forming modal beliefs about what is certain or likely to be true, we may be able to form many more true empirical beliefs. In *Three Roads*, for example, we may realize that the dog's quarry must have gone down the third road. Our ancestors did much better than this intelligent dog. By forming modal beliefs, they formed advantageous true beliefs about many other ways of getting food, such as sowing seeds of rice or wheat. And we can now form true beliefs that do not help us to survive and reproduce, such as beliefs about prime numbers, atoms, and how life began.

These facts about our modal beliefs also support the view that there are modal truths. Our modal beliefs help us to form true empirical beliefs only when and because these modal beliefs are true. If the facts of which we are aware did *not* make it certain or likely that various empirical beliefs are true, that would undermine this explanation of our ability to form these true empirical beliefs. This ability would involve a massive coincidence. Though modal properties are not empirically discoverable features of the spatio-temporal world, and are therefore doubted by some Naturalists, these people should admit that, to explain our knowledge of the world, we should and can appeal to various modal truths. Similar claims apply, we can add, to our scientific theories, and our philosophical arguments about these theories. These theories and

arguments must all appeal to claims about what is certain, probable, possible, or impossible.

We can now return to our normative beliefs, and to normative properties and truths. As we shall see, several people argue that if our moral and other evaluative beliefs can be explained in evolutionary terms, this fact would undermine these beliefs, showing that we have no reason to have them. Such an explanation, Nozick writes

> threatens to bypass moral rightness or bestness completely. . . This type of explanation. . . . would also seem to show that it is unreasonable to believe that there are any such (objective) evaluative facts.

When some evolutionary account might undermine beliefs of some kind, we have two questions, which give us three possibilities:

Q1: Is it often evolutionarily advantageous to have such beliefs?

	Yes	No
Q2: Are such beliefs advantageous because they are often true?		(3)
	Yes	No
	(1)	(2)

As we have just seen, our modal beliefs are of type (1). Having such beliefs is often advantageous, so evolution can explain why we have them. These modal beliefs are advantageous only because they are often true, so this evolutionary explanation does not undermine but supports these beliefs.

Consider next beliefs of type (2), which are advantageous whether or not they are true. When evolution explains why we have such beliefs, that *may* cast doubt on these beliefs. On such evolutionary accounts, it might be claimed, we would have such beliefs even if they *weren't* true, and that may undermine, or weaken, our reasons to have these beliefs. We can call these *debunking* explanations.

Beliefs of type (3) avoid such objections. When certain beliefs are not advantageous, they cannot be challenged by appeals to evolutionary psychology. We cannot argue

that, because such beliefs would be selected by evolution, we would have them whether or not they are true.

I shall return to the question whether, if evolution could explain our moral beliefs, and our beliefs about practical reasons, that would undermine these beliefs. But we should first try to reach conclusions about normative epistemic reasons, and our normative beliefs about such reasons.

When we believe that

(1) some fact F makes some belief B certain or likely to be true,

many of us also believe that

(2) F counts in favour of our believing B.

We should ask whether evolution can explain why we have normative beliefs like (2). That depends on whether these beliefs are evolutionarily advantageous, since they helped our ancestors to survive and reproduce. Is that true?

The answer is not obvious. To form many true empirical beliefs, it is enough to realize that various facts make these beliefs certain or likely to be true. We do not need to have the further normative belief that these facts count in favour of our having these empirical beliefs. It is similarly true, however, that to form many true empirical beliefs, we don't need to have modal beliefs about whether these empirical beliefs are certain or likely to be true. In many other cases, though, it is advantageous to form true modal beliefs, since that greatly increases our ability to form true empirical beliefs. Similar claims might apply to normative epistemic beliefs, like (2). If we can form such normative beliefs, that might further increase our ability to form true empirical beliefs.

It will help to distinguish here between certainty and probability. When we realize that some belief is certain to be true, that might nearly always be enough to lead us to form this belief. In many cases, however, what we realize is that some belief is probable or likely to be true. In many such cases, this realization may *not* be enough to lead us to form this belief. We may, for example, fail to form this belief because we don't want this belief to be true, or we hope that it isn't. When we have such desires or hopes, we may be more likely to form such true beliefs if, as well as realizing that these beliefs are probably true, we also believe that this fact counts in favour of our having these beliefs, or we believe that we ought to have these beliefs. This is like the way in which our normative beliefs may help us to act in certain ways despite having conflicting desires.

Such normative epistemic beliefs may not be strongly advantageous. But evolutionary explanations do not need to appeal to great advantages. In most cases,

natural selection occurs when some animals have a slight advantage over their contemporaries, as is true of the cheetahs who can run slightly faster than others, and the giraffes who have slightly longer necks. If there were any advantages in our believing that we have strong reasons to have certain beliefs, or that we ought to have these beliefs, that would be enough to explain how evolution made us able to form such advantageous normative beliefs. We have a further reason to believe that such normative beliefs are evolutionarily advantageous. Many of us often have such beliefs. If we assume that our brains and minds are the products of evolution, the fact that we often have such beliefs gives us a further reason to expect that such beliefs are at least slightly advantageous.

For these reasons, I conclude that our normative epistemic beliefs are at least slightly advantageous, and that evolution could therefore explain how we became able to form such beliefs.

We should next ask

Q3: Are these normative beliefs advantageous because they are often true?

As several writers claim, the answer seems to be No. Suppose again that we have both the modal belief that

(1) some fact F makes some belief B certain or likely to be true,

and the normative belief that

(2) F therefore counts in favour of our believing B.

As we have seen, modal beliefs like (1) are advantageous only when they are true. If it was never true that some fact made some belief certain or likely to be true, our having such modal beliefs would not help us to form other true beliefs. No such claim applies to normative beliefs like (2). If we believe that, when fact F makes B likely to be true, this fact counts in favour of our believing B, this normative belief may make us more likely to believe B. When B is true, believing B may be advantageous. For normative beliefs like (2) to have such advantageous effects, however, these beliefs don't need to be true. A similar claim applies to some non-normative beliefs. We would be more likely to act in certain ways, for example, if we believed that our government or God would punish our failure to act in these ways. These beliefs would have these effects whether or not they were true.

Suppose that, as I have argued, evolution could explain why we have these normative epistemic beliefs. We should now ask

> Q4: Would this evolutionary explanation debunk these beliefs? Should we conclude that our normative epistemic beliefs are *merely* advantageous, since we have no reason to assume that such beliefs are often true?

In considering this question, we can first note that, though we often have normative epistemic beliefs, these are not like empirical beliefs each of which would need its own justification. It is enough to consider a small group of normative beliefs, which are all closely related to certain modal beliefs. Two examples are the beliefs that

> (3) when some belief is probably true, this fact gives us a normative reason, in the sense that it counts in favour of our having this belief,

and that

> (4) when some belief is certainly true, this fact counts *decisively* in favour of our having this belief.

If some evolutionary account could explain why we believe (3) and (4), would that undermine these beliefs?

The answer, I believe, is No. If we consider (3) and (4), most of us would believe that

> (5) it is certain or very likely that (3) and (4) are true.

(5) states the kind of modal belief that, on the evolutionary account, we are able to able to form, because such beliefs are advantageous. This evolutionary account also supports such beliefs, by implying that they are often true. We can assume, I believe, that most of our modal beliefs are true. In ordinary circumstances, most of us are fairly good at judging what is certain, likely, unlikely, or impossible. On these assumptions, we can justifiably believe that (5) is true. And the truth of (5) gives us a strong modal reason to believe that (3) and (4) are true. This reason would not be undermined by the fact that some evolutionary account could explain why we have these beliefs.

Some people would reject these claims. These people argue that, if our normative beliefs could be explained by evolution, this fact *would* undermine these beliefs. When they give such arguments, however, some of these people use misleading analogies. In what we can call

> *Joyce's imagined case*, we believe that Napoleon lost the battle of Waterloo, but we can't remember why we have this belief. We then learn that we were given and swallowed one of two belief-inducing pills. Though this pill caused us to believe that Napoleon lost this battle, the other pill would have caused us to believe that Napoleon won. We also learn that the pill we were given was picked at random.

As Joyce points out, this information would undermine our belief that Napoleon lost this battle. We would have no reason to believe either that Napoleon lost this battle, or that he won. Similar remarks would apply, Joyce claims, if we learnt that our moral beliefs could be explained by evolutionary psychology.

This argument fails, I believe, because it ignores the difference between

> empirical beliefs for which we would need some kind of evidence,

and

> beliefs that are *intrinsically* plausible.

Beliefs are in this sense plausible when it is the *content* of these beliefs that justifies our having these beliefs. Such beliefs are plausible, we can often claim, because their content makes it certain or likely that these beliefs are true. Since we don't need evidence for such beliefs, nor do we need to infer these beliefs from other true beliefs, such beliefs can be justified even when we have no reason, or no *other* reason, to have these beliefs. Two examples are the beliefs that

> (3) when some belief is probably true, this fact counts in favour of our having this belief,

and

> (4) when some belief is certainly true, this fact counts *decisively* in favour of our having this belief.

If we understand the modal concepts *probable* and *certain* and the normative concept *counts in favour*, we shall be likely to find these beliefs intrinsically plausible. The content of (3) and (4) makes these beliefs at least very likely to be true. These are the kinds of belief that Sidgwick and others call *intuitive*.

Suppose next that we learnt that

> (6) evolution could explain why we believe (3) and (4).

This would not be like learning that our belief about Napoleon was caused in the random way that Joyce imagines. Given their intrinsic plausibility, our beliefs in (3) and (4) would *not* be undermined by (6). If we learnt that (6) was true, we would have learnt that there is one kind of influence on our beliefs that would have been sufficient to cause us to believe (3) and (4) even if these claims had *not* seemed to us intrinsically plausible. But (3) and (4) would still seem to us intrinsically plausible, because they would seem certain or at least very likely to be true. This justification for these beliefs would be unaffected.

For our beliefs in (3) and (4) to be undermined, we would have to learn instead that

> (7) evolution could explain why we believe that (3) and (4) are intrinsically plausible, and are certain or very likely to be true. We would have had these beliefs even if these claims were *not* plausible, and were *not* certain or likely to be true.

We can imagine a variant of Joyce's case, in which we came to believe that (7) was true. Descartes imagined a similar case, in which an evil demon caused Descartes' reasoning abilities to go widly astray. If we learnt that we had taken some delusion-inducing pill, which made us enable to form true beliefs about which claims were certain or likely to be true, we *would* lose our reason to believe that (3) and (4) were true. But, for reasons that I give above and below, we could not defend (7).

Evolutionary Naturalists might reject what I have just claimed. Street, for example, writes:

> if we adopt the non-naturalist realist's conception of normative truth—as independent and yet lacking in causal powers—there is no reason to think that natural selection, or for that matter any other causal process, would shape us in such a way that we would be able to track such truths. . . [We have] no *reason* to think that the causal forces described by our best scientific explanations shaped our normative judgments in ways that might have led those judgments to track the truth.

For Street to defend this claim, she would need to make similar claims about our *modal* judgments or beliefs. One such belief is

> (5) It is certain or very likely that
>
> > (3) when some belief is probably true, this fact counts in favour of our having this belief,
>
> and that
>
> > (4) when some belief is certainly true, this fact counts decisively in favour of our having this belief.

If we could justifiably believe (5), we could justifiably believe (3) and (4). To defend her view, Street would therefore have to claim that

> (8) we have no reason to think that the causal forces described by our best scientific explanations shaped either our normative beliefs *or our modal beliefs* in ways that might have led these beliefs to be true.

But Street could not defend (8). First, as I have argued, evolution explains how we can form true modal beliefs, and this explanation supports the view that such beliefs are often true. Second, our scientific theories must appeal to our modal beliefs. We could not give scientific explanations without assuming that we can form true beliefs about what is certain, probable, or unlikely to be true. Third, as Street herself concedes, our conclusions about these questions must depend on which claims we find more plausible. We should compare (5) with

> (9) It is not certain nor even likely that
>
> > (3) when some belief is probably true, this fact counts in favour of our having this belief,
>
> or that
>
> > (4) when some belief is certainly true, this fact counts decisively in favour of our having this belief.
>
> We have no reason to think that these beliefs are true. We believe (3) and (4) only because such beliefs helped our ancestors to survive and reproduce.

Compared with (5), (9) is much less plausible. We do not believe (3) and (4) only because, in Joyce's phrase, such beliefs helped our ancestors to have more babies. We have these beliefs because they are intrinsically very plausible, being certain or very likely to be true.

106 Practical and Moral Truths

We can now turn to our beliefs about practical reasons, and about what we should or ought to do in the reason-implying, rational, and moral senses.

Many people argue:

> (A) Our normative beliefs were deeply influenced by evolution.
>
> (B) We have no reasons to believe that evolution would lead us to have true normative beliefs.
>
> Therefore
>
> We have no reasons to believe that our normative beliefs are true.

We can call this *the Evolutionary Debunking Argument*.

This argument, I believe, fails. We can reply:

> (C) We have some normative beliefs that are intrinsically plausible, because they are very likely to be true.
>
> (D) Even if these normative beliefs were influenced by evolution, that would not show that these beliefs are not intrinsically plausible.

Therefore

> We can justifiably assume that these beliefs are true.

Since (C) appeals to our intuitions about the plausibility of these beliefs, we can call this *the Intuitionist Reply*. Two such beliefs are

> (E) We have reasons to try to stay alive, and to promote our future well-being,

and

> (F) We have reasons to take the most effective means of achieving these and our other rational aims.

These beliefs are intrinsically very plausible. They are also evolutionarily advantageous, since people who have and act on these beliefs are more likely to survive and reproduce. Since these beliefs are advantageous, evolution might be able to explain why we have these beliefs. But that would not make these beliefs less plausible. So we can justifiably assume that (E) and (F) are true.

When Street defends the Debunking Argument, she considers something like this Intuitionist Reply. Such a reply, she writes

> assumes the very thing called into question by my argument---namely that we are not hopeless as normative judges. The reply trivially assumes that we are *correct* to think that staying alive, developing one's capacities, family and friendship, and so on, are independently worth pursuing. In so doing, it utterly fails to address the question posed by my argument: what is the relation between evolutionary influences on our normative judgments. . . and the independent normative truth . . . [when] considered without presupposing substantive views on what that independent truth is. . . When we think about *that* question, we see we have reason to be pessimistic that we have any ability to track the independent normative truth.

Street here claims that, to defend our belief that we are not 'hopeless as normative judges', we must show that evolution would be likely to lead us to have true

normative beliefs, without assuming that any particular normative beliefs are, or are likely to be, true.

What Street here requires us to do is logically impossible. Some whimsical despot might require us to show that some clock is reliable, because it always tells the correct time, without making any assumptions about what *is* the correct time. Or this despot might require us to show that someone gives true answers to certain questions, without making any assumptions about which answers are true. Though we could not meet these requirements, that would not show that this clock was unreliable, or that this person did not give true answers. Just as we couldn't possibly show that some clock tells the correct time without making any assumptions about the correct time, we couldn't possibly show that evolution tends to produce true normative beliefs, without making any assumptions about which normative beliefs are true, or are at least likely to be true. That does not help to show that, as Street claims, we may be 'hopeless as normative judges', who could never justifiably believe that our normative beliefs are true.

Other writers make similar claims. It is a deep problem, Gibbard writes

> whether we can see why, for beings like us, finding things to be of value should go with their genuinely *being* of value.

To justify our value judgments, we must try to show how we might be able to get things right. Our judgments would be *deeply vindicated*, Gibbard writes,

> if some correct non-trivial account could be given of why we aren't hopeless judges of what's to be sought in life.

This account, Gibbard claims, 'must fit our best current scientific understandings of the nature of humanity'. Since evolutionary psychology is the most relevant science, we should try to tell some story which provides 'a deep Darwinian vindication' of our ability to judge what is worth seeking. After some original and subtle attempts, Gibbard writes

> I have been baffled in seeing how such a story could be told.

Gibbard is right to be baffled, since no such story could be told. On Gibbard's assumptions, this Darwinian vindication would have to claim that

> (G) we are not hopeless judges of what's to be sought in life, because our normative judgments were all produced by evolution, and we can assume that evolution would lead us to make true judgments about these questions.

We could not hope to defend this claim. If evolution caused us to have our normative beliefs, it would cause us to have those beliefs that would enable us to have more

children, thereby spreading our genes. We cannot assume that, if our normative beliefs are in this way reproductively advantageous, that makes them likely to be true. As Street writes,

> the right conclusion seems to be that evolutionary forces pushed us in ways that simply bear no relation to the independent normative truth. . . [or] were as good as random with respect to the truth.

Though we cannot give what Gibbard calls a deep Darwinian vindication of our normative beliefs, I have suggested a less direct and partly Darwinian defence of these beliefs. Since we cannot assume that evolution would lead us to form true normative beliefs, we must describe some other way in which we could form such beliefs. Like Sidgwick and others, I believe that

> (H) we can recognize the intrinsic plausibility of many normative beliefs.

Since we cannot be causally affected by reason-implying normative properties, we must suggest how we are able assess which normative beliefs are more plausible, and more likely to be true. Evolution, I have claimed, may provide the answer. Evolution can explain how we became able to form true intuitive modal beliefs about what is certain, probable, possible, or impossible. And we can form such beliefs about which normative beliefs are certain or very likely to be true. Such beliefs are intrinsically plausible. Two examples are

> (I) when some belief is probably true, this fact counts in favour of our having this belief,

and

> (J) when some belief is certainly true, this fact counts decisively in favour of our having this belief.

We can also recognize that, compared with the belief that

> (K) what's to be sought in life is happiness, achievement, and mutual love,

it is less plausible to believe that

> (L) what's to be sought is suffering, failure, and mutual hate.

Of the people who have claimed that evolution undermines our normative beliefs, some might accept this partly Darwinian account, and conclude that we are *not* hopeless in judging what's to be sought in life.

Other people would reject these claims, since they deny that normative beliefs can be intrinsically plausible or implausible. This seems to be what Street assumes. When

she discusses the view that there are some normative truths which are not mind-dependent, Street claims that we would be most unlikely to be able to recognize such truths. On such views, she writes:

> the independent normative truth could be *anything*. For all we know as a conceptual matter. . what's ultimately worth pursuing could well be hand-clasping, or writing the number 587 over and over again, or counting blades of grass. But if there are innumerable things such that it's conceptually possible they're ultimately worth pursuing, and yet our values have been shaped from the outset by forces that are as good as random with respect to the normative truth, then what are the odds that our values will have hit, as a matter of sheer coincidence, on those things which are independently really worth pursuing?

Street here assumes that, if we ask whether what's worth pursuing is happiness, suffering, or counting blades of grass, we should regard these three answers as being equally likely to be true. She writes elsewhere:

> Either the realist is forced to embrace a skeptical conclusion---acknowledging that our normative judgments are in all likelihood hopelessly off track. . . or else the realist must hold that an astonishing coincidence took place.

But if it is true that happiness is more worth pursuing than suffering, and we believe this truth, that would not be an astonishing coincidence.

Street also claims that, since it is logically possible that there are infinitely many people who all have conflicting normative beliefs, we would be 'crazy' to think that *we* are the people whose beliefs are true. This view, she writes, would be

> a strange form of religion—a religion stripped clean of everything except the bare conviction that there are independent normative truths that one is capable of recognizing.

If Street assumed that some such normative beliefs are intrinsically plausible, and very likely to be true, she would not make these claims.

Gibbard makes similar claims. Intuitionists like Sidgwick, Gibbard writes, believe that there are

> facts of what is rational and what is not. A person of normal mental powers can discern these facts. Judgments of rationality are thus straightforward apprehensions of fact, not through sense perception but through a mental faculty analogous to sense perception.

Of these 'facts of what is rational', one example, we can claim, is

> (M) If we believe that we have decisive reasons to have some belief, or decisive reasons to act in some way, we ought rationally to have this belief, or to act in this way.

When Gibbard discusses the view that we can recognize such facts, he writes

> If this is what anyone seriously believes, then I simply want to debunk it.

What Gibbard rejects so firmly here may be only the belief that we have, in intuition, a mental faculty that is like sense perception. As I have said, that was not Sidgwick's view. If Gibbard accepted my account of how evolution might make us able to recognize facts like (M) without our having to use such a mysterious quasi-sensory faculty, he might accept this vindication of our normative beliefs.

I shall not try to defend the view that normative beliefs or claims can be intrinsically plausible or implausible. I assume that, when Street rejects this view, she is really rejecting the metaphysics or epistemology to which, she assumes, normative realists must appeal. Almost everyone finds some normative claims intrinsically plausible. Two examples are the claims that, when some belief is certain to be true, that counts in favour of our having this belief, and that compared with happiness, suffering is less worth pursuing.

Street suggests another argument, however, against this Intuitionist view. In what we can call

> *Street's imagined case*: We have various beliefs about the planet Jupiter, such as beliefs about Jupiter's size and its number of moons. To our surprise we learn that we were caused to have these beliefs by some hypnotist, who picked them at random out of a hat.

As Street points out, this information should lead us to give up these beliefs. Suppose that we are stubborn, and we claim that

> (N) since our beliefs were picked at random, and these beliefs are true, picking beliefs at random is a reliable way of forming true beliefs.

This claim would be viciously circular. We cannot defensibly argue both that our beliefs are true because our method of forming them is reliable, and that this method is reliable because our beliefs are true. Similar remarks apply, Street claims, to those who believe that they can recognize independent normative truths. In her words:

> suppose you hold certain views about how you have reason to live (and you think that there are independent truths about such matters). Then one day you learn that you acquired these normative views in large part by having had them shaped by evolutionary forces. You're concerned that this might not have been

> a reliable method for arriving at them. In answer to this concern, it is no help to repeat your views about how you have reason to live and then point out that these are the very same ones that evolutionary forces shaped you to have. No method of arriving at your views about how you have reason to live---no matter how bizarre and unreliable---could fail this test, since if having led to your actual views is the test of a method's reliability, then whatever method you actually use will come out as reliable.

This objection to Intuitionism does not, I believe, succeed. Suppose first that Street had made a stronger claim, asking us to imagine that we have learnt that

> (O) our normative beliefs are *entirely* shaped by evolutionary forces.

If (O) were true, the Evolutionary Debunking Argument would have considerable force. On this assumption, our normative beliefs would not even in part depend on our ability to recognize the intrinsic plausibility of these beliefs. Our intuitions about such plausibility would either make no difference, or would themselves be entirely shaped by evolutionary forces. We would have learnt that our normative beliefs were all formed by a single method: natural selection. On this disturbing view, we have only those beliefs that help us to survive and reproduce, thereby spreading our genes. If that were true, Street's imagined case *would* be a relevant analogy. We could not defensibly claim that

> (P) since our normative beliefs were entirely produced by evolution, and these beliefs are true, evolution produces true beliefs.

That would be viciously circular, like the claim that, since our beliefs were picked at random, but these beliefs are true, picking beliefs at random is a reliable way of forming true beliefs.

Street does not, however, claim (O), which is clearly false. Street claims that

> (Q) our normative beliefs were *partly* shaped by evolutionary forces.

If (Q) is true, we have had *two* methods of reaching our normative beliefs. At a conscious level, when we are engaged in normative thinking, we are led to form those beliefs which are intuitively most plausible, and seem to us most likely to be true. Unconsciously, however, we have also been influenced by evolutionary forces. The Intuitive method, we can plausibly believe, leads us towards the normative truth. The evolutionary forces, as Street claims, are random with respect with truth. On these assumptions, we must ask which of these influences on our beliefs has proved more effective. Of these truth-directed and truth-ignoring methods of forming normative beliefs, which has prevailed?

We ought, I have claimed, to reject the Debunking Argument as stated above. But this argument might now take another, better form. It might be claimed that

> (R) though we can sometimes make such intuitive judgments, most of our normative beliefs have been so strongly influenced by evolution that it would be most unlikely if these beliefs were true.

This *Second* Debunking Argument might, I agree, succeed.

When Street argues that evolution has had 'a tremendous influence' on our normative beliefs, she gives examples of many beliefs that are

> (1) reproductively advantageous, widely held, and plausible.

Two examples are the beliefs that we have reasons to promote our own future well-being, and the well-being of our children. Such beliefs provide weak evidence, however. These beliefs might be widely held either because they are plausible, or because they are advantageous, or both. It would be hard to choose between these explanations.

Better evidence would be beliefs that are

> (2) reproductively advantageous, widely held, and implausible.

Suppose it was widely believed that men ought to commit rape or adultery as often as they can, thereby spreading their genes. We would have good reason to believe that *these* beliefs were produced by evolution. Darwin gives another example. If humans had evolved to be more similar to bees, Darwin writes,

> unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters, and no one would think of interfering.

But these beliefs are not widely held, so they do not support (R). For actual examples of type (2), we might suggest the beliefs that homosexuality and birth control are wrong. But when we ask why these beliefs have been widely held, the explanations do not, I believe, give strong support to (R).

Street suggests an example of type (2), which is the widely held belief that we should give priority to the well-being of people who are members of our group, such as people to whom we are genetically related, or members of the same community, or those who are most likely to reciprocate. This belief is reproductively advantageous, Street writes, but 'many of us are coming to think that [this belief] is not true.' But this belief is not of type (2) since it is not implausible.

Consider next beliefs that are

> (3) reproductively disadvantageous, widely held, and plausible.

Two examples are the beliefs that it is wrong for men to rape women, and that we have no duty to have children. Since these beliefs are reproductively disadvantageous, their being widely held counts against (R).

We can next note that, when we consider how people's moral beliefs have changed over many centuries, we find slow but accelerating progress towards the beliefs that everyone's well-being matters equally, and that everyone has equal rights. Though these beliefs may not be strongly disadvantageous, they are clearly *not* the product of evolution. This fact suggests an alternative to (R). We can claim, I believe, that

> (S) though mankind's earliest moral beliefs were in several ways distorted by the influence of evolution, those distortions are being overcome, so that true moral beliefs are becoming more and more widely held.

These few remarks do not refute (R) or establish (S). But there is, I believe, little evidence that our normative beliefs have been massively influenced by evolution, in ways that undermine our reasons to have these beliefs, and to assume that they are often true.

I shall now summarize some of these claims.

Evolution can explain how we became able to form true modal beliefs, about what is certain, likely, or unlikely to be true. This explanation itself gives us further reason to believe that our modal beliefs are often true.

We have such modal beliefs about which normative beliefs or claims are certain or likely to be true. One example is the modal belief that

> (T) it is certain or very likely that
>
> > (U) when some belief is probably true, this fact gives us a normative epistemic reason, by counting in favour of our having this belief.

This evolutionary account does not itself support (T) and (U), since these beliefs would be reproductively advantageous whether or not they are true. But this evolutionary account also gives us no reasons to doubt that these beliefs are true. Since we have reasons to believe that we can form true modal beliefs, we can justifiably believe that (T) is at least very likely true, and that (U) is therefore very likely to be true. Since (U) is a normative belief, (U) is one example of a normative belief that we can justifiably

claim to be true, and evolution itself explains how we can form such true normative beliefs.

We also have modal beliefs when we compare various claims about practical reasons and what we ought to do. Two examples are the beliefs that

(V) it is very likely to be true that

> (W) when some act is our only way to avoid great pain, this fact gives us a reason to act in this way,

and that

> (X) when we believe that we have decisive reasons to act in some way, we ought rationally to act in this way.

As before, since beliefs (W) and (X) would be advantageous whether or not they are true, this evolutionary account does not itself support these beliefs. But this account also gives us no reason to doubt that these beliefs are true. Since we can form true modal beliefs, we can justifiably believe that (V) is true, and that (W) and (X) are therefore very likely to be true. Similar claims apply to other beliefs about practical reasons, and about what we ought to do.

This vindication of these normative beliefs is less Darwinian than the vindication that Gibbard tried to give. Though evolution can explain how we can form true modal and normative beliefs, evolution does not explain the content of these beliefs, or the normativity of epistemic and practical reasons. As Nagel claims, evolution cannot create the possible forms that living beings can take, since evolution can only select between these possibilities. But if there can be animals who are rational, and can respond to reasons, evolution can explain how this possibility became actual. Such evolutionary accounts can also answer the epistemic objections to Non-Metaphysical Non-Naturalist Cognitivism. Just as evolution explains how we can form true arithmetical beliefs without being causally affected by numbers or their properties, evolution can explain how, without being causally affected by irreducibly normative properties, we can form true beliefs about epistemic and practical reasons, and about what we ought to believe, and want, and do.

107 On What Matters

To be written. The conclusion might be:

Williams writes:

> It is not a trivial question, Socrates said: what we are talking about is how one should live. Or so Plato reports him, in one of the first books written about this subject. Plato thought that philosophy could answer the question. . . The aims of moral philosophy, and any hopes it may have of being worth serious attention, are bound up with the fate of Socrates' question.

Plato believed that there might be truths about how we should live, and that we should try to reach these truths, by considering various arguments about this question. Other writers, such as Sidgwick and Nagel, make similar claims. Commenting on such claims, Williams remarks:

> the writer's note of urgency suggests. . . that what will happen could turn on the outcome of these arguments, that the justification of the ethical life could be a *force*. If we are to take this seriously, then it is a real question, who is supposed to be listening. Why are they supposed to be listening? What will the professor's justification do, when they break down the door, smash his spectacles, take him away?

Williams may here be intending only to remind us of what he calls the 'glistening contempt for philosophy' that is shown by some of the characters in Plato's dialogues. But Williams himself believes, though less contemptuously, that philosophy cannot help us to decide how to live. On his view, we could not be helped by reaching truths about we what have decisive reasons, or most reason, to want and to do.

Only such truths, I believe, could help us to decide how to live. And if there were no such truths, we would have no reason to try to decide how to live. As Gibbard fears, these decisions would be arbitrary, since there would not be any better or worse ways to live. We would not be the animals that can understand and respond to reasons. In a world without reasons, we would act only on our instincts and desires, living as other animals live. The Universe could not contain rational beings.

That is not, I believe, the fate of Socrates' question. Some things matter, and there are better and worse ways to live. After thousands or millions of years of responding to reasons in ways that helped them to survive and reproduce, human beings can now respond to other reasons. We are a part of the Universe that is starting to understand itself. And we can partly understand, not only what is in fact true, but also what ought to be true, and what we might be able to make true. We can have hopes for what Kant calls 'the noblest ends of mankind in all future ages'. And we can each brighten the lives of one or two people, a little.

PART SEVEN APPENDICES

APPENDIX A 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?

These questions, some believe, may have causal answers. Suppose first that the Universe has always existed. Some believe that, if all events were caused by earlier events, everything would be explained. That, however, is not so. Even an infinite series of events cannot explain itself. We could ask why this series occurred, rather than some other series, or no series. Of the supporters of the Steady State Theory, some welcomed what they took to be this theory's atheistic implications. They assumed that, if the Universe had no beginning, there would be nothing for a Creator to explain. But there would still be an eternal Universe to explain.

Suppose next that the Universe is not eternal, since nothing preceded the Big Bang. That first event, some physicists suggest, may have obeyed the laws of quantum mechanics, by being a random fluctuation in a vacuum. This would causally explain, they say, how the Universe came into existence out of nothing. But what physicists call a vacuum isn't really nothing. We can ask why it exists, and has the potentialities it does. In Hawking's phrase, 'What breathes fire into the equations?'

Similar remarks apply to all suggestions of these kinds. There could not be a causal explanation of why the Universe exists, why there are any laws of nature, or why these laws are as they are. Nor would it make a difference if there is a God, who caused the rest of the Universe to exist. There could not be a causal explanation of why God exists.

Many people have assumed that, since these questions cannot have causal answers, they cannot have any answers. Some therefore dismiss these questions, taking them to be not worth considering. Others conclude that

they do not make sense, assuming that, as Wittgenstein wrote, 'doubt can exist only where there is a question; and a question only where there is an answer'.

These assumptions are, I believe, mistaken. Even if these questions could not have answers, they would still make sense, and they would still be worth considering. Such thoughts take us into the aesthetic category of the *sublime*, which applies to the highest mountains, raging oceans, the night sky, the interiors of some cathedrals, and other things that are superhuman, awesome, limitless. No question is more sublime than why there is a Universe: why there is anything rather than nothing. Nor should we assume that answers to this question must be causal. And, even if reality cannot be fully explained, we may still make progress, since what is inexplicable may become less baffling than it now seems.

1

One apparent fact about reality has recently been much discussed. Many physicists believe that, for life to be possible, various features of the Universe must be almost precisely as they are. As one example of such a feature, we can take the initial conditions in the Big Bang. If these conditions had been more than very slightly different, these physicists claim, the Universe would not have had the complexity that allows living beings to exist. Why were these conditions so precisely right? [819]

Some say: 'If they had not been right, we couldn't even ask this question.' But that is no answer. It could be baffling how we survived some crash even though, if we hadn't, we could not be baffled.

Others say: 'There had to be some initial conditions, and the conditions that make life possible were as likely as any others. So there is nothing to be explained.' To see what is wrong with this reply, we must distinguish two kinds of case. Suppose first that, when some radio telescope is aimed at most points in space, it records a random sequence of incoming waves. There might be nothing here that needed to be explained. Suppose next that, when the telescope is aimed in one direction, it records a sequence of waves whose pulses match the number π, in binary notation, to the first ten thousand digits. That particular number is, in one sense, just as likely as any other. But there *would* be something here that needed to be explained. Though each long number is unique, only a very few are, like π, mathematically special. What would need to be explained is why this sequence of waves exactly matched such a special number. Though this matching might be a coincidence, which had been randomly produced, that would be most unlikely. We could be

almost certain that these waves had been produced by some kind of intelligence.

On the view that we are now considering, since any sequence of waves is as likely as any other, there would be nothing to be explained. If we accepted this view, intelligent beings elsewhere in space would not be able to communicate with us, since we would ignore their messages. Nor could God reveal himself. Suppose that, with some optical telescope, we saw a distant pattern of stars which spelled out in Hebrew script the first chapter of Genesis. This pattern of stars, according to this view, would not need to be explained. That is clearly false.

Here is another analogy. Suppose first that, of a thousand people facing death, only one can be rescued. If there is a lottery to pick this one survivor, and I win, I would be very lucky. But there might be nothing here that needed to be explained. Someone had to win, and why not me? Consider next another lottery. Unless my gaoler picks the longest of a thousand straws, I shall be shot. If my gaoler picks that longest straw, there would be something to be explained. It would not be enough to say, 'This result was as likely as any other.' In the first lottery, nothing special happened: whatever the result, someone's life would be saved. In this second lottery, the result *was* special, since, of the thousand possible results, only one would save a life. Why was this special result *also* what happened? Though this might be a coincidence, the chance of that is only one in a thousand. I could be almost certain that, like Dostoyevsky's mock execution, this lottery was rigged.

The Big Bang, it seems, was like this second lottery. For life to be possible, the initial conditions had to be selected with great accuracy. This *appearance of fine-tuning*, as some call it, also needs to be explained.

It may be objected that, in regarding conditions as special if they allow for life, we unjustifiably assume our own importance. But life *is* special, if only because of its complexity. An earthworm's brain is more complicated than a lifeless galaxy. Nor is it only life that requires this fine-tuning. If the Big Bang's initial conditions had not been almost precisely as they were, the Universe would have either almost instantly recollapsed, or expanded so fast, and with particles so thinly spread, that not even stars or heavy elements could have formed. That is enough to make these conditions very special.

It may next be objected that these conditions cannot be claimed to be improbable, since such a claim requires a statistical basis, and there is only one Universe. If we were considering all conceivable Universes, it would indeed be implausible to make judgments of statistical probability. But our question is much narrower. We are asking what would have happened if, with the

same laws of nature, the initial conditions had been different. That provides the basis for a statistical judgment. There is a range of values that these conditions might have had, and physicists can work out in what proportion of this range the resulting Universe could have contained stars, heavy elements, and life.

This proportion, it is claimed, is extremely small. Of the range of possible initial conditions, fewer than one in a billion billion would have produced a Universe with the complexity that allows for life. If this claim is true, as I shall here assume, there is something that cries out to be explained. Why was one of this tiny set *also* the one that actually obtained?

On one view, this was a mere coincidence. That is conceivable, since coincidences happen. But this view is hard to believe since, if it were true, the chance of this coincidence occurring would be below one in a billion billion.

Others say: 'The Big Bang *was* fine-tuned. In creating the Universe, God chose to make life possible.' Atheists may reject this answer, thinking it improbable that God exists. But this is not as improbable as the view that would require so great a coincidence. So even atheists should admit that, of these two answers to our question, the answer that invokes God is more likely to be true.

This reasoning revives one of the traditional arguments for belief in God. In its strongest form, this argument appealed to the many features of animals, such as eyes or wings, that look as if they have been designed. Paley's appeal to such features much impressed Darwin when he was young. Darwin later undermined this form of the argument, since evolution can explain this appearance of design. But evolution cannot explain the appearance of fine-tuning in the Big Bang.

This argument's appeal to probabilities can be challenged in a different way. In claiming it to be most improbable that this fine-tuning was a coincidence, the argument assumes that, of the possible initial conditions in the Big Bang, each was equally likely to obtain. That assumption may be mistaken. The conditions that allow for complexity and life may have been, compared with all the others, much more likely to obtain. Perhaps they were even certain to obtain.

To answer this objection, we must broaden this argument's conclusion. If these life-allowing conditions were either very likely or certain to obtain, then---as the argument claims---it would be no coincidence that the Universe allows for complexity and life. But this fine-tuning might have been the work, not of some existing being, but of some impersonal force, or fundamental law. That is what some theists believe God to be.

A stronger challenge to this argument comes from a different way to explain the appearance of fine-tuning. Consider first a similar question. For life to be possible on the Earth, many of the Earth's features have to be close to being as they are. The Earth's having such features, it might be claimed, is unlikely to be a coincidence, and should therefore be regarded as God's work. But such an argument would be weak. The Universe, we can reasonably believe, contains many planets, with varying conditions. We should expect that, on a few of these planets, conditions would be just right for life. Nor is it surprising that we live on one of these few.

Things are different, we may assume, with the appearance of fine-tuning in the Big Bang. While there are likely to be many other planets, there is only one Universe. But this difference may be less than it seems. Some physicists suggest that the observable Universe is only one out of many different worlds, which are all equally parts of reality. According to one such view, the other worlds are related to ours in a way that solves some of the mysteries of quantum physics. On the different and simpler view that is relevant here, the other worlds have the same laws of nature as our world, and they are produced by Big Bangs that are broadly similar, except in having different initial conditions.

On this *Many Worlds Hypothesis,* there is no need for fine-tuning. If there were enough Big Bangs, we should expect that, in a few of these, conditions would be just right to allow for complexity and life; and it would be no surprise that our Big Bang was one of these few. To illustrate this point, we can revise my second lottery. Suppose my gaoler picks a straw, not once, but very many times. That would explain his managing, once, to pick the longest straw, without that's being an extreme coincidence, or this lottery's being rigged.

On most versions of the Many Worlds Hypothesis, these many worlds are not, except through their origins, causally related. Some object that, since our world could not be causally affected by such other worlds, we can have no evidence for their existence, and can therefore have no reason to believe in them. But we do have such a reason, since their existence would explain an otherwise puzzling feature of our world: the appearance of fine-tuning.

Of these two ways to explain this appearance, which is better? Compared with belief in God, the Many Worlds Hypothesis is more cautious, since its claim is merely that there is more of the kind of reality that we can observe around us. But God's existence has been claimed to be intrinsically more probable. According to most theists, God is a being who is omnipotent, omniscient, and wholly good. The uncaused existence of such a being has been claimed to be simpler, and less arbitrary, than the uncaused existence of

many highly complicated worlds. And simpler hypotheses, many scientists assume, are more likely to be true.

If such a God exists, however, other features of our world become hard to explain. It may not be surprising that God chose to make life possible. But the laws of nature could have been different, so there are many possible worlds that would have contained life. It is hard to understand why, out of all these possibilities, God chose to create our world. What is most baffling is the problem of evil. There appears to be suffering which any good person, knowing the truth, would have prevented if he could. If there is such suffering, there cannot be a God who is omnipotent, omniscient, and wholly good.

To this problem, theists have proposed several solutions. Some suggest that God is not omnipotent, or not wholly good. Others suggest that undeserved suffering is not, as it seems, bad, or that God could not prevent such suffering without making the Universe, as a whole, less good.

We must ignore these suggestions here, since we have larger questions to consider. I began by asking why things are as they are. Before returning to that question, we should ask *how* things are. There is much about our world that we have not discovered. And, just as there may be other worlds that are like ours, there may be worlds that are very different.

2

It will help to distinguish two kinds of possibilities. *Cosmic* possibilities cover everything that ever exists, and are the different ways that the whole of reality might be. Only one such possibility can be actual, or be the one that *obtains*. *Local* possibilities are the different ways that some part of reality, or *local world*, might be. If some local world exists, that leaves it open whether other worlds exist.

One cosmic possibility is, roughly, that *every* possible local world exists. This we can call the *All Worlds Hypothesis*. Another possibility, which might have obtained, is that nothing ever exists. This we can call the *Null Possibility*. In each of the remaining possibilities, the number of local worlds that exist is between none and all. There are countless of these possibilities, since there are countless combinations of possible local worlds.

Of these different cosmic possibilities, one must obtain, and only one can obtain. So we have two questions: Which obtains, and Why?

These questions are connected. If some possibility would be easier to explain, we may have more reason to believe that this possibility obtains. This is how, rather than believing in only one Big Bang, we have more reason to believe in many. Whether we believe in one or many, we have the question why any Big Bang has occurred. Though this question is hard, the occurrence of many Big Bangs is not more puzzling than the occurrence of only one. Most kinds of thing, or event, have many instances. We also have the question why, in the Big Bang that produced our world, the initial conditions allowed for complexity and life. If there has been only one Big Bang, this fact is also hard to explain, since it is most unlikely that these conditions merely happened to be right. If instead there have been many Big Bangs, this fact is easy to explain, since it is like the fact that, among countless planets, there are some whose conditions allow for life. Since belief in many Big Bangs leaves less that is unexplained, it is the better view.

If some cosmic possibilities would be less puzzling than others, because their obtaining would leave less to be explained, is there some possibility whose obtaining would be in no way puzzling?

Consider first the Null Possibility, in which nothing ever exists. To imagine this possibility, it may help to suppose first that all that ever existed was a single atom. We then imagine that even this atom never existed.

Some have claimed that, if there had never been anything, there wouldn't have been anything to be explained. But that is not so. When we imagine how things would have been if nothing had ever existed, what we should imagine away are such things as living beings, stars, and atoms. There would still have been various truths, such as the truth that there were no stars or atoms, or that 9 is divisible by 3. We can ask why these things would have been true. And such questions may have answers. Thus we can explain why, even if nothing had ever existed, 9 would still have been divisible by 3. There is no conceivable alternative. And we can explain why there would have been no such things as immaterial matter, or spherical cubes. Such things are logically impossible. But why would *nothing* have existed? Why would there have been no stars or atoms, no philosophers or bluebell woods?

We should not claim that, if nothing had ever existed, there would have been nothing to be explained. But we can claim something less. Of all the cosmic possibilities, the Null Possibility would have needed the least explanation. As Leibniz pointed out, it is much the simplest, and the least arbitrary. And it is the easiest to understand. It can seem mysterious, for example, how things could exist without their existence having some cause, but there cannot be a causal explanation of why the whole Universe, or God, exists. The Null

Possibility raises no such problem. If nothing had ever existed, that state of affairs would not have needed to be caused.

Reality, however, does not take its least puzzling form. In some way or other, a Universe has managed to exist. That is what can take one's breath away. As Wittgenstein wrote, 'not how the world is, is the mystical, but *that* it is.' Or, in the words of a thinker as unmystical as Jack Smart: 'That anything should exist at all does seem to me a matter for the deepest awe.'

Consider next the All Worlds Hypothesis, on which every possible local world exists. Unlike the Null Possibility, this may be how things are. And it may be the next least puzzling possibility. This hypothesis is not the same as---though it includes---the Many Worlds Hypothesis. On that more cautious view, the many other worlds have the same elements as our world, and the same fundamental laws, and differ only in such features as their constants and initial conditions. The All Worlds Hypothesis covers every conceivable kind of world, and most of these other worlds would have very different elements and laws.

If all these worlds exist, we can ask why they do. But, compared with most other cosmic possibilities, the All Worlds Hypothesis may leave less that is unexplained. For example, whatever the number of possible worlds that exist, we have the question, 'Why *that* number?' That question would have been least puzzling if the number that existed were *none*, and the next least arbitrary possibility seems to be that *all* these worlds exist. With every other cosmic possibility, we have a further question. If ours is the only world, we can ask: 'Out of all the possible local worlds, why is *this* the one that exists?' On any version of the Many Worlds Hypothesis, we have a similar question: 'Why do just *these* worlds exist, with *these* elements and laws?' But, if *all* these worlds exist, there is no such further question.

It may be objected that, even if all possible local worlds exist, that does not explain why our world is as it is. But that is a mistake. If all these worlds exist, each world is as it is in the way in which each number is as it is. We cannot sensibly ask why 9 is 9. Nor should we ask why our world is the one it is: why it is *this* world. That would be like asking, 'Why are *we* who we are?', or 'Why is it *now* the time that it is?' Those, on reflection, are not good questions.

Though the All Worlds Hypothesis avoids certain questions, it is not as simple, or unarbitrary, as the Null Possibility. There may be no sharp distinction between worlds that are and are not possible. It is unclear what counts as a

kind of world. And, if there are infinitely many kinds, there is a choice between different kinds of infinity.

Whichever cosmic possibility obtains, we can ask why it obtains. All that I have claimed so far is that, with some possibilities, this question would be less puzzling. Let us now ask: Could this question have an answer? Might there be a theory that leaves nothing unexplained?

3

It is sometimes claimed that God, or the Universe, make themselves exist. But this cannot be true, since these entities cannot do anything unless they exist.

On a more intelligible view, it is logically necessary that God, or the Universe, exist, since the claim that they might not have existed leads to a contradiction. On such a view, though it may seem conceivable that there might never have been anything, that is not really logically possible. Some people even claim that there may be only one coherent cosmic possibility. Thus Einstein suggested that, if God created our world, he might have had no choice about which world to create. If such a view were true, everything might be explained. Reality might be the way it is because there was no conceivable alternative. But, for reasons that have been often given, we can reject such views.

Consider next a quite different view. According to Plato, Plotinus and others, the Universe exists because its existence is good. Even if we are confident that we should reject this view, it is worth asking whether it makes sense. If it does, that may suggest other possibilities.

This *Axiarchic View* can take a theistic form. It can claim that God exists because his existence is good, and that the rest of the Universe exists because God caused it to exist. But in that explanation God, *qua* Creator, is redundant. If God can exist because his existence is good, so can the whole Universe. This may be why some theists reject the Axiarchic View, and insist that God's existence is a brute fact, with no explanation.

In its simplest form, this view makes three claims:

(1) It would be best if reality were a certain way.

(2) Reality is that way.

(3) (1) explains (2).

(1) is an ordinary evaluative claim, like the claim that it would be better if there was less suffering. The Axiarchic View assumes, I believe rightly, that such claims can be in a strong sense true. (2) is an ordinary empirical or scientific claim, though of a sweeping kind. What is distinctive in this view is claim (3), according to which (1) explains (2).

Can we understand this third claim? To focus on this question, we should briefly ignore the world's evils, and suspend our other doubts about claims (1) and (2). We should suppose that, as Leibniz claimed, the best possible Universe exists. Would it then make sense to claim that this Universe exists *because* it is the best?

That use of 'because', Axiarchists should admit, cannot be easily explained. But even ordinary causation is mysterious. At the most fundamental level, we have no idea why some events cause others; and it is hard to explain what causation is. There are, moreover, non-causal senses of 'because' and 'why', as in the claim that God exists because his existence is logically necessary. We can understand that claim, even if we think it false. The Axiarchic View is harder to understand. But that is not surprising. If there is some explanation of the whole of reality, we should not expect this explanation to fit neatly into some familiar category. This extra-ordinary question may have an extra-ordinary answer. We should reject suggested answers which make no sense; but we should also try to see what might make sense.

Axiarchy might be expressed as follows. We are now supposing that, of all the countless ways that the whole of reality might be, one is both the very best, and is the way that reality is. On the Axiarchic View, *that is no coincidence*. This claim, I believe, makes sense. And, if it were no coincidence that the best way for reality to be is *also* the way that reality is, that might support the further claim that this was *why* reality was this way.

This view has one advantage over the more familiar theistic view. An appeal to God cannot explain why the Universe exists, since God would himself be part of the Universe, or one of the things that exist. Some theists argue that, since nothing can exist without some cause, God, who is the First Cause, must exist. As Schopenhauer objected, this argument's premise is not like some cab-driver whom theists are free to dismiss once they have reached their destination. The Axiarchic View appeals, not to an existing entity, but to an explanatory law. Since such a law would not itself be part of the Universe, it might explain why the Universe exists, and is as good as it could be. If such a law governed reality, we could still ask why it did, or why the Axiarchic View was true. But, in discovering this law, we would have made some progress.

It is hard, however, to believe the Axiarchic View. If, as it seems, there is much pointless suffering, our world cannot be part of the best possible Universe.

4

Some Axiarchists claim that, if we reject their view, we must regard our world's existence as a brute fact, since no other explanation could make sense. But that, I believe, is not so. If we abstract from the optimism of the Axiarchic View, its claims are these:

> Of the countless cosmic possibilities, one both has some very special feature, and is the possibility that obtains. That is no coincidence. This possibility obtains because it has this feature.

Other views can make such claims. This special feature need not be that of being best. Thus, on the All Worlds Hypothesis, reality is *maximal*, or as full as it could be. Similarly, if nothing had ever existed, reality would have been *minimal*, or as empty as it could be. If the possibility that obtained were either maximal, or minimal, that fact, we might claim, would be most unlikely to be a coincidence. And that might support the further claim that this possibility's having this feature would be *why* it obtained.

Let us now look more closely at that last step. When it is no coincidence that two things are both true, there is something that explains why, given the truth of one, the other is also true. The truth of either might make the other true. Or both might be explained by some third truth, as when two facts are the joint effects of a common cause.

Suppose next that, of the cosmic possibilities, one is both very special and is the one that obtains. If that is no coincidence, what might explain why these things are both true? On the reasoning that we are now considering, the first truth explains the second, since this possibility obtains because it has this special feature. Given the kind of truths these are, such an explanation could not go the other way. This possibility could not have this feature because it obtains. If some possibility has some feature, it could not fail to have this feature, so it would have this feature whether or not it obtains. The All Worlds Hypothesis, for example, could not fail to describe the fullest way for reality to be.

While it is necessary that our imagined possibility has its special feature, it is not necessary that this possibility obtains. This difference, I believe, justifies

the reasoning that we are now considering. Since this possibility must have this feature, but might not have obtained, it cannot have this feature because it obtains, nor could some third truth explain why it both has this feature and obtains. So, if these facts are no coincidence, this possibility must obtain *because* it has this feature.

When some possibility obtains because it has some feature, its having this feature may be why some agent, or some process of natural selection, made it obtain. These we can call the *intentional* and *evolutionary* ways in which some feature of some possibility may explain why it obtains.

Our world, theists claim, can be explained in the first of these ways. If reality were as good as it could be, it would indeed make sense to claim that this was partly God's work. But, since God's own existence could not be God's work, there could be no intentional explanation of why the whole of reality was as good as it could be. So we could reasonably conclude that this way's being the best explained *directly* why reality was this way. Even if God exists, the intentional explanation could not compete with the different and bolder explanation offered by the Axiarchic View.

Return now to other explanations of this kind. Consider first the Null Possibility. This, we know, does not obtain; but, since we are asking what makes sense, that does not matter. If there had never been anything, would that have had to be a brute fact, which had no explanation? The answer, I suggest, is No. It might have been no coincidence that, of all the countless cosmic possibilities, what obtained was the simplest, and least arbitrary, and the only possibility in which nothing ever exists. And, if these facts had been no coincidence, this possibility would have obtained because--or partly because---it had one or more of these special features. This explanation, moreover, could not have taken an intentional or evolutionary form. If nothing had ever existed, there could not have been some agent, or process of selection, who or which made this possibility obtain. Its being the simplest or least arbitrary possibility would have been, directly, why it obtained.

Consider next the All Worlds Hypothesis, which may obtain. If reality is as full as it could be, is that a coincidence? Does it merely happen to be true that, of all the cosmic possibilities, the one that obtains is at this extreme? As before, that is conceivable, but this coincidence would be too great to be credible. We can reasonably assume that, if this possibility obtains, that is because it is maximal, or at this extreme. On this *Maximalist View*, it is a fundamental truth that being possible, and part of the fullest way that reality could be, is sufficient for being actual. That is the highest law governing reality. As before, if such a law governed reality, we could still ask *why* it did. But, in discovering this law, we would have made some progress.

Here is another special feature. Perhaps reality is the way it is because its fundamental laws are, on some criterion, as mathematically beautiful as they could be. That is what some physicists are inclined to believe.

As these remarks suggest, there is no clear boundary here between philosophy and science. If there is such a highest law governing reality, this law is of the same kind as those that physicists are trying to discover. When we appeal to natural laws to explain some features of reality, such as the relations between light, gravity, space, and time, we are not giving causal explanations, since we are not claiming that one part of reality caused another part to be some way. What such laws explain, or partly explain, are the deeper facts about reality that causal explanations take for granted.

There would be a highest law, of the kind that I have sketched, if some cosmic possibility obtained because it had some special feature. This feature we can call the *Selector*. If there is more than one such feature, they are all partial Selectors. Just as there are various cosmic possibilities, there are various *explanatory* possibilities. For each of these special features, there is the explanatory possibility that this feature is the Selector, or is one of the Selectors. Reality would then be the way it is because, or partly because, this way had this feature.

There is one other explanatory possibility: that there is *no* Selector. If that is true, it is random that reality is as it is. Events may be in one sense random, even though they are causally inevitable. That is how it is random whether a meteorite strikes the land or the sea. Events are random in a stronger sense if they have no cause. That is what most physicists believe about some features of events involving sub-atomic particles. If it is random what reality is like, the Universe not only has no cause. It has no explanation of any kind. This claim we can call the *Brute Fact View*.

Few features can be plausibly regarded as possible Selectors. Though plausibility is a matter of degree, there is a natural threshold to which we can appeal. If we suppose that reality has some special feature, we can ask which of two beliefs would be more credible: that reality merely happens to have this feature, or that reality is the way it is because this way has this feature. If the second would be more credible, this feature can be called a *credible Selector*. Return for example to the question of how many possible local worlds exist. Of the different answers to this question, *all* and *none* give us, I have claimed, credible Selectors. If either all or no worlds existed, that would be unlikely to be a coincidence. But suppose that 58 worlds existed. This number has some special features, such as being the smallest number that is the sum of seven different primes. It may be just conceivable that this would be why 58 worlds

existed; but it would be more reasonable to believe that the number that existed merely happened to be 58.

There are, I have claimed, some credible Selectors. Reality might be some way because that way is the best, or the simplest, or the least arbitrary, or because its obtaining makes reality as full and varied as it could be, or because its fundamental laws are, in some way, as elegant as they could be. Presumably there are other such features, which I have overlooked.

In claiming that there are credible Selectors, I am assuming that some cosmic and explanatory possibilities are more probable than others. That assumption may be questioned. Judgments of probability, it may again be claimed, must be grounded on facts about our world, so such judgments cannot be applied either to how the whole of reality might be, or to how reality might be explained.

This objection is, I believe, unsound. When we choose between scientific theories, our judgments of their probability cannot rest only on predictions based on established facts and laws. We need such judgments in trying to decide what these facts and laws are. And we can justifiably make such judgments when considering different ways in which the whole of reality may be, or might have have been. Compare two such cosmic possibilities. In the first, there is a lifeless Universe consisting only of some spherical iron stars, whose relative motion is as it would be in our world. In the second, things are the same, except that the stars move together in the patterns of a minuet, and they are shaped like either Queen Victoria or Cary Grant. We would be right to claim that, of these two possibilities, the first is more likely to obtain.

In making that claim, we would not mean that it is more likely *that* the first possibility obtains. Since this possibility is the existence of a lifeless Universe, we know that it does not obtain. We would be claiming that this possibility is intrinsically more likely, or that, to put it roughly, it had a greater chance of being how reality is. If some possibility is more likely to obtain, that will often make it more likely that it obtains; but though one kind of likelihood supports the other, they are quite different.

Another objection may again seem relevant here. Of the countless cosmic possibilities, a few have special features, which I have called credible Selectors. If such a possibility obtains, we have, I have claimed, a choice of two conclusions. Either reality, by an extreme coincidence, merely happens to have this feature, or---more plausibly---this feature is one of the Selectors. It may be objected that, when I talk of an extreme coincidence, I must be assuming that these cosmic possibilities are all equally likely to obtain. But I

have now rejected that assumption. And, if these possibilities are *not* equally likely, my reasoning may seem to be undermined.

As before, that is not so. Suppose that, of the cosmic possibilities, those that have these special features are much more likely to obtain. As this objection rightly claims, it would not then be amazing if such a possibility merely happened to obtain. But that does not undermine my reasoning, since it is another way of stating my conclusion. It is another way of saying that these features are Selectors.

These remarks do show, however, that we should distinguish two ways in which some feature may be a Selector. *Probabilistic* Selectors make some cosmic possibility more likely to obtain, but leave it open whether it does obtain. On any plausible view, there are some Selectors of this kind, since some ways for reality to be are intrinsically more likely than some others. Thus of our two imagined Universes, the one consisting of spherical stars is intrinsically more likely than the one with stars that are shaped like Queen Victoria or Cary Grant. Besides Probabilistic Selectors, there may also be one or more *Effective* Selectors. If some possibility has a certain feature, this may make this possibility, not merely intrinsically more likely, but the one that obtains. Thus, if simplicity had been the Effective Selector, that would have made it true that nothing ever existed. And, if maximality is the Effective Selector, as it may be, that is what makes reality as full as it could be. When I talk of Selectors, these are the kind I mean.

5

There are, then, various cosmic and explanatory possibilities. In trying to decide which of these obtain, we can in part appeal to facts about our world. Thus, from the mere fact that our world exists, we can deduce that the Null Possibility does not obtain. And, since our world seems to contain pointless evils, we have reason to reject the Axiarchic View.

Consider next the Brute Fact View, on which reality merely happens to be as it is. No facts about our world could refute this view. But some facts would make it less likely that this view is true. If reality is randomly selected, what we should expect to exist are many varied worlds, none of which had features that, in the range of possibilities, were at one extreme. That is what we should expect because, in much the largest set of cosmic possibilities, that would be what exists. If our world has very special features, that would count against the Brute Fact View.

Return now to the question whether God exists. Compared with the uncaused existence of one or many complicated worlds, the hypothesis that God exists has been claimed to be simpler, and less arbitrary, and thus more likely to be true. But this hypothesis is not simpler than the Brute Fact View. And, if it is random which cosmic possibility obtains, we should not expect the one that obtains to be as simple, and unarbitrary, as God's existence is claimed to be. Rather, as I have just said, we should expect there to be many worlds, none of which had very special features. Ours may be the kind of world that, on the Brute Fact View, we should expect to observe.

Similar remarks apply to the All Worlds Hypothesis. Few facts about our world could refute this view; but, if all possible local worlds exist, the likely character of our world is much the same as on the Brute Fact View. That claim may seem surprising, given the difference between these two views. One view is about *which* cosmic possibility obtains, the other is about *why* the one that obtains obtains. And these views conflict, since, if we knew that either view was true, we would have strong reason not to believe the other. If all possible worlds exist, that is unlikely to be a brute fact. But, in their different ways, these views are both *non-selective*. On neither view do certain worlds exist *because* they have certain special features. So, if either view is true, we should not expect our world to have such features.

To that last claim, there is one exception. This is the feature with which we began: that our world allows for life. Though this feature is, in some ways, special, it is one that we cannot help observing. That restricts what we can infer from the fact that our world has this feature. Rather than claiming that being life-allowing is one of the Selectors, we can appeal to some version of the Many Worlds Hypothesis. If there are very many worlds, we would expect a few worlds to be life-allowing, and our world is bound to be one of these few.

Consider next other kinds of special feature: ones that we are not bound to observe. Suppose we discover that our world has such a feature, and we ask whether that is no coincidence. It may again be said that, if there are many worlds, we would expect a few worlds to have this special feature. But that would not explain why that is true of *our* world. We could not claim---as with the feature of being life-allowing---that our world is bound to have this feature. So the appeal to many worlds could not explain away the coincidence. Suppose, for example, that our world were very good, or were wholly law-governed, or had very simple natural laws. Those facts would count against both of the unselective views: both the All Worlds Hypothesis and the Brute Fact View. It is true that, if all worlds exist, or there are very many randomly selected worlds, we should expect a few worlds to be very good, or wholly law-governed, or to have very simple laws. But that would not explain why

our world had those features. So we would have some reason to believe that our world is the way it is because this way has those features.

Does our world have such features: ones that count against the unselective views? Our world's moral character seems not to count against these views, since it seems the mixture of good and bad that, on the unselective views, we should expect. But our world may have the other two features: being wholly law-governed, and having very simple laws. Neither feature seems to be required in order for life to be possible. And, among possible life-containing worlds, a far greater range would not have these features. Thus, for each law-governed world, there are countless variants that would fail in different ways to be wholly law-governed. And, compared with simple laws, there is a far greater range of complicated laws. So, on both the unselective views, we should not expect our world to have these features. If it has them, as physicists might discover, that would give us reasons to reject both the All Worlds Hypothesis and the Brute Fact View. We would have some reason to believe that there are at least two partial Selectors: being law-governed and having simple laws.

There may be other features of our world from which we can try to infer what reality is like, and why. But observation can take us only part of the way. If we can get further, that will have to be by pure reasoning.

6

Of those who accept the Brute Fact View, many assume that it must be true. According to these people, though reality merely happens to be some way, *that* it merely happens to be some way does not merely happen to be true. There could not be an explanation of why reality is the way it is, since there could not be a causal explanation, and no other explanation would make sense.

This assumption, I have argued, is mistaken. Reality might be the way it is because this way is the fullest, or the most varied, or obeys the simplest or most elegant laws, or has some other special feature. Since the Brute Fact View is not the only explanatory possibility, we should not assume that it must be true.

When supporters of this view recognize these other possibilities, they may switch to the other extreme, claiming that their view's truth is another brute fact. If that were so, not only would there be no explanation of reality's being as it is, there would also be no explanation of there being no such explanation.

As before, though this might be true, we should not assume that it must be true. If some explanatory possibility merely happens to obtain, the one that obtains may not be the Brute Fact View. If it is randomly selected *whether* reality is randomly selected, and there are other possibilities, random selection may not be selected.

There is, moreover, another way in which some explanatory possibility may obtain. Rather than merely happening to obtain, this possibility may have some feature, or set of features, which explains why it obtains. Such a feature would be a Selector at a higher level, since it would apply not to factual but to explanatory possibilities. It would determine, not that reality be a certain way, but that it be determined in a certain way how reality is to be.

If the Brute Fact View is true, it may have been selected in this way. Of the explanatory possibilities, this view seems to describe the simplest, since its claim is only that reality has no explanation. This possibility's being the simplest might make it the one that obtains. Simplicity may be the higher Selector, determining that there is no Selector between the ways that reality might be.

Once again however, though this may be true, we cannot assume its truth. There may be some other higher Selector. Some explanatory possibility may obtain, for example, because it is the least arbitrary, or is the one that explains most. The Brute Fact View has neither of those features. Or there may be no higher Selector, since some explanatory possibility may merely happen to obtain.

These alternatives are the different possibilities at yet another, higher explanatory level. So we have the same two questions: Which obtains, and Why?

We may now become discouraged. Every answer, it may seem, raises a further question. But that may not be so. There may be some answer that is a necessary truth. With that necessity, our search would end.

Some truth is logically necessary when its denial leads to a contradiction. It cannot be in this sense necessary either that reality is a brute fact, or that there is some Selector. Both these claims can be denied without contradiction.

There are also non-logical necessities. The most familiar, causal necessity, cannot give us the truth we need. It could not be causally necessary that reality is, or isn't, a brute fact. Causal necessities come lower down. Similar remarks apply to the necessities involved in the essential properties of particular things, or natural kinds. Consider next the metaphysical necessity

that some writers claim for God's existence. That claim means, they say, that God's existence does not depend on anything else, and that nothing else could cause God to cease to exist. But these claims do not imply that God must exist, and that makes such necessity too weak to end our questions.

There are, however, some kinds of necessity that would be strong enough. Consider the truths that undeserved suffering is bad, and that, if we believe the premises of a sound argument, we ought rationally to believe this argument's conclusion. These truths are not logically necessary, since their denials would not lead to contradictions. But they could not have failed to be true. Undeserved suffering does not merely happen to be bad.

When Leslie defends the Axiarchic View, he appeals to this kind of non-logical necessity. Not only does value rule reality, Leslie suggests, it could not have failed to rule. But this suggestion is hard to believe. While it is inconceivable that undeserved suffering might have failed to be in itself bad, it is clearly conceivable that value might have failed to rule, if only because it seems so clear that value does *not* rule.

Return now to the Brute Fact View, which is more likely to be true. If this view is true, could its truth be non-logically necessary? Is it inconceivable that there might have been some Selector, or highest law, making reality be some way? The answer, I have claimed, is No. Even if reality is a brute fact, it might not have been. Thus, if nothing had ever existed, that might have been no coincidence. Reality might have been that way because, of the cosmic possibilities, it is the simplest and least arbitrary. And, as I have also claimed, just as it is not necessary that the Brute Fact View is true, it is not necessary that this view's truth be another brute fact. This view might be true because it is the simplest of the explanatory possibilities.

We have not yet found the necessity we need. Reality may happen to be as it is, or there may be some Selector. Whichever of these is true, it may happen to be true, or there may be some higher Selector. These are the different possibilities at the next explanatory level, so we are back with our two questions: Which obtains, and Why?

Could these questions continue for ever? Might there be, at every level, another higher Selector? Consider another version of the Axiarchic View. Reality might be as good as it could be, and that might be true because its being true is best, and that in turn might be true because its being true is best, and so on for ever. In this way, it may seem, everything might be explained. But that is not so. Like an infinite series of events, such a series of explanatory truths could not explain itself. Even if each truth were made true by the next,

we could still ask why the whole series was true, rather than some other series, or no series.

The point can be made more simply. Though there might be some highest Selector, this might not be goodness but some other feature, such as non-arbitrariness. What could select between these possibilities? Might goodness be the highest Selector because that is best, or non-arbitrariness be this Selector because that is the least arbitrary possibility? Neither suggestion, I believe, makes sense. Just as God could not make himself exist, no Selector could make itself the one that, at the highest level, rules. No Selector could settle *whether* it rules, since it cannot settle anything unless it does rule.

If there is some highest Selector, this cannot, I have claimed, be a necessary truth. Nor could this Selector make itself the highest. And, since this Selector would be the highest, nothing else could make that true. So we may have found the necessity we need. If there is some highest Selector, that, I suggest, must merely happen to be true.

Supporters of the Brute Fact View may now feel vindicated. Have we not, in the end, accepted their view?

We have not. According to the Brute Fact View, reality merely happens to be as it is. That, I have argued, may not be true, since there may be some Selector which explains, or partly explains, reality's being as it is. There may also be some higher Selector which explains there being this Selector. My suggestion is only that, at the end of any such explanatory chain, some highest Selector must merely happen to be the one that rules. That is a different view.

This difference may seem small. No Selector could *explain* reality, we may believe, if it merely happened to rule. But this thought, though natural, is a mistake. If some explanation appeals to a brute fact, it does not explain that fact; but it may explain others.

Suppose, for example, that reality is as full as it could be. On the Brute Fact View, this fact would have no explanation. On the Maximalist View, reality would be this way because the highest law is that what is possible is actual. If reality were as full as it could be, this Maximalist View would be better than the Brute Fact View, since it would explain reality's being this way. And this view would provide that explanation even if it merely happened to be true. It makes a difference where the brute fact comes.

Part of the difference here is that, while there are countless cosmic possibilities, there are few plausible explanatory possibilities. If reality is as full as it could be, that's being a brute fact would be very puzzling. Since there are countless

cosmic possibilities, it would be amazing if the one that obtained merely happened to be at the maximal extreme. On the Maximalist View, this fact would be no coincidence. And, since there are few explanatory possibilities, it would not be amazing if the Maximalist highest law merely happened to be the one that rules.

We should not claim that, if some explanation rests on a brute fact, it is not an explanation. Most scientific explanations take this form. The most that might be true is that such an explanation is, in a way, merely a better a description.

If that were true, there would be a different defence of the kind of reasoning that we have been considering. Even to discover *how* things are, we need explanations. And we may need explanations on the grandest scale. Our world may seem to have some feature that would be unlikely to be a coincidence. We may reasonably suspect that this feature is the Selector, or one of the Selectors. That hypothesis might lead us to confirm that, as it seemed, our world does have this feature. And that might give us reason to conclude either that ours is the only world, or that there are other worlds, with the same or related features. We might thus reach truths about the whole Universe.

Even if all explanations must end with a brute fact, we should go on trying to explain why the Universe exists, and is as it is. The brute fact may not enter at the lowest level. If reality is the way it is because this way has some feature, to know *what* reality is like, we must ask *why*.

7

We may never be able to answer these questions, either because our world is only a small part of reality, or because, though our world is the whole of reality, we could never know that to be true, or because of our own limitations. But, as I have tried to show, we may come to see more clearly what the possible answers are. Some of the fog that shrouds these questions may then disappear.

It can seem astonishing, for example, how reality could be made to be as it is. If God made the rest of reality be as it is, what could have made God exist? And, if God does not exist, what else could have made reality be as it is? When we think about these questions, even the Brute Fact View may seem unintelligible. It may be baffling how reality could be even randomly selected. What kind of *process* could select whether, for example, time had no beginning,

or whether anything ever exists? When, and how, could any selection be made?

This is not a real problem. Of all the possible ways that reality might be, there must be one that is the way reality actually is. Since it is logically necessary that reality be some way or other, it is necessary that one way be picked to be the way that reality is. Logic ensures that, without any kind of process, a selection is made. There is no need for hidden machinery.

Suppose next that, as many people assume, the Brute Fact View must be true. If our world has no very special features, there would then be nothing that was deeply puzzling. If it were necessary that some cosmic possibility be randomly selected, while there would be no explanation of why the selection went as it did, there would be no mystery in reality's being as it is. Reality's features would be inexplicable, but only in the way in which it is inexplicable how some particle randomly moves. If a particle can merely happen to move as it does, reality could merely happen to be as it is. Randomness may even be *less* puzzling at the level of the whole Universe, since we know that facts at this level could not have been caused.

The Brute Fact View, I have argued, is not necessary, and may not be true. There may be one or more Selectors between the ways that reality might be, and one or more Selectors between such Selectors. But, as I have also claimed, it may be a necessary truth that it be a brute fact whether there are such Selectors, and, if so, which the highest Selector is.

If that is a necessary truth, similar remarks apply. On these assumptions, there would again be nothing that was deeply puzzling. If it is necessary that, of these explanatory possibilities, one merely happens to obtain, there would be no explanation of why the one that obtains obtains. But, as before, that would be no more mysterious than the random movement of some particle.

The existence of the Universe can seem, in another way, astonishing. Even if it is not baffling that reality was made to be some way, since there is no conceivable alternative, it can seem baffling that the selection went as it did. Why is there a Universe at all? Why doesn't reality take its simplest and least arbitrary form: that in which nothing ever exists?

If we find this astonishing, we are assuming that these features should be the Selectors: that reality should be as simple and unarbitrary as it could be. That assumption has, I believe, great plausibility. But, just as the simplest cosmic possibility is that nothing ever exists, the simplest explanatory possibility is that there is no Selector. So we should not expect simplicity at both the

factual and explanatory levels. If there is no Selector, we should not expect that there would also be no Universe. That would be an extreme coincidence.
820

APPENDIX B STATE-GIVEN REASONS

According to what we can call

> *the State-Given Theory:* Whenever certain facts would make it better if we had some belief or desire, these facts give us a reason to have this belief or desire.

To decide whether we have such *state-given* reasons, we can first ask how we might respond to such reasons.

Suppose that, in

> *Case One*, some whimsical Despot credibly threatens that I shall be tortured for ten minutes unless, one hour from now, I both believe that 2 + 2 = 1, and want to be tortured. Some lie-detector test will reveal whether I really have this belief and desire. [821]

On the State-Given Theory, this man's threat gives me strong state-given reasons to have this belief and desire, since that is my only way to avoid being tortured. But I could not respond to such reasons by choosing to have this belief and desire.

One problem here is that I have *object*-given reasons that count decisively *against* believing that 2 + 2 = 1, and *against* wanting to be tortured. Suppose that, because I fail to have this belief and desire, this Despot tortures me. Someone might say: 'You idiot! Why didn't you believe that 2 + 2 = 1?' But this remark would be absurd. I could not help believing that 2 + 2 does *not* = 1. It would also be absurd to claim that I was an idiot in not wanting to be tortured. I might want to be tortured if I knew that this would be my only way to achieve some great good. That might be true, for example, if I have some life-threatening illness, and great pain would trigger some healing process in my body. But this example is not of that kind. This Despot will carry out his threat unless I want to be tortured, not as a means to some end, but as an end, or for the sake of being tortured. Since I am rational, I could not want to be tortured for its own sake. Given the awfulness of being tortured, I have a decisive object-given reason *not* to have this desire, and I could not help responding to this reason in the non-voluntary way.

Suppose next that this Despot gives me an easier task. In

> *Case Two*, I shall be tortured unless, one hour from now, I believe that a certain closed box is empty.

On the State-Given Theory, this threat gives me a state-given reason to have this

belief. And this reason would be unopposed, since I have no object-given epistemic reason *not* to believe that this box is empty. But as before, I could not respond to this alleged state-given reason by choosing to have this belief. Since I am rational, I could not choose to believe that this box is empty simply because I know that it would be better for me if I had this belief.

There are other possibilities. When it would be better for us if we had some belief, there are three main ways in which we might be able to cause ourselves to have this belief. One method is to make this belief true. In *Case Two*, for example, I might be able to open the closed box and take out anything that it contains. That would make me believe that this box is empty, thereby saving me from my Despot's threat.

In some other cases, we might cause ourselves to have some beneficial belief by finding evidence or arguments that gave us strong enough epistemic reasons to have this belief. This method is risky, since we might find evidence or arguments that gave us strong reasons *not* to have this belief. But we might reduce this risk by trying to avoid becoming aware of such reasons. If we are trying to believe that God exists, for example, we might read books written by believers, and avoid books by atheists. While we are acting in this way, it is worth adding, we may be fully rational not only practically but also epistemically. We may always respond rationally to our awareness of any epistemic reason or apparent reason. This may be why we have to take such care to avoid becoming aware of epistemic reasons not to believe what we are trying to believe.

In a third kind of case, it would be better if we had some belief that we know to be false, because we are aware of facts that give us decisive epistemic reasons not to have this belief. If we are rational, we could not have this belief while we are aware of these decisive reasons not to have it. But we might be able to make ourselves have this belief by using some technique like self-hypnosis. We could not choose to give ourselves beliefs whose content makes them too obviously false. When my Despot makes his first threat, I could not make myself believe that 2 + 2 = 1. No one could both understand this mathematical equation and believe it to be true. But suppose that, in

> *Case Three*, this Despot threatens that I shall be tortured unless, one hour from now, I believe that he is the world's greatest genius.

I might be able to hypnotize myself into having this false belief. I would have to make myself forget my epistemic reasons to believe that this man is *not* a genius. I might also have to make myself forget how and why I had caused myself to have this new, false belief, since my remembering these facts would be likely to undermine this belief. Since I am rational, I could not believe what I knew that I had no epistemic reasons to believe. For similar reasons, I might also have to give myself some false apparent memories of this Despot's brilliant achievements. But if I am a skilled self-

hypnotist, I might be able to do these things. I would then rationally come to believe that this man is the world's greatest genius, because these false apparent memories would give me decisive apparent reasons to have this belief.

Most of us do not have such self-hypnotic powers. But we can imagine coming to have them. We could then make ourselves have many false beliefs at will, just as directly as we can perform various other mental acts.

Return now to the view that we can have state-given reasons. State-Given Theorists claim that

> (1) whenever certain facts would make it better if we had some belief, these facts give us a reason to have this belief.

In cases of the kinds that I have just described, we would have no need to appeal to such reasons. It would be enough to claim that we have reasons to want to have such beneficial beliefs, and to cause ourselves to have them, if we can. These would be like any other reasons to want something to happen, and to make it happen if we can. There would be no point in adding that, as well as having reasons to *cause* ourselves to have such beliefs, we would have reasons to *have* them.

We can imagine another change in our psychology. It might become true that, when we believed that it would be better if we had some epistemically irrational belief, we sometimes didn't need to make ourselves have this belief with some voluntary mental act, like self-hypnosis. We might find ourselves coming to have such beneficial beliefs, with supporting sets of false apparent memories, in a non-voluntary way.

It may seem that, in *these* cases, we *could* significantly claim that we had state-given reasons to have these beliefs. As I have said, when we are aware of facts that give us decisive epistemic reasons to *have* some belief, we respond to most of these reasons, not by voluntarily *causing* ourselves to have this belief, but by coming to have this belief, and then continuing to have it, in a non-voluntary way. We might similarly claim that, when we found ourselves coming to have such irrational but beneficial beliefs, we would be responding to practical reasons to *have* these beliefs.

We ought, I suggest, to reject these claims. There would be two other, better ways to describe such cases.

On one description, in coming to have these beneficial beliefs, we would still be responding, though in a non-voluntary way, to our reasons to cause ourselves to have these beliefs. We often find ourselves doing something that we could also voluntarily do. For example, we might find ourselves suddenly trying to catch some object that we have just dropped, or moving our body to regain our balance, or raising our arms when we are falling so as to protect our head. If we saw some hand grenade that was about to explode, we might find ourselves throwing ourselves onto

this grenade, to save the lives of those around us. These would be non-voluntary responses to our reasons to act in certain ways. Suppose that, when my Despot makes his third threat, I find myself coming to believe that this man is a genius. I might here be responding in this non-voluntary way to my practical reason to cause myself to have this beneficial belief. This may be what happens in some actual cases of unconscious self-deception.

We might instead claim that, when we found ourselves coming to have such beneficial beliefs, we would not be responding to any reasons. The truth might be only that, when we believed that it would be better if we had some other belief, this belief would cause us to have this other belief. This would be partly like the way in which, when we believe that we are in danger, this belief causes adrenalin to be released into our blood stream, thereby helping us to respond more effectively to this danger. This release of adrenalin, though beneficial, does not involve a response to some reason. Nor, perhaps, do some cases of wishful thinking.

Return now to the claim that, in such cases, we would be responding to our reasons to *have* these beneficial beliefs. We ought, I have suggested, to reject this claim. If we were *causing* ourselves to have these beliefs, this process might be rational, and involve responses to reasons. We would be responding to reasons for *acting*, which would be provided by the facts that would make it good if we had these beliefs. But if we were merely *passively* coming to have these beliefs, this process would not be rational, or involve any response to reasons. Suppose that I cannot hypnotize myself into believing that my Despot is a genius. As a result, he tortures me. Someone might say: 'You idiot! Why didn't you respond to your reasons to believe this man to be a genius?' When we are aware of facts that give us decisive *epistemic* reasons to have some belief, we are less than fully rational if we fail to respond to these reasons by coming to have this belief. But if we cannot cause ourselves to have some beneficial but irrational belief, we would not be open to the slightest criticism if we failed to have this belief. And if we would be in no way irrational despite our failure to respond to our awareness of certain alleged reasons, this counts against the view that we have any such reasons.

We have other reasons to reject the State-Given Theory. Two reasons, we can say,

> *compete* when we could not successfully respond to both these reasons,

and they

> *conflict* when they support different answers to the same question.

If we have a moral reason to keep some promise, for example, and a self-interested reason to break this promise, these reasons compete, since we couldn't both keep and

break this promise. These reasons also conflict, since they support different answers to the question of what we have most reason to do.

Suppose next that we are aware of facts that give us decisive epistemic reasons *not* to have some beneficial belief. According to the State-Given Theory, the benefits of having this belief would also give us state-given reasons to have it. These two sets of reasons would compete, since we could not both have and not have this belief. On one version of this view, these reasons would also conflict. When we ask what we had most reason to believe, these reasons would support different answers to this question. We would have to decide whether our state-given reasons to have this belief were stronger than, or outweighed, our epistemic reasons not to have this belief.

We would not, I believe, have such conflicting reasons. When my Despot makes this third threat, I would be aware of facts that gave me decisive epistemic reasons *not* to believe falsely that this man is the world's greatest genius. If I had a state-given reason to have this belief, this reason would be provided by the facts that would make it bad to be tortured. I might ask whether, compared with being tortured, it would be worse to have such a false belief. But I would here be asking which of two outcomes I had more reason to want to prevent and to try to prevent. That is a question about the strength of two *practical* reasons, like any other reasons for wanting to prevent and trying to prevent some bad outcome. I could not rationally ask whether my state-given reason to have this false belief is stronger than, or outweighs, my *epistemic* reasons *not* to have it. It makes no sense to compare the strength of my evidence for the falsity of this belief with the badness of my being tortured.

Having seen that such comparisons make no sense, State-Given Theorists might turn to the claim that these two kinds of reason do not conflict, since they support answers to different questions. When we ask whether we ought to have some belief, we might be asking either

Q1: Is this a belief that I *ought epistemically* to have?

or

Q2: Is this a belief that I *ought practically* to have?

On this view, in answering Q1, we should consider only epistemic reasons; and in answering Q2, we should consider only practical state-given reasons. Since these are different questions, we cannot ask what we ought to believe, or what we have most reason to believe, *all things considered*.

These claims are partly right. There are, indeed, two questions here. But these claims do not help to show that we can have practical state-given reasons to have beliefs. Q2 needs to be explained, since it is unclear what it means to ask whether we *ought practically* to have some belief. This question could be more clearly stated, I

suggest, as

> Q3: What would it be best for me to believe? In other words, what do I have most reason to want to believe, and to cause myself to believe, if I can?

And this question is not about what I have reasons to *believe*. Like other practical questions, this question is about what I have reasons to *want*, and to *do*.

Since Q1 and Q3 are different questions, we never need to compare the strength of practical and epistemic reasons. [822] We *respond* to reasons. And we could never have practical reasons to respond in a certain way, while having epistemic reasons *not* to respond in this same way. When my Despot makes his third threat, I might respond to my practical reasons by acting in a way that would make me believe that this man is the world's greatest genius. I have no epistemic reasons *not* to act in this way, since epistemic reasons are not reasons for *acting*. I do have decisive epistemic reasons not to *believe* that this man is such a genius, and while I remember the facts that give me *these* reasons, I might respond to them in a non-voluntary way by losing this belief. But I have no practical reasons *not* to respond in this non-voluntary way. My practical reasons are to act in ways that would make me keep this belief until I have passed this Despot's lie-detector test, so that he will not torture me. These practical and epistemic reasons do *compete*, in the sense that I could not successfully respond to both sets of reasons. But these reasons do not conflict.

It is easy to overlook, or misunderstand, the distinctions that I have just drawn. As I have said, theoretical reasoning is a voluntary activity, in which we often engage for practical reasons. When we are doing mathematics, for example, we may have a practical reason to check some part of some proof, or to redo some calculation in a different way. These are reasons for acting in ways that may help us to reach the truth. While we are acting in these ways, for these practical reasons, we shall also respond to many epistemic reasons. While we are checking some proof, for example, we respond to epistemic reasons whenever we see what follows from what, and what must be true. Coming to have some such particular belief is not a voluntary mental act. Theoretical reasoning, we might say, involves both *practical* and *pure* epistemic rationality.

There are other close connections between practical reasons and certain epistemic reasons. Much of our practical reasoning consists in theoretical reasoning about practical questions. When we ask what we have most reason to do, we may be trying to reach some true answer to this question. And some facts may give us both a decisive practical reason to act in some way, and a decisive epistemic reason to believe that we have this practical reason. Return to the case in which your hotel is on fire, and you could save your life only by jumping into some canal. This fact would give you a decisive reason to jump, and a decisive reason to believe that you ought to jump. But though our practical and epistemic reasons are often very closely

related, and these kinds of reason can compete, they cannot ever conflict.

State-Given Theorists also claim that

> (2) whenever certain facts would make it better if we had some desire, these facts give us a reason to have this desire.

Compared with the claim that we can have state-given reasons to have beliefs, this claim is more plausible. We can object that, since beliefs aim at the truth, our reasons to have beliefs must all be epistemic, or truth-related. No such claim applies to desires. So it may seem that, just as we have an object-given reason to have some desire when, and because, *what we want* would be relevantly good, we have a state-given reason to have some desire when, and because, *our wanting something* would be good.

We do not, I suggest, have such reasons. Suppose that, in

> *Case Four*, my Despot declares that I shall be tortured for ten minutes unless, one hour from now, I want him to kill me. If I have this desire, and ask him to kill me, he will refuse, and set me free. As I know, this man always does what he declares that he will do.

Suppose next that the rest of my life would be well worth living. I would then find it difficult to want this man to kill me. But I might be able to hypnotize myself into having this desire during the next few hours. That would be what I had most reason to do, and what I ought rationally to do. This mental act would be a riskless way to avoid some intense pain.

State-Given Theorists might claim that their view explains why I ought to act in this way. They might argue:

> (A) I have a decisive reason to want this Despot to kill me, since that would save me from being tortured.
>
> (B) When we have a decisive reason to have some desire, this fact gives us a decisive reason to make ourselves have this desire, if we have some riskless way of doing that.
>
> (C) I have such a way of making myself want this man to kill me.

Therefore

> I ought to make myself have this desire.

Premise (A), however, is false. I have object-given reasons to want this Despot *not* to kill me, and these are also reasons not to want this man to kill me. These reasons are clearly stronger than my alleged state-given reason to want this man to kill me. Losing a life worth living is much worse than being tortured for ten minutes. So I do not have a decisive reason to want this man to kill me.

State-Given Theorists might reply that I don't have any reason not to want this man to kill me. If I had this desire, this man would not kill me but set me free. Since I have a reason to have this desire, and no reason not to have it, I ought rationally to cause myself to have this desire. On this view, all reasons to have desires are state-given, or provided by the benefits of having these desires.

To assess this view, we can suppose that, because my attempt to have this desire fails, this Despot tortures me. Someone might say: 'You idiot! Why didn't you want him to kill you?' But this remark would be unjustified. As before, if I am rational, I could not want this man to kill me merely because I know that, if I had this desire, that would be better for me. This point is clearer in a simpler case. If I learnt that I was fatally ill, it might be better for me if I wanted to die. But that wouldn't show that I had no reason to want not to die. It would be absurd for others to say 'You idiot! Why don't you want to die?' We should admit that, even after this Despot has made his threat, I have decisive object-given reasons to want this man not to kill me.

State-Given Theorists might next suggest that, since these reasons are of different kinds, they do not conflict. On this view, we can ask two questions:

Q4: What do I have the strongest object-given reasons to want?

Q5: What do I have the strongest state-given reasons to want?

But this suggestion fails. We can also ask

Q6: What do I have most reason to want all things considered?

If we have reasons for and against having the same desire, these reasons *do* conflict, since they support different answers to this wider question. It is irrelevant that these reasons are of different kinds. It might be similarly claimed that moral and self-interested reasons are of different kinds: but, when we ask what we have most reason to do all things considered, these reasons can conflict, by supporting different answers to this question.

In cases of the kind that we are now discussing, there *are* two questions that are worth asking. But these are not questions about two kinds of reason for or against having the same desire. Q6 can be restated as

> Q7: Which desires do I have most reason to have?

We can also ask

> Q8: Which desires do I have most reason to want to have, and to cause myself to have, if I can?

In *Case Four*, I could ask:

> If I wanted this Despot to kill me, would I be wanting something that I have decisive reasons to want?
>
> If I caused myself to have this desire, would I be doing something that I have decisive reasons to do?

My answers should be No and Yes. If I wanted this man to kill me, this desire would be in itself irrational, since I have decisive reasons *not* to want this man to kill me. But it would be rational for me to cause myself briefly to have this irrational desire, since this act would save me from being tortured.

There is another kind of case that gives us reasons to deny that we have state-given reasons to have desires. Suppose that, in

> *Self-defeating Desire*, I have a strong desire to get to sleep, because I need to sleep to improve my performance in some interview tomorrow. But I have one kind of insomnia. Whenever I strongly want to get to sleep, this desire makes me anxious about my failure to become sleepy, thereby keeping me awake. So I shall get the sleep I need only if I lose my desire to get to sleep.

My need for sleep gives me an object-given reason to want to get to sleep. According to the State-Given Theory, this need also gives me a state-given reason *not* to have this desire, since that would be my only way to get to sleep. These reasons would conflict, since they would be reasons for and against having the same desire. On this view, to decide whether I ought to have this desire, I should compare the strength of these two reasons. I should ask what I have most reason to want, all things considered.

I could easily compare the strength of these two reasons. My object-given reason to want to get to sleep is provided by the fact that I need sleep to improve my performance in my interview tomorrow. My alleged state-given reason *not* to have this desire would be provided by this same fact, together with the fact that having this desire would keep me awake. Since these reasons would both get their normative force from my need for sleep, their strength would be precisely equal. Since these reasons would also conflict, they would cancel each other out. The State-Given Theory therefore implies that, on balance, I have no reason to want to get to sleep. If

that were true, I would have no reason to have the aim of getting to sleep, and no reason to cause myself to lose this desire, so that I could achieve this aim. These claims are clearly false.

We ought, I suggest, to reject this State-Given Theory. I have no state-given reason not to have my desire to get to sleep. What I have are *object*-given reasons to *want* not to have this desire, and to *cause* myself to lose this desire, if I can. Unlike my alleged state-given reason *not* to have this desire, these reasons do not conflict with my object-given reason to *have* this desire. On this view, we reach the right conclusion. My need for sleep gives me a strong and unopposed reason to want to get to sleep, and this need also gives me a strong and unopposed reason to cause myself to lose this desire, since that is my only way to fulfil this same desire, thereby getting the sleep I need.

Whenever it would be better if we had certain beliefs or desires, we have reasons to want to have these beliefs or desires, and to make ourselves have them, if we can. But we do not, I suggest, have *state*-given reasons to have beliefs or desires.

We may have state-given reasons to be in some other kinds of state. I might truly claim, for example, that I have a reason to be in Paris next April. But as I have argued, such reasons would have no importance. It would be enough to claim that I have reasons to want to be in Paris next April, and to go there, if I can.

APPENDIX C RATIONAL IRRATIONALITY AND GAUTHIER'S THEORY

In an early article, Gauthier argued that, to act rationally, we must act morally. [823] I tried to refute that argument. [824] Since Gauthier was not convinced, I shall try again. [825]

1

Gauthier assumes that, to be rational, we must maximize our own expected utility. Though he distinguishes between 'utility' and 'benefit', this distinction does not affect his main arguments. We can regard him as appealing to Rational Egoism. [826]

Many writers have argued that, in self-interested terms, it is always rational to act morally. According to most of these writers, morality and self-interest coincide. But that is not Gauthier's line. Gauthier concedes that acting morally may be, and be known to be, worse for us. He claims that, even in such cases, it is rational to act morally.

If we appeal to Rational Egoism, it may seem impossible to defend that claim. How can our acts be rational, in self-interested terms, if we know them to be worse for us? But Gauthier *revises* Rational Egoism. On the standard version of this theory, an act is rational if it will maximize our expected benefit---or be *expectably-best* for us. [827] On Gauthier's version, it is rational to benefit ourselves not with our *acts* but with our *dispositions*. A disposition is rational if having it will be expectably-best for us. An act is rational if it results from such a disposition. In making these claims, Gauthier's view is like a version of Indirect Consequentialism.

Besides revising Rational Egoism, Gauthier restricts the scope of morality. To act morally, Gauthier claims, we must honour our agreements. In the cases with which he is concerned, each of us promises that, at some cost to ourselves, we shall give a greater benefit to others. If we all kept such promises, we would all gain. The cost to each would be outweighed by the greater benefits that each received from others.

Though such agreements are mutually advantageous, it would often be better for each of us if he or she broke this promise. Either we could break it secretly, or the damage to our reputation would be outweighed by what we would gain. We may think that, in self-interested terms, it is rational to break such promises. But Gauthier argues that, if we do, we are fools.

Gauthier's argument starts with a prediction. If we were straightforwardly self-interested---or, for short, *prudent*---we would intend to break such promises. Other

people, knowing this, would exclude us from these advantageous agreements. That would be worse for us. It would be better for us if we were trustworthy, since we would then be admitted to these agreements.

It would be even better for us, as I pointed out, if we merely *appeared* to be trustworthy but were really prudent. We would still be admitted to these agreements, but we would break our promises whenever we could expect that to benefit us. [828] Gauthier replied that we are too *translucent* to be capable of such deceit. When we were negotiating such agreements, we would sometimes be unable to conceal our true intentions. He therefore claimed that, on balance, it would be better for us if we were really trustworthy. [829]

Gauthier then appealed to his variant of Rational Egoism---which I shall call *Gauthier's view*. On this view, since it is in our interests to be trustworthy, it is rational for us to act upon this disposition. It is rational to keep our promises, even when we know that what we are doing will be worse for us.

Should we accept this argument? I believe not. When applied to trustworthiness, this argument may seem plausible. But we should reject Gauthier's view. It could be in our interests to have some disposition, and rational to cause ourselves to have it, but be irrational to act upon it.

2

One problem for Gauthier's view is that, at different times, different dispositions can be in our interests. This makes it hard to state Gauthier's view in a way that might achieve his aims.

In his earliest statements of his view, Gauthier assumed

> (A) If we have acquired some disposition because we reasonably believed that, by doing so, we would make our lives go better, it is rational to act upon this disposition. [830]

I challenged (A) as follows. [831] Just as it could be in our interests to be trustworthy, it could be in our interests to be disposed to fulfil our threats, and to ignore threats made by others. As before, it would be best to appear to have these dispositions, while remaining really prudent. But to test Gauthier's view, we should accept his claim that we are too translucent to be able to deceive others. It might then be better for us if we really had these dispositions. But it might not be rational for us to act upon them. [832]

I gave the following example, which I shall here call *Your Fatal Threat*. Suppose that you and I are on a desert island, and we are both transparent. You become a *threat-*

fulfiller. By regularly threatening to explode some bomb, you aim to make me your slave. My only way to preserve my freedom is to become a *threat-ignorer*, who is disposed never to give in to your threats. Since I am translucent, I can reasonably expect you to be aware of my disposition, which would be best for me. I manage to acquire this disposition. But I have bad luck. In a momentary lapse, you threaten that, unless I give you a coconut, you will blow us both to pieces. According to (A), it would be rational for me to ignore your threat. This would be rational even though I know that, if I do, you will explode your bomb, killing us both.

Gauthier once accepted this conclusion. [833] But he later revised his view, moving from (A) to

> (B) If we have reason to believe that, in acquiring some disposition, we made our lives go better, it is rational to act upon this disposition.

According to (B), for it to be rational to act upon some disposition, it is not enough that we *did* have reason to believe that, by acquiring this disposition, we would make our lives go better. We must *still* have reason to believe that this past belief was true. We need not 'adhere to a disposition in the face of its known failure to make one's life go better'. [834]

Gauthier intended (B) to handle my example. When you make your fatal threat, I lose my reason to believe that, in becoming a threat-ignorer, I made my life go better. On Gauthier's revised view, I need not 'adhere' to my disposition.

We can revise the example. Suppose I know that, if I had not become a threat-ignorer, I would have died some time ago. [835] Gauthier's view again implies that I should ignore your threat. Since my disposition once saved my life, my acquiring of this disposition made my life go better. True, this disposition will now kill me. But that is not what counts. According to (B), I should deny you the coconut, and be blown to pieces. [836]

As this example shows, even if some disposition has become disastrous, (B) can still imply that it is rational to act upon it. This would be rational if this disposition brought past benefits that were greater than its future costs. Gauthier claims that we should 'adhere' to such dispositions. We should be true to our 'commitment'.

When applied to promises, such a view has some appeal. If we have gained from trustworthiness, we may think it rational to act upon this disposition, even if it becomes a burden. Talk of *commitment* here makes sense. But in the case of threat-behaviour, it makes little sense. Why should I remain a threat-ignorer, at the cost of death, merely because this disposition once saved my life? [837]

If my alternative was to be your slave, my death might hardly be a cost. But we can add a further detail to the case. Suppose that a rescue party has just landed on the

beach. I know that, if I give you the coconut, I shall soon be freed.

To handle this version of the case, Gauthier must again change his view. It may have been rational for me to become a threat-ignorer. But as Gauthier must agree, it would now be rational for me to try to lose this disposition. [838] If I could cause myself to lose this disposition, it would be irrational to allow myself to keep it. Since that is so, Gauthier cannot claim that it must still be rational to act upon it. Now that I could soon be free, it would be irrational for me knowingly to bring about my death. [839]

How should Gauthier revise his view? He might restate claim (B) so that it covered temporary dispositions. But there is a simpler formulation. Gauthier could turn to

> (C) If we have reason to believe that, in having some disposition, we are making our lives go better, it is rational for us to act upon this disposition.

If he appealed to (C), Gauthier's view would not be challenged by my example. When I see that my disposition has become disastrous, (C) does not imply that it must still be rational for me to act upon it. [840]

I gave another example, which I shall here call *Schelling's Case*. A robber threatens that, unless I unlock my safe and give him all my money, he will start to kill my children. It would be irrational for me to ignore this robber's threat. But even if I gave in to his threat, there is a risk that he will kill us all, to reduce his chance of being caught. I claimed that, in this case, it would be rational for me to take a drug that would make me very irrational. The robber would then see that it was pointless to threaten me; And since he could not commit his crime, and I would not be capable of calling the police, he would also be less likely to kill either me or my children.

When Gauthier considered this example, he seemed to accept (C). He agreed that it would be rational for me to make myself, for a brief period, insane; and he claimed that it would be rational for me to act upon this disposition. [841]

If he turned to (C), however, Gauthier would pay a price. In his defence of contractual morality, Gauthier compared only permanent dispositions. He thought it enough to show that, if we are trustworthy, this will on the whole make our lives go better. [842] But if he appealed to (C), he would need to show more than this. According to (C), for it to be rational to act upon a disposition, it is not enough that it was earlier in our interests to acquire this disposition. We must have reason to believe that, *at the time of acting*, it is in our interests to have this disposition. Gauthier must therefore show that, if we are trustworthy, this disposition is in our interests when we are *keeping* our agreements.

He does not, I believe, show this. What he shows is, at most, that trustworthiness is in our interests when we are negotiating our agreements. In some cases, when the time comes to keep one agreement, we are negotiating some new agreement.

Gauthier's argument might then apply. But in other cases there is no such overlap. There are some promises that we could secretly and swiftly break, to our own advantage. When this is possible, it would be worse for us if we were trustworthy. It would be better for us if we lost that disposition, and became self-interested, even if only for just long enough to break our promise. [843]

To defend his view that it is always rational to act morally, Gauthier must claim that it would be rational to keep such promises. If he appealed to (C), however, he would lose his argument for that claim. (C) implies that it would be rational to break such promises, since we would then be acting on the disposition that we could reasonably believe to be, at the time, best for us.

Gauthier might try a different reply. He might claim that, if we are trustworthy, we would be unable to lose, or to overcome, this disposition. In the sense that is relevant here, this claim may not be true. [844] But suppose that it were true. Suppose that, because I am trustworthy, I would find it impossible to break some promise. Gauthier might appeal to the claim that 'ought' implies 'can'. He might say that, since I cannot break my promise, it cannot be true that it would be rational for me to do so. And he might say that, given the strength of my disposition, it would be rational for me to act upon it. [845]

Is this an adequate reply? Return to the case in which I am disposed to ignore your fatal threat. If I overcome my disposition, and thereby manage to remain alive until I can be rescued, Gauthier must agree that my act is rational. But suppose that my disposition proves too strong. I find that I cannot bring myself to give you the coconut. Could Gauthier claim that, since I cannot overcome my disposition, it cannot be true that it would be rational for me to do so? Could he claim that, since it is causally impossible for me to act differently, it is rational for me to bring about my death?

I believe not. For reasons that I give above, and as Gauthier elsewhere claims, what it would be rational for us to do does not depend, in this way, on what is causally possible. [846] We could have acted otherwise, in the relevant sense, if nothing stopped us from doing so except our desires or dispositions. If it would have been rational for me to have acted differently, it is irrelevant that, given my desires and dispositions, acting differently would have been causally impossible. Nor could I defend my act by appealing to the strength of my disposition. That may exempt *me* from certain kinds of criticism. But it cannot show that my *act* is rational. [847]

Gauthier admits as much in retreating from claim (A). Suppose that, though it was rational for me to acquire some disposition, I have learnt that doing so was a terrible mistake. Gauthier no longer claims that it must still be rational to act upon such dispositions. He agrees that, from the fact that I rationally acquired some disposition, and that I cannot now overcome it, we cannot infer that it is rational for me to act

upon it.

3

I have described one problem for Gauthier's view. Since it can be in our interests to have temporary dispositions, it is hard to state Gauthier's view in a way that might achieve his aims. Let us now ignore this problem, and turn to the central question. Should we accept Gauthier's view? Should we believe that, if it is in our interests to have some disposition, or rational to cause ourselves to have it, it is rational to act upon it?

In the cases with which we are concerned, though it is in our interests to have some disposition, it is against our interests to act upon it. Only here does Gauthier's view make a difference.

Reconsider *Schelling's Case*. Because I am temporarily insane, the robber knows that, even if he starts to injure my children, he would not thereby induce me to unlock my safe. That gives him reasons to give up and leave, which will be much better for me. [848] But while I am in my drug-induced state, and before the robber leaves, I act in damaging and self-defeating ways. I beat my children because I love them. I burn my manuscripts because I want to preserve them.

Gauthier objects that my crazy acts are, in fact, better for me. They are what persuades this man that I am immune to his threats. Since these acts are better for me, they are, on any view, rational. So this is not, as I claimed, a case of rational irrationality. [849]

To answer this objection, we can add one feature to the case. We can suppose that, to convince this man that I am crazy, I don't need to act in crazy ways. He sees me take this drug, and he knows that it produces temporary madness. Since the robber already knows that I am in this state, my destructive acts have no good effects.

Though my acts have only bad effects, they result from an advantageous disposition. That is enough, on Gauthier's view, to make these acts rational. [850]

Hume notoriously claimed that it would not be contrary to reason to prefer our own total ruin to the least uneasiness of some stranger. But Gauthier's view is more extreme. Hume at least required that, for our acts to be rational, we must be trying to achieve our aims. On Gauthier's view, we could be trying to frustrate our aims. When I burn my manuscript, or beat my children, I might be doing what I believe to be irrational, and *because* I believe it to be irrational. My acts could be as crazy as we can imagine. They could still, on Gauthier's view, be rational. [851] That is clearly false.

4

Of Gauthier's arguments for his view, one appeals to the claim that, if we accept his view, this will be better for us. We can first ask whether that is true.

Gauthier assumes that, to be rational, we should maximize our own expected utility. He compares two versions of this view. According to the standard version of Rational Egoism, which we can call *E*, we should maximize at the level of our acts. An act is rational if it maximizes our benefits or expected benefits. According to Gauthier's view, we should maximize only at the level of our dispositions. An act is rational if it results from a benefit-maximizing disposition. This view we can now call *G*. [852]

In the cases with which we are concerned, we cannot always maximize expected benefits at both levels. If we try to maximize with all our acts, we cannot have benefit-maximizing dispositions. Thus, if we break our promises whenever we can expect this to be better for us, we cannot be trustworthy, which will be bad for us. [853]

When we cannot maximize at both levels, it would be better for us if we had maximizing dispositions. The good effects of these dispositions would outweigh the bad effects of our acts. [854]

Gauthier claims that, given this fact, it will be better for us if we accept not E but G. [855] In making this claim, Gauthier assumes that, if we accept E, we would maximize with our acts *rather than* our dispositions.

This assumption may be incorrect. Since it would be better for us if we had maximizing dispositions, E would tell us, if we could, to acquire them. E agrees with G that we should try to *have* these dispositions.[856] What E denies is only that it must be rational to act upon them.

Gauthier may think that, if we accept E, we would always do what E claims to be rational.[857] Or he may think that, in judging any theory about rationality, we should ask what would happen if we always successfully followed this theory. This may be why he assumes that we would always maximize with our acts. But if we can change our dispositions, we cannot always do what E claims to be rational. Acquiring these dispositions would itself be a maximizing act. If we maximize with all our other acts, we shall have acted irrationally in failing to acquire these dispositions. If instead we acquire these dispositions, we cannot always maximize with our other acts. [858]

Since we cannot always do what E claims to be rational, we must do the best we can. And E implies that, rather than maximizing with our other acts, we should acquire maximizing dispositions. This is the way of acting that we can expect to be best for us. The disagreement between E and G is not over the question of whether we

should *acquire* maximizing dispositions. Like G, E claims that we should acquire such dispositions. The disagreement is only about whether, when we *act* on such dispositions, what we are doing is rational. [859]

Gauthier might now say that, if we accept E, we would be *unable* to acquire these dispositions. We would believe that, in some cases, acting on these dispositions would be irrational. And we might be unable to make ourselves disposed to do what we believe to be irrational. Perhaps, to acquire these dispositions, we must accept Gauthier's view, and believe that it is rational to act upon them.

When he discusses nuclear deterrence, Gauthier does make such a claim. [860] He supposes that it would be in our interests to form an intention to retaliate, if we are attacked. Forming this intention might be what protects us from attack. Gauthier then claims that, if we believed that such retaliation would be irrational, we would be unable to form this intention. [861]

It would be implausible to claim that we could *never* acquire some disposition if we believed that acting upon it would be irrational. *Schelling's Case* is one exception, and there are many others. But Gauthier would not need so strong a claim. He might say that it would often be impossible to acquire such dispositions. Or he might say that, if we believe that it would be irrational to act in some way, it would be more difficult for us to become disposed to act in this way. We might have to use some indirect method, such as taking drugs, or hypnosis, both of which have disadvantages. Things might be easier if we believed that it would be rational to act in this way. We might then be able simply to decide to do so. [862]

This may only shift the problem. How could we acquire this belief? Suppose that, as Gauthier claims, we could not intend to retaliate unless we believed that retaliation would be rational. If retaliation would be both pointless and suicidal, as Gauthier concedes, how could we persuade ourselves that, as Gauthier also claims, such retaliation would be rational? How could we make ourselves believe Gauthier's view? It is not easy to acquire some belief if our only ground for doing so is that this belief would be in our interests. Here too, we might need some costly indirect method. Let us, however, ignore this problem. Suppose next that it would be impossible for us to acquire some useful disposition unless we can somehow manage to believe that it would be rational to act upon it. It might then be in our interests to make ourselves acquire this belief. [863] It would then be worse for us if we accepted the standard version of Rational Egoism. It would be better for us if we accepted Gauthier's view. That would not yet show that Gauthier's view is true, or is the best view. To reach that conclusion, Gauthier needs another premise.

In the original version of his argument, Gauthier's other premise was---surprisingly---the standard version of Rational Egoism. He assumed that we should start by accepting E. We should believe that an act is rational if it will be expectably-best for

us. He then claimed that it would be better for us if we changed our own conception of rationality, by moving from E to G. Since it would be better for us if we made this change, E implies that it would be rational to do so. S tells us to believe that the true theory is not E but G. Gauthier concluded that the true theory *is* G.

Shelly Kagan suggested the following objection. [864] If E is true, G must be false, since E is incompatible with S. If E is false, G might be true, but G would not be supported by the fact that E tells us to believe G. It is irrelevant what a false theory tells us to believe. Either way, Gauthier's argument cannot support his conclusion.

Gauthier later revised his argument. He no longer claimed that we should first accept E, and then move to his view. He argued directly that we should accept his view. [865]

In this version of his argument, Gauthier's main claim still seems to be that, if we accept his view, this will be better for us. What should his other premise be?

Though he no longer appeals to E, Gauthier might still say that, if it is in our interests to accept some belief, it is rational to do so. He could then keep his claim that it is rational for us to accept G.

As before, such a claim does not imply that G is true. It could be rational to accept a false theory. But Gauthier might think it enough to show that it would be rational to accept his view. He might say that, even in the sciences, we cannot prove our theories to be true. We can at most show that it is rational to believe them.

Such an argument, however, would conflate two kinds of rationality. When we claim that it would be rational to have some belief, we usually mean that this belief would be *theoretically* or *epistemically* rational, since we have sufficient epistemic reasons to have it. Such reasons *support* this belief, since they are provided by facts which either entail this belief, or make it likely that this belief is true. But Gauthier's argument does not appeal to epistemic reasons. His claim would be that, since it is in our interests to believe his view, this belief would be *practically* rational. When we have practical reasons to cause ourselves to have some belief, these reasons do not support this belief, since they are not related, in relevant ways, to this belief's truth.

The point could be put like this. Gauthier claims that it is in our interests to believe that certain acts are rational. He concludes that such acts *are* rational. This argument assumes

> (D) If it is in our interests to believe that certain acts are rational, this belief is true.

Gauthier, however, rightly rejects (D). He imagines a demon who rewards various beliefs about rationality. He then claims that, if there were such a demon, it would

be 'rational to hold false beliefs about rationality'. [866] Gauthier here concedes that, though it would be in our interests to hold these beliefs, they would still be false. The fact that they would be in our interests could not make them true.

Could Gauthier withdraw this claim, and appeal to (D)? [867] It seems clear that he could not. Suppose that Gauthier's demon rewarded the belief that, for our acts to be rational, we must be called 'Bertie', and be wearing a pink bow tie. Gauthier could not claim that, if there were such a demon, this belief would be true. Nor do we need fantastic cases to refute (D). It might be in the interests of some people to have one belief about rationality, and in the interests of others to have some contradictory belief. Gauthier could not claim that these beliefs would both be true.

Since we should reject (D), we should reject this argument for Gauthier's view. Even if it were in our interests to believe Gauthier's view, or rational to cause ourselves to believe this view, this would not show that Gauthier's view was true.

This argument might show something. Gauthier might still claim that it would be practically rational to believe his view. But unless he claimed that his view was true, Gauthier would have to abandon his main aim. He could not argue that it *is* rational to act morally. He could only argue that this belief is a useful illusion. [868]

5

In his discussion of nuclear deterrence, Gauthier gave a second argument for his view. Gauthier assumed that it could be rational to form the intention to retaliate, if we are attacked. He then claimed that, since it would be rational to form this intention, it would be rational, if deterrence failed, to act upon it.

David Lewis rejected this inference. While agreeing that it could be rational to intend to retaliate, Lewis denied that retaliation would itself be rational. [869]

In his reply, Gauthier denied 'that actions necessary to a rational policy may themselves be irrational'. If we accept deterrent policies, he wrote, we 'cannot consistently reject the actions they require.' Since we 'cannot claim that such actions should not be performed', we cannot call them irrational. 'To assess an action as irrational is. . . to claim that it should not be. . . performed.' [870]

These retaliatory acts cannot be *necessary* to deterrent policies since, if these policies succeed, these acts won't even be performed. But this is a special feature of deterrence, which we can set aside. In most of the cases with which we are concerned, the relevant acts *would* be performed. Thus, if I become trustworthy, because this disposition will be in my interests, I must expect that I shall keep my promises. Similarly, in *Schelling's Case*, I must expect my drug-induced state to affect

my acts. In both cases, if I adopt the policy that will be good for me, I must expect to act in ways that will be bad for me.

Note next that, even in these cases, my acts aren't *required* by my policy. They aren't necessary to my policy's success. If they were, and my policy was good for me, my acts could not be bad for me. What is necessary to my policy is not my acts, but only my intention, or my disposition. My acts are merely the unwelcome side-effects.

This distinction, I believe, undermines Gauthier's reply to Lewis. If some policy is justified despite having bad effects, we may agree that, in one sense, these effects 'should occur'. But this only means, 'Things should be such that they occur'. And in accepting that claim, we need not endorse, or welcome, these effects. If we are giving a dinner party, things should be such that we later have to do the washing up. We can still have reasons to regret having to wash up. Similar claims apply to the acts that result from an advantageous disposition. We can agree that, in one sense, these acts should be performed. Things should be such that these acts will be performed. But we can still, consistently, believe these acts to be regrettable and irrational.

6

Gauthier suggests another argument in favour of his view. This view avoids, he claims, 'some of the unwelcome consequences' of Rational Egoism. The chief such consequence is that, on that theory, it could be a curse to be rational. [871]

This argument does not, I believe, support Gauthier's view. Gauthier admits that, even on his view, it might be a curse to be epistemically rational. That would be true if epistemic irrationality were directly rewarded. This unwelcome consequence, Gauthier claims, could not be avoided by any theory. [872] But that is not true. Gauthier could extend his view. He could similarly claim that our theoretical reasoning is epistemically rational if and only if it is in our interests. On this version of Gauthier's view, epistemic rationality could never be a curse. This revision would not, however, improve Gauthier's view. When crazy reasoning would be in our interests, that does not make it rational.

Epistemic irrationality could be in our interests, as any good theory should admit. So could practical irrationality. Both kinds of irrationality could be rewarded. It is no objection to Rational Egoism that it assumes or accepts these facts.

Gauthier makes one other claim in support of his view. He admits that, when his view is applied to *Schelling's Case*, it may seem counterintuitive. We may hesitate to claim that my crazy acts are rational. But Gauthier suggests that this is no objection, since 'whatever we might intuitively be inclined to say. . . "rationality" is a technical term in both Parfit's enquiry and my critique.' [873]

That is not so. I was asking what, in the ordinary sense, it is rational to want and do. And Gauthier claims that *Schelling's Case* 'shows that our ordinary ideas about rationality. . . . are sometimes mistaken.' Since Gauthier is arguing that we should revise our ordinary ideas, he cannot defend his use of 'rational' by making it a mere stipulation, which is true by definition. And that would also make his view trivial.

On Gauthier's view, acts are rational if they result from an advantageous disposition. Such acts are rational even if they are merely the regretted side-effects of this disposition, and are as crazy as we can imagine. That is very hard to believe. I have discussed what seem to me all of Gauthier's arguments for this view. None, I suggest, succeed. I conclude that we should reject this view. It could be in our interests to have some disposition, and be rational to cause ourselves to have it, but be irrational to act upon it.

Gauthier proposes a Hobbesian version of Contractualism, and defends a minimal morality, because he believes he can then show that, even in self-interested terms, we are rationally required never to act wrongly. No other moral theory, Gauthier claims, achieves this aim. [874] If Gauthier's argument fails, as I have claimed, we lose our main reason to accept Gauthier's minimal morality.

APPENDIX D DEONTIC REASONS

In defending premise (E) of the Kantian Argument for Rule Consequentialism, I suggest that

> (X) if the optimific principles require certain acts that we believe to be wrong, the features or facts that, in our opinion, make these acts wrong would not give us decisive *non*-deontic reasons not to act in these ways. What might be true is only that, by making these acts wrong, these facts would give us decisive deontic reasons not to act in these ways.

It may seem that, to defend (X), we could appeal to the claim that

> (1) if these acts were not wrong, we would not have decisive reasons not to act in these ways.

But it may be difficult to defend this claim. [875] If certain facts would make certain acts wrong, it is hard to suppose that such acts are not wrong, since there may be no possible world in which that is true. And even if we could appeal to (1), that would not show that it is the wrongness of these acts that gives us decisive reasons not to act in these ways. There may be facts that would make certain acts wrong if and only if these facts also gave us decisive non-deontic reasons not to act in these ways.

I know of no quick argument for (X), which is why I merely suggest that (X) is true. But one argument against (X) is worth discussing. When some people claim that some act is wrong, these people mean that we have decisive moral reasons not to act in this way. Though these people appeal to *moral* reasons, they would deny that there are any *deontic* reasons. On this view,

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

This claim conflicts with (X), since (2) implies that

> (3) if the optimific principles required some acts that are wrong, we would have decisive non-deontic reasons not to act in these ways.

Most of us, I believe, do not use 'wrong' in this *decisive-moral-reason* sense. Since we use 'wrong' in some other sense, we could justifiably reject (2). And (2), I believe, is least plausible in precisely the cases that we are now considering. If the optimific

principles did require some acts that are wrong, it is acts of the kind that we are now considering whose wrongness could most plausibly be claimed to give us a further, independent reason not to act in these ways. In some of these cases, we might even claim, the wrongness of these acts would give us our *only* reason not to act in these ways. If some method of contraception would be artificial, for example, this fact, when considered by itself, seems to give us no reason not to act in this way.

This example does not show that (2) is false if, as most of us believe, such methods of contraception are not wrong. In asking whether (2) is true, we cannot usefully consider acts that are clearly wrong, and ask what would be true if such acts were not wrong. As I have said, this counterfactual may be impossible, or at least too hard to imagine. But it may help to consider how certain people have changed their moral view. In describing this change of view, I shall redescribe these people's beliefs so that they apply to my imagined cases rather than to the slightly different versions of these cases which these people actually considered. Suppose first that, in

> *Bomb*, the runaway train is headed for the tunnel in which it would kill the five. You could save the five by throwing a bomb in front of the train. But I am standing nearby, so this bomb's explosion would also kill me.

Many people would believe this act to be wrong. After considering such cases, certain people accepted

> *the Priority Principle*: The negative duty not to kill has priority over the positive duty to save people's lives.

In explaining this principle, these people claimed that

> (4) it would be wrong to save several people's lives in some way that would also kill someone else.

Remember next that, in

> *Tunnel*, you could redirect the runaway train onto another track so that it would kill me rather than the five.

This imagined case has been much discussed, though it has little practical importance, because this case seems to many people a counter-example to the Priority Principle. When they considered *Tunnel*, several supporters of this principle changed their mind. These people ceased to believe (4). On their view, you would be morally permitted to save the five by redirecting the train, even though your act would also kill me. These people then supposed that, in

> *Bridge*, you could save the five only causing me to fall onto the track, thereby killing me but stopping the train.

This act, these people believed, *would* be wrong. These people concluded that, though it would *not* be wrong to save several people's lives by redirecting some threat so that it would kill fewer people, it *would* be wrong to save these people *by* killing someone else. [876]

According to (2), an act's wrongness does not give us a further, independent reason not to do it. That is true, some people believe, because the claim that some act is wrong adds nothing to the claim that we have decisive moral reasons not to act in this way. If these claims were true, it would always be enough to ask whether we have such decisive reasons not to act in some way. We would never need to ask, as a separate question, whether some act would be wrong.

These claims are, as I have said, least plausible in precisely the kinds of case that we are now discussing. I have just described how, when comparing cases like *Bomb, Tunnel,* and *Bridge*, several people changed their moral view. This was not a change of view about the strength of our reasons to act in certain ways. When these people considered *Tunnel*, they did not first decide that you would have sufficient reasons to save the five by redirecting the train, and then conclude that, since you would have such reasons, this act would not be wrong. What struck them first was that this way of saving the five would not be wrong. Some of these people then concluded that, since this act would not be wrong, the fact that you would be saving several people's lives would give you sufficient reasons to act in this morally permissible way. Similar claims apply to *Bridge*. When these people considered this example, they did not first decide that you would have a decisive reason not to save the five by killing me, and only then conclude that this act would be wrong. These people were struck first by the belief that this act would be wrong, and only then concluded that the wrongness of this act gave you a further, and perhaps decisive reason not to act in such a way.

Some of us, I have claimed, use the word 'wrong' in an indefinable sense, which I express with the phrase 'mustn't-be-done'. It is in cases like *Tunnel*, *Bomb*, and *Bridge* that we can most plausibly believe that certain acts are in this sense wrong. In both *Tunnel* and *Bridge*, you could save the five by acting in a way that would also kill me. From my point of view, being killed as a means in *Bridge* would be no worse than being killed as a side-effect in *Tunnel*. But of these similar acts, many people believe, it is only killing as a means that has the distinctive property of being something that mustn't-be-done. Such acts are *out*, or *impermissible*. And if some act mustn't-be-done, we can plausibly believe, this fact gives us a further, independent reason not to act in this way. These are the cases in which it seems least plausible to claim that, when some act is wrong, this fact doesn't give us any further reason not to do it.

If we can justifiably reject (2), as I have just argued, we can reject this argument against (X). I am therefore inclined to believe that, when the optimific principles

require certain acts, we would never have decisive *non*-deontic reasons not to act in these ways.

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

1

In the second section of the *Groundwork*, Kant writes:

> (A) All imperatives command either *hypothetically* or *categorically*. The former represent the practical necessity of a possible action as a means of attaining something else that one wills (or might will). The categorical imperative would be one which represented an action as objectively necessary of itself, without reference to another end. (G 414)

Kant here asserts that there are only two kinds of claim about what is practically necessary, or what we are required to do. Imperatives are *hypothetical* if they require us to do something as a means of achieving some end whose achievement we have willed. Imperatives are *categorical* if they require us to do something, not as a means of achieving any other end, but as an end, or for its own sake.

These are not, as Kant asserts, the only two kinds of imperative. Kant's remarks draw two distinctions, which combine to give us four possibilities. Some imperative may require us to act in some way either

as a means of achieving some end,	or	not as a means, but as an end or for its own sake

and either

if we will this act or the achievement of this end,	(1)	(2)
or		
whatever we will	(3)	(4)

All imperatives, Kant claims, are of types (1) or (4). Kant ignores (2) and (3). It

does not matter if we ignore imperatives of type (2), which require us to do something for its own sake, if and because we will this act. It matters greatly, however, if we ignore imperatives of type (3). Categorical imperatives are unconditional, in the sense that they apply to us whatever we want or will. All such imperatives, Kant's remarks imply, require us to act in some way, not as a means of achieving some end, but only as an end, or for the sake of acting in this way. That is not true. Of the imperatives which apply to us whatever we want or will, some might require us to act in some way as a means of achieving some unconditionally required end.

At one point, Kant seems to acknowledge that there might be such imperatives. He writes:

> What serves the will as the objective ground of its self-determination is an *end*, and this, if it is given by reason alone, must hold equally for all rational beings. . . The subjective ground of desire is an *incentive*; the objective ground of volition is a *motive*, hence the distinction between subjective ends, which rest on incentives, and objective ends, which depend on motives, which hold for every rational being. (G 427-8)

Kant here claims that, while some ends are subjective, there are also *objective ends*, which reason gives to all rational beings. Some of these might be ends in the ordinary sense of 'end', which refers to anything that, in acting in some way, we might be trying to achieve. These are what Kant calls *ends-to-be-produced*. Since Kant distinguishes between such objective ends and merely subjective ends, we would expect that, after describing a class of imperatives which are hypothetical, because they appeal to our subjective ends, Kant would describe a class of imperatives that are categorical, because they give us objective ends-to-be-produced. But Kant claims instead that all categorical imperatives declare some act to be necessary of itself, without reference to another end. This claim implies that there are no objective ends-to-be-produced given by reason to all rational beings. And in both the *Groundwork* and the *Second Critique*, Kant assumes that there are no such ends. Kant's formal Categorical Imperative may *indirectly* require us to try to achieve certain ends, as when Kant argues that his Formula of Universal Law implies that we are required to develop our talents. But that does not make this formula an imperative of type (3). Only ten years later, in his *Metaphysics of Morals*, does Kant claim that there are two such ends: our own perfection and the happiness of others. [877]

Since Kant later came to believe that there are two such objective ends-to-be-produced, it may seem not to matter that, in the *Groundwork* and the *Second Critique*, Kant assumes that there are no such ends. But this does matter. Kant's assumption makes a great difference to his arguments in these earlier, more important books.

To help us to assess these claims and arguments, we can next distinguish various senses in which Kant uses two of his most important terms: 'material' and 'formal'. These senses partly overlap with Kant's uses of 'hypothetical' and 'categorical'. In his most explicit definition, Kant writes:

> Practical principles are *formal* when they abstract from all subjective ends; they are *material* when they are grounded upon subjective ends, and hence on certain incentives (G 427-8).

Some imperative or principle 'abstracts' from our subjective ends, if this principle applies to us, or requires something from us, whatever we want or will. We can call such principles *normatively formal in sense 1*. Other principles apply to us only if we have certain desires, or subjective ends. We can call such principles *normatively material in sense 1*.

When some principle is in this sense normatively material, we can be *moved* to act on this principle, Kant assumes, only by a desire to achieve some subjective end. So we can also call such principles *motivationally material*. But when some principle is normatively formal in sense 1, because it applies to us whatever we want or will, our acceptance of this principle can move us to act, Kant claims, without the help of any the ordinary desires that Kant calls 'incentives'. We can call such principles *motivationally formal*.

We can call principles *teleological* if they require us to act in certain ways as a means of achieving some end. Kant sometimes uses the word 'matter' to refer, not only to subjective ends, but to any end-to-be-produced. Thus he defines the 'matter' of an action as 'what is to result from it' (G 428). Since teleological principles have a 'matter' in this wider sense, we can call such principles *normatively material in sense 2.*

There are also principles which are not teleological. Since these principles are not normatively *material* in sense 2, we can call them *normatively formal in sense 2*. These principles are *deontological* if they require us to act in some way as an end, or for its own sake, rather than as a means of achieving some other end. Two examples might be requirements not to lie, and not to injure anyone as a means of benefiting others. [878]

Some principles are neither purely teleological nor purely deontological, since these principles require us to act in certain ways partly as an end, or for its own sake, and partly as a means of achieving some other end. That is true, for example, of the principles that require us to keep our promises, and pay our debts. Such principles are often called 'deontological' in a sense that means 'not purely teleological'.

There is another kind of non-teleological principle. Rather than requiring us to act in certain ways, some principles impose some merely formal constraint on our decisions and our acts. One example is Kant's Formula of Universal Law, which requires us to

act only on maxims that we could will to be universal laws. We can call such principles *normatively formal in sense 3*.

Principles that are not, in this sense, normatively formal we can call *substantive*, or *normatively material in sense 3*. Deontological principles, we should note, are in this sense material, since they require us to act in certain ways. Kant claims that his formula requires 'mere conformity to law as such, without appeal to any law that requires acting in certain ways' (G 402). Deontological principles *are*, precisely, laws that require us to act in certain ways.

We have, then, three normative senses of both 'formal' and 'material', and one motivational sense. When applied to principles, these senses can be summed up as follows:

motivationally material:	*motivationally formal:*
motivates us only with the help of some desire	motivates us all by itself

normatively material in sense 1, or *hypothetical:*	*normatively formal in sense 1*, or *categorical:*
applies to us only if and because there is something that we want or will	applies to us whatever we want or will

normatively material in sense 2, or *teleological:*	*normatively formal in sense 2:*
tells us to act in a certain way as a means of achieving some end	not teleological

normatively material in sense 3, or *substantive:*	*normatively formal in sense 3:*
tells us to act in a certain way	imposes only a general constraint on our maxims or our acts. [879]

2

We can now turn to some of Kant's arguments for his Formula of Universal Law, which Kant also calls his *Formal Principle*, as I shall sometimes do below.

One of Kant's arguments, in *Groundwork 2*, assumes one of the claims that I have already discussed. Kant writes:

> all imperatives command either hypothetically or categorically. The former represent the practical necessity of a possible action as a means of achieving something else that one wills (or might will). The categorical imperative would be one which represented an action as objectively necessary of itself, without reference to another end. (G 414)

Kant later writes:

> we want first to enquire whether the mere concept of a categorical imperative may not also provide its formula containing the proposition which alone can be a categorical imperative. . . When I think of a *hypothetical* imperative in general I do not know before hand what it will contain. . . But when I think of a *categorical* imperative, I know at once what it contains. For since the imperative contains, beyond the law, only the necessity that the maxim be in conformity with this law, while the law contains no condition to which it would be limited, nothing is left with which the maxim of the action should conform but the universality of a law as such, and this conformity alone is what the imperative properly represents as necessary. Hence there is only one categorical imperative, and it is this: Act only in accordance with that maxim through which you can at the same time will that it become a universal law. (G 420-1)

In these passages, Kant argues:

> (1) 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.
>
> (2) 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

> (3) There is only one categorical imperative, which requires us to act only on maxims that we could will to be universal laws.

This argument fails. Kant's premises are false, And even if they were true, Kant's conclusion would not follow.

Both of Kant's premises, as we have seen, overlook those categorical imperatives which are teleological, requiring us to try to achieve some objective end-to-be-produced.

Kant's second premise also overlooks those categorical imperatives which are deontological, requiring us to act in some way partly or wholly for its own sake. Two examples would be requirements to keep our promises and not to lie. Such imperatives do not impose only a formal constraint.

As several writers note, Kant's conclusion involves a third mistake. Kant assumes that, if some imperative imposes only a formal constraint, this imperative must be his formula, which requires us to act only on maxims that we could rationally will to be universal laws. That is not true, since there are other possible formal constraints. One example is a requirement to act only in ways in which we believe that it would be rational for everyone to act. This requirement is quite different from Kant's Formula. If we are Rational Egoists, for example, we shall believe that everyone is rationally required to try to do whatever would be best for themselves, though we could not rationally will it to be true that everyone acts in this way. [880]

This mistake might be reparable. Kant might argue that, of the possible formal constraints, only his Formula of Universal Law meets some further requirement that any acceptable principle must meet. But this argument's other premises cannot be repaired. There is no hope of showing that, if some imperative is categorical, it must impose only a formal constraint.

Why did Kant make these mistakes? He may have had in mind, but failed to distinguish, the three senses in which imperatives can be normatively formal. If Kant had distinguished these senses, he would have seen that his argument assumes that being formal in sense 1 implies being formal sense 2, which implies being formal in sense 3. Kant could not have believed that these inferences are valid. The first inference assumes that, if some imperative applies to us whatever we want or will, it cannot require us to act in some way as a means of achieving some required end. That is obviously false. The second inference assumes that, if some imperative does not require us to try to achieve some end, it cannot require us to act in certain ways, but must impose only a formal constraint. That is also obviously false. Kant's failure to notice these points may be due to his preference for thinking at the most abstract level. Only that could explain how, in giving this argument, Kant overlooks the possibility of both teleological and deontological categorical imperatives. Kant

thereby overlooks most of the moral principles that other people accept.

We can turn next to *Groundwork 1*. Consider first these remarks:

> an action from duty has its moral worth. . . in the principle of volition in accordance with which the act is done without regard for any object of the faculty of desire. . . For the will stands between its a priori principle, which is formal, and its a posteriori incentive, which is material, as at a crossroads; and since it must still be determined by something, it must be determined by the formal principle of volition if it does an action from duty, since every material principle has been withdrawn from it. . . [Hence] mere conformity to law as such, without having as its basis some law determined for certain actions, is what serves the will as its principle, and must so serve it if duty is not to be everywhere an empty delusion. . . (G 399-402)

Kant's argument here is this:

> (1) An act has moral worth only when the agent's motive is to do his duty.
>
> (2) Such an agent acts on a principle which is not material, since it does not appeal to any of his desires.
>
> (3) Such a principle must be formal, requiring mere conformity to law as such.

Therefore

> (4) This requirement is the only moral law.

In explaining his first premise, Kant compares two philanthropists (398). The first helps other people out of sympathy, or because he wants to make them happy. The second helps others because he believes that to be his duty. Of these people, Kant claims, the first is lovable, and deserves praise, but only the acts of the second have moral worth.

This may be Kant's least popular claim, damaging his reputation even more than his claim that we should not lie to prevent a murder. Kant's view about moral worth has, however, been well defended. And we do not need to consider such defences, since this argument need not appeal to Kant's view about moral worth. Kant's first two premises could become

> (5) When we act in some way because we believe this act to be our duty, we are acting on some principle which does not appeal to our desires.

With some qualifications which we can here ignore, this claim is true.

According to this argument's other premise, if some principle does not appeal to our desires, it must require what Kant calls mere conformity to law. That is not true. Such a principle might require us either to try to achieve some end, or to act in certain ways. Kant's argument again overlooks all teleological or deontological principles.

Why did Kant assume that, if some principle does not appeal to our desires, it must require mere conformity to law? He may again have been misled by his failure to distinguish between his different uses of the words 'material' and 'formal'. The will, Kant writes:

> must be determined by the formal principle of volition if it does an action from duty, since every material principle has been withdrawn from it. . .

Kant here assumes that, if some principle is not normatively material in sense 1, because it does not appeal to our desires, this principle must be normatively formal in sense 3, imposing only a formal constraint on what we will. That is not true. Though such a principle must be normatively formal in sense 1, it might not be normatively formal in either sense 3, or sense 2. Kant's use of the word 'formal' blurs these distinctions.

There is another way in which Kant may have gone astray. In the same passage, Kant writes:

> the purposes we may have for our actions, and their effects as ends and incentives of the will, can give no actions unconditional and moral worth. . . In what, then, can this worth lie. . ? It can lie nowhere else than in the principle of the will without regard for the ends that can be brought about by such an action. (G 399-400)

In the first sentence here, Kant's use of the word 'ends' must refer to our subjective or desire-based ends. An act's moral worth lies, Kant claims, not in the agent's subjective end, but in the agent's motive, which is to do his duty. But when Kant later writes 'without regard for the ends that can be brought about by such an action', he seems to shift, without noticing this, to the wider use of 'end' that would cover all possible ends-to-be-produced, including ends that are objective, or categorically required. This may be why Kant mistakenly concludes that the moral law must be formal in the sense of having no 'regard for the ends' that our acts might bring about.

Groundwork 1 suggests another argument. Kant writes:

> . . . an action from duty is to put aside entirely the influence of inclination and with it every object of the will; hence there is left for the will nothing that could determine it except objectively the law and subjectively pure respect for this

> practical law. . . But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it. . ? Since I have deprived the will of every impulse that could arise for it from obeying some law, nothing is left but the conformity of actions as such with universal law, which alone is to serve the will as its principle, that is: I ought never to act except in such a way that I could also will that my maxim should become a universal law. (G 400-402)

Kant here 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

> (4) This requirement is the only moral law.

Kant's first premise here is true. Humeans might claim that, when our motive in acting is to do our duty, we must be moved by a desire to do our duty. But even if that were true, we would not be being moved by a desire for our act's effects.

Premise (2), however, is false. Return to Kant's philanthropist who promotes the happiness of others, not because he wants to make them happy, but because he believes this act to be his duty. Kant's argument implies that, since this person is not moved by a desire for his act's effects, he must be acting on some principle which is purely formal, requiring only that our acts conform with universal law. That is not so. This person might be acting on a principle that requires us to promote the happiness of others.

Premise (3), as we have seen, is also false, since a principle could be purely formal without requiring that we act on universalizable maxims.

Though premise (3) might be repaired, nothing can be done with premise (2). There is no hope of showing that, when our motive is to do our duty, we must be acting on some principle which is purely formal.

Why did Kant make this assumption? When our motive is to do our duty, this motive is purely formal in the sense that it does not involve, or abstracts from, the *content* of

our duty. This feature of our *motive* Kant may have mistakenly transferred to the *principle* on which we act. Jerome Schneewind writes that, on Kant's view, a moral agent acts on principle, and that

> the only principle available, because she is not moved by the content of her action, must be formal. The agent of good will must therefore be moved by the bare lawfulness of the act. [881]

Though such a person may be, in one sense, moved by 'the bare lawfulness' of her act, this sense is only that this person's motive is to do her duty. That leaves it open what this person believes her duty to be. She may be acting on some principle which is *not* formal, since it requires her either to try to achieve some end, or to act in some way for its own sake.

Kant may also be again misled by overlooking his distinctions between different kinds of end. In another summary of Kant's argument, Nelson Potter writes:

> All action to which we are determined by some subjective end. . . is action whose maxim is without 'moral content'. . . So the maxim of action from duty must be a maxim which is determined by no such end. . . The only other thing which could determine us to action would be some 'formal' principle, i.e. a principle containing no reference to any end.[882]

As Potter fails to note, there is here a fatal slide from the claim that acts from duty must not be determined by *subjective* ends, to the claim that such acts must be determined by a principle which does not refer to *any* end, not even an objectively required end-to-be-produced. Schneewind similarly writes:

> Given Kant's claim that means-ends necessity is inadequate for morality, it is plain that he must think there is another law of rational willing, and so another kind of 'ought' or 'imperative'. The kind of 'ought' that does not depend on the agent's ends arises from the moral law. . . [This law] Kant holds, can only be the form of lawfulness itself, because nothing else is left once all content has been rejected. [883]

There is here the same unnoticed slide. If some law does not depend on the agent's ends, it may still have *content*, requiring more than the mere form of lawfulness. And this law might require the agent to try to achieve some end. Mary Gregor similarly writes:

> [if] principles of reason based on a desire for some end are all conditioned principles, the unconditioned necessity of duty implies that the principle prescribing duty must be a merely formal principle. . . it follows. . . that this principle says nothing at all about our ends. It neither commands nor forbids the adoption of any end, but merely sets a limiting condition on our actions. . . [884]

These claims assume that, if some principle does not appeal to our desire for some subjective end, it cannot say anything about our ends, and can neither command nor forbid the adoption of any end. That does not follow.

It may be suggested that, in making these remarks, I have misinterpreted Kant. When Kant claims that moral principles must be purely formal, he may not mean that these principles cannot be material in the sense of requiring us to try to achieve certain ends. Kant may be making some other point. Consider, for example, these remarks in the *Second Critique*:

> a free will must find a determining ground in the law but independently of the *matter* of the law. But besides the matter of the law, nothing further is contained in it than the lawgiving form. (CPR 29)

Kant may seem here to assume that any practical law *has* matter, which is what this law tells us to try to achieve. His point may seem to be only that, though any law is, in this sense, 'material', our motive in following this law---or the determining ground of our will----should be provided not by this law's matter, but by the fact that it has *the form of a moral law*. And this may seem to be Kant's point, in the *Groundwork*, when he discusses his unsympathetic philanthropist. When Kant claims that, to act out of duty, we must be moved by a principle's law-giving form, he may mean only that we must be moved by our belief that our act is a duty. That could be true of Kant's philanthropist even if this person is acting on a principle which has 'matter' in the sense that it requires him to promote the happiness of others.

This suggested reading seems to me doubtful. Nor could this suggestion repair Kant's arguments. After discussing this philanthropist, Kant takes his argument to show that his Formal Principle is the only moral law. That could not be shown if Kant meant only that this man is moved by a belief that his act is a duty.

Consider next another passage in the *Second Critique*:

> The matter of a practical principle is the object of the will. This is either the determining ground of the will or it is not. If it is the determining ground of the will, then the rule of the will is subject to an empirical condition. . . and so is not a practical law. Now if we abstract from the law everything material, that is, every object of the will (as its determining ground), all that remains is the mere *form* of giving universal law. Therefore, either a rational being cannot think of his . . . maxims, as being at the same time universal laws, or he must assume that their mere form, by which they are fit for a giving of universal law, of itself and alone makes them practical laws. (CPR 27)

When Kant refers here to 'the mere *form* of giving universal law', he cannot mean 'the mere form of a moral law'. His point cannot be that, if principles have the form of a

moral law, that alone makes them practical laws. Kant takes this argument to show that, since we must 'abstract from the law everything material', we ought to act only on maxims that we could will to be universal, because only these maxims 'are fit for a giving of universal law'. [885] Kant must be referring here to his Formula of Universal Law.

In the paragraph just quoted, Kant comes close to seeing that his argument is invalid. The *Second Critique* was the fastest written of Kant's major works, and this paragraph shows the speed with which Kant wrote. What Kant calls the 'matter' of a principle, or the 'object of the will', is the object or aim which this principle tells us to try to achieve. This object would be the will's 'determining ground' if we were moved to act upon this principle by a desire to achieve this object. After remarking that this object either is *or is not* the will's determining ground, Kant claims that, if we abstract from the law every object of the will which is its determining ground, we are left only with the mere form of giving universal law. That is not so, as Kant's earlier remark implies. We may be left with some object of the will which is *not* the will's determining ground. One such object might be the happiness of others. We might be moved to try to achieve this object, not because we want to make others happy, but out of duty and our belief that the happiness of others is a categorically required end. We would not then be acting on a principle that was purely formal. So Kant's argument again fails to support his conclusion.

Consider next Kant's summary of his view:

> The sole principle of morality consists in independence from all matter of the law (i.e. a desired object) and in the accompanying determination of choice by the mere form of giving universal law which a maxim must be capable of having. (CPR 33)

Kant here forgets the difference between his two uses of the phrase 'the matter of the law'. On Kant's narrower use, this 'matter' is a desired object. On Kant's wider use, a law's 'matter' is whatever this law tells us to try to achieve, which might be some categorically required end. Kant assumes that, if some moral principle does not have 'matter' in his narrower sense, it cannot have 'matter' in this wider sense. This leads him to conclude that, if some moral principle does not appeal to a desired object, it must require the mere form of giving universal law. That is not true. As before, Kant overlooks all substantive categorical principles.

3

Near the end of *Groundwork 2*, Kant reviews all possible alternatives to his Formula of Universal Law. Some of these principles Kant calls 'empirical' in the sense that they

appeal to our desires. Other principles he calls 'rational' in the sense that they appeal to 'grounds of morality' which are 'based on reason'. Kant gives, as one example, a principle that requires us to promote our own perfection.

Kant defends his Formula by arguing against all other principles. The concept of perfection, he objects, is too vague. But Kant could not claim that *all* principles which are 'based on reason' must be too vague; so he must give some other argument against these other principles. At this critical point, Kant writes:

> I believe that I may be excused from a lengthy refutation of all these doctrines. That is so easy. . . that it would be merely superfluous labour. (G 443)

Kant's 'refutation' of all other principles takes only one paragraph. This begins:

> Whenever an object of the will has to be laid down as the basis for prescribing the rule that determines the will, there the rule is none other than heteronomy; the imperative is conditional, namely: *if* or *because* one wills this object, one ought to act in such or such a way; hence it can never command morally, that is, categorically. Whether the object determines the will by means of inclination, as with the principle of one's own happiness, or by means of reason directed to objects of our possible volition in general, as with the principle of perfection, the will never determines itself *directly*, just by the representation of an action, but only by means of an incentive that the anticipated effect of the action has upon the will. . . (G 444)

Kant here claims that all other principles can provide only hypothetical imperatives. To defend this claim, Kant first repeats his distinction between the two ways in which we can be moved to act on these other principles. When we are moved to act on these principles, Kant writes, our will may be determined either by means of inclination, as in the case of empirical principles, 'or *by means of reason*', as in the case of rational principles. But Kant then forgets this second possibility, since he goes on to claim that, in both these cases, our will would be determined by means of an 'incentive' which the anticipated effect of our act had upon our will. Kant distinguished earlier between *incentives*, which he defines as the 'subjective grounds of desire', and *motives*, which he defines as 'objective ends' or 'grounds of volition', which are 'given by reason alone' to all rational beings. So, when Kant claims that it can be only some *incentive* which moves us to act on these rational principles, he is inconsistently denying that, as he has just conceded, we could be moved to act on such principles not by an inclination but by reason.

Kant's argument requires him to deny that, when acting on such a rational principle, we could be moved by reason. To justify this denial, Kant might claim that reason does not give us any objective ends-to-be-produced. But though Kant's arguments in the *Groundwork* assume that reason gives us no such ends, Kant says nothing that

supports this claim. And if some rational principle requires us to try to achieve such an objective end, we could act upon this principle in the same reason-provided way in which we can act upon Kant's Formula of Universal Law. [886]

The *Second Critique* contains another version of Kant's 'refutation'. [887] Kant writes:

> If we now compare our *formal* supreme principle of pure practical reason. . . with all previous *material* principles of morality, we can set forth all the rest, as such, in a table in which all possible cases are actually exhausted, except the one formal principle. . .
>
> Practical Material Determining Grounds
> in the principle of morality:
>
> | | *Subjective* | |
> |---|---|---|
> | *External* | | *Internal* |
> | Education (Montaigne) | | Physical feeling (Epicurus) |
> | The civil constitution (Mandeville) | | Moral feeling (Hutcheson) |
> | | *Objective* | |
> | *External* | | *Internal* |
> | Perfection (Wolff and the Stoics) | | The will of God (Crusius and others) |
>
> Those in the first group are without exception empirical and obviously not at all qualified for the universal principle of morality. But those in the second group are based on reason. . . . the concept of perfection in the *practical* sense is the fitness or adequacy of a thing for all sorts of ends. This perfection, as a characteristic of the human being. . . is nothing other than talent and. . . skill. The supreme perfection in *substance*, that is, God. . . is the adequacy of this being to all ends in general. Now, if ends must first be given to us, in relation to which alone the concept of *perfection*. . . can be the determining ground of the will; and if an end as an *object* which must precede the determination of the will. . . is always empirical; then it can serve as the Epicurean principle of the doctrine of happiness but never as the pure rational principle of the doctrine of morals. . . so too, talents and their development. . . or the will of God if agreement with it is taken as the object of the will without an antecedent practical principle independent of this idea, can become motives of the will only by means of the happiness we expect

> from them; from this it follows, *first*, that all the principles exhibited here are *material;* second, that they include all possible material principles; and, finally. . that since material principles are quite unfit to be the supreme moral law. . . the formal practical principle of pure reason. . . is the *sole* principle that can *possibly* be fit for categorical imperatives. . . (CPR 39-41)

In this passage, Kant argues:

> There are only two material principles which might be objective and based on reason: the principles of perfection and of obedience to God's will.
>
> The concept of *perfection* is the concept of something's fitness or adequacy as a means of achieving ends. God is supremely perfect because he is an adequate means to every end.
>
> Since the idea of perfection cannot move us to act unless we have some end to which this perfection is a means, and since all such ends are empirical, or given by our desires, the principle of perfection cannot be moral, but can serve only as the Epicurean principle of pursuing our own happiness.
>
> The principle of obeying God's will also cannot move us to act except through the expectation of our own happiness.

Therefore

> These principles are material, and are the only possible material principles.
>
> Material principles cannot be moral laws.

Therefore

> Kant's Formula is the only moral law.

Kant's premises are all false; and even if they were true, Kant's conclusions would not follow. Kant writes, rather charmingly, that his table 'proves visually' that there are no other possible objective material principles; but 'possible' does not mean 'shown in Kant's table'. Perfection is not all instrumental. God's perfection could not be that of an ideal Swiss army knife, or all-purpose tool. It is not true that all of our ends are given by our desires, since we can have objective ends that are given to us by reason. If we act on some principle either of perfection or of obedience to God's will, our motive can be something other than a desire for our own happiness. Even if our motive would have to be this desire, that would not show that these are the only possible material principles. It is not true that material principles cannot be moral laws. And even if that were true, Kant's Formula is not the only formal principle, so this argument could not show that Kant's Formula is the only moral law.

Kant gives some other arguments for his Formula of Universal Law. These other arguments, I believe, also fail. But that does not matter. Moral principles can be justified by their intrinsic plausibility, and by their ability to support and guide our other moral beliefs. I have argued that, with some revisions, Kant's Formula provides a remarkably successful version of Contractualism, which Kant could defensibly, though not undeniably, claim to be the supreme moral law.

APPENDIX F KANT'S CLAIMS ABOUT THE GOOD

The Latin language has a defect, Kant writes, since it uses the words *bonum* and *malum* in two senses, which German distinguishes. Kant's claims can also be applied to the English words *good* and *bad*. When widened in this way, Kant's claims would be these. Where Latin has to use the same word *bonum*, and English has to use the same word *good*, German distinguishes between *das Gute* and *das Wohl*. And, where Latin has to use *malum*, and English has to use *bad*, German distinguishes between *das Böse* and *das Übel* (or *das Weh*). (CPR 59-60)

These claims are mistaken. Latin and English have words whose meaning is similar to 'das Wohl'. Two such words in English are 'well-being' and 'happiness'. And Latin and English have words whose meaning is similar to 'das Übel' and 'das Weh'. Three such words in English are 'ill-being', 'suffering', and 'woe'. The language which is impoverished is not, as Kant claims, Latin, or English, but Kant's own version of German. Kant uses 'Gute' and 'Böse' to mean only 'morally good' and 'morally bad'. In English and other versions of German, we can express the thought that, if someone suffers, that is both bad for this person, and a bad event. Kant's version of German cannot express such thoughts, and Kant seems not to understand them.

Consider, for example, Kant's remarks about the Latin sentence:

> *Nihil appetimus nisi sub ratione boni, nihil aversamur nisi sub ratione mali,*

or, in English,

> We want nothing except what we believe to be good, and we try to avoid nothing except what we believe to be bad.

Kant complains that, given the ambiguity of the words 'boni' and 'mali', this 'scholastic formula' is 'detrimental to philosophy'. This formula, Kant writes,

> is at least very doubtful if it is translated as:
>
> > we desire nothing except with a view to our well-being or woe,
>
> whereas if it is translated:
>
> > we will nothing under the direction of reason except insofar as we hold it to be morally good or bad,
>
> it is indubitably certain and at the same time quite clearly expressed.

Kant's translations are both incorrect. This 'scholastic formula' does not use 'boni' and 'mali' to mean 'well-being' and 'woe'. Nor does it use these words to mean only 'morally good' and 'morally bad'. This formula rightly assumes that we want many things because we believe them to be either morally or *non*-morally good. On Kant's second proposed translation, this formula would not be, as Kant claims, 'indubitably certain'. It would be seriously mistaken. That is well shown by the case of woe, or suffering. On Kant's proposal, for us to have a reason to want ourselves not to suffer---or, in his words, for us to 'will' this 'under the direction of reason'---our suffering would have to be morally bad. Since suffering is not morally bad, Kant's view implies that we have no such reason.

It might be suggested that I am misreading Kant, since Kant may use 'das Böse' in a way that covers non-moral badness. The word 'evil' is so used in many discussions of the problem of evil, since most theologians rightly regard suffering as part of this problem. My reading, however, seems to be correct. Kant continues:

> . . . good or evil is, strictly speaking, applied to actions, not to the person's state of feeling. . . Thus one may always laugh at the Stoic who in the most intense pains of gout cried out, 'Pain, however you torment me, I will still never admit that you are something evil (*kakon*, *malum*)', nevertheless, he was right. He felt that it was something bad, and he betrayed that in his cry; but that anything evil attached to him he had no reason to concede. . . (CPR 60)

As Terence Irwin notes, Kant misunderstands this Stoic claim. [888] This Stoic didn't mean that the pains of gout aren't morally bad, in the sense that applies only to agents and to acts. That claim would be trivial, since no one believes that pain is in that sense bad. The Stoic was making the controversial claim that his pain isn't even *non-morally* bad for him, or a bad state to be in.

Consider next Kant's remarks about Hedonism. Kant writes that, since good and evil must

> always be appraised by reason and hence through concepts, which can be universally communicated, not through mere feeling. . . a philosopher who believed that he had to put a feeling of pleasure at the basis of his practical appraisal would have to call that good which is a means to the agreeable, and evil that which is a cause of disagreeableness and of pain; for appraisal of the relation of means to ends certainly belongs to reason. (CPR 58)

Kant's thinking here is close to Hume's. Kant assumes that, since pleasure and pain are feelings, they cannot be appraised by reason, and judged to be good or bad. The most that hedonists could claim, he says, is that things are good if they produce pleasure, and bad if they produce pain, since reason is capable of judging that one thing produces another. Kant understates the implications of this view. If pleasure

cannot be in itself good, hedonists could not call something good because it produces pleasure. For something to be good because of its effects, its effects must be good. Hedonists could at most claim that some things are good, because they are effective, as a *means* of producing pleasure. But Hedonists would have to admit that other things are in the same sense good as a means of producing pain. So, on Kant's view, no form of normative Hedonism would make sense.

Why does Kant believe that, since pleasure and pain are feelings, they cannot be appraised by reason? Kant writes:

> the usage of language. . . demands that good and evil be judged by reason and thus through concepts which alone can be universally communicated and not by mere sensation which is limited to individual subjects and their susceptibility. (CPR 58)

This remark suggests that we could not rationally judge that it was bad to be in pain, since such a judgment would have to be made with public and communicable concepts, and not with a private sensation. But when we judge that pain is bad, that judgment is not a sensation. It is a judgment *about* a sensation, made with the communicable concepts *pain* and *bad*. Nor could Kant be assuming that, since the word 'pain' refers to a private sensation, this word has no communicable meaning. Kant does not deny that we can refer to pain. Kant's point must be that the concept *bad* cannot be applied to a sensation. As he explicitly claims,

> good or evil is, strictly speaking, applied to actions, not to the person's state of feeling (CPR 60).

Kant seems to make this claim because he either lacks, or rejects, the concept of something's being in itself non-morally good or bad. If we believe that events or states can be non-morally bad, we have no reason to deny that it can be bad to be in pain. Nothing is more clearly bad, in this non-moral sense, than being in extreme agony.

Kant's views about what is good or bad may be in part explained by the fact that he makes little use of the concept of a normative reason. Kant's main normative concepts are *required*, *permitted*, and *forbidden*. These concepts cannot express the thought that some things are in themselves good, or worth achieving, and others are in themselves bad, or worth avoiding or preventing. Kant says that he uses 'good' to mean 'practically necessary'. That is not what 'good' means. Something can be good, even though some available alternative would be even better. To understand this kind of goodness, or badness, we must be able to have the thought that certain properties or facts give us reasons, by counting in favour of our having some desire, or acting in some way. Pain is bad in the sense that its nature gives us reasons to want and to try to avoid being in pain.

Kant may, at certain points, have such thoughts. Thus he writes:

> What we are to call good must be an object of the faculty of desire in the judgment of every reasonable human being, and evil an object of aversion in the eyes of everyone (CPR 61).

And he writes:

> Someone who submits to a surgical operation feels it no doubt as an ill, but through reason he and everyone else pronounces it good (CPR 61).

Kant is unlikely to mean that such an operation is morally good, and he may not mean only that this operation is, like a murderer's poison, good as a means. Kant may mean that this operation has effects which are good in the non-moral sense, since it saves this person's life. And in writing 'feels it. . . as an ill, but through reason. . . pronounces it good', Kant seems to suggest that, in being an ill, this pain is bad. But despite such passages, Kant often claims that 'good' or 'evil' cannot be applied to states of feeling, and that well-being and woe cannot be in themselves good or bad. Thus he writes:

> The end itself, the enjoyment that we seek, is. . . not a *good* but a state of *well-being*, not a concept of reason but an empirical concept of an object of feeling. . . (CPR 62)

This feature of Kant's view is well shown by his claims about the principle of prudence. Kant often calls this principle a merely hypothetical imperative, assuming that it applies to us only because we want to promote our own future happiness. In its only important form, the principle of prudence is *not* hypothetical. According to this principle, even if we don't care about some act's likely effects on our future happiness---as some young smokers don't care about the cancer they may cause themselves to have in forty years---we have reasons to care, and we ought rationally to care. Dying early from lung cancer is not morally bad. But such deaths, and the suffering they cause, are in themselves bad for people, and impersonally bad. In much of his writing, as I have said, Kant seems not to have recognized these kinds of badness, and our non-moral reasons to care about them, and to prevent them if we can. This creates a huge gap in Kant's view. Practical reason, Kant suggests, makes only two kinds of claim. At one extreme, there is moral duty; at the other, instrumental rationality. There is little but a wasteland in between. If we are taught such a view, but we then cease to believe in moral duty, we shall believe only in instrumental rationality. That is the only kind of rationality in which many people now believe.

APPENDIX G AUTONOMY AND CATEGORICAL IMPERATIVES

The moral law, Kant claims, is a categorical imperative. We are subject to this law, Kant also claims, only if we give it to ourselves. If these claims are taken seriously, they cannot both be true.

Kant writes:

> If we look back upon all previous efforts that have ever been made to discover the principle of morality, we need not wonder why all of them had to fail. It was seen that the human being is bound to laws by his duty; but it never occurred to them that he is subject only to laws given by himself but still universal and that he is obligated only to act in conformity with his own will. . . I shall call this basic principle the principle of the **autonomy** of the will in contrast with every other, which I accordingly count as **heteronomy**. . . (G 432-2)

According to this 'basic principle', which we can call 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.

There are two other relevant possibilities. According to Nihilists, we are not subject to any principles, even if we give them to ourselves as laws. We can ignore that possibility here. According to what we can call

> *The Heteronomy Thesis*: We are subject to certain principles, and obligated to act in conformity with them, whether or not we give these principles to ourselves as laws, and whatever we will.

Though Kant does not explicitly refer to this thesis, he says that he will 'count as heteronomy' all principles which are not compatible with his Autonomy Thesis, and the Heteronomy Thesis is what all such other principles have in common.

We are *subject* to some principle when this principle applies to us. So we can call principles

> *autonomous* when they apply to us only if we give them to ourselves as laws,

and

> *heteronomous* when they apply to us whether or not we give them to ourselves as laws.

I shall return to the question of what Kant means by our *giving* ourselves some principle *as a law*.

As we have seen, Kant draws another, partly similar distinction. Principles are

> *hypothetical* imperatives if they require us to act in some way as a means of achieving some end whose achievement we have willed,

and

> *categorical* imperatives if they require us to act in some way whether or not we have willed the achievement of some end.

Hypothetical imperatives, Kant also writes, say that

> I ought to do something *because I will something else*. The moral and therefore *categorical* imperative in contrast says: I ought to do something even though I have not willed anything else. (G 441)

Kant's second sentence is ambiguous. He may mean that a categorical imperative applies to us unconditionally, whatever we have willed. But this sentence could be read more literally. Kant may instead mean that, though a categorical imperative applies to us only because we have willed that to be so, this imperative applies to us even if we have not *also* willed something *else*. On this reading, unlike hypothetical imperatives, a categorical imperative applies to us even if we have not also willed the achievement of some end.

With these distinctions we can describe four kinds of imperative:

Some imperative may
apply to us either

	only if and because we have willed that to be so	or	whether or not we have willed that to be so
and either			
only if and because we have willed the achievement of some end	strongly hypothetical		weakly hypothetical
or			
whether or not we have willed the achievement of some end	weakly categorical		strongly categorical

According to Kant's Autonomy Thesis, we are subject only to principles or imperatives that we give to ourselves as laws, and obligated only to act in conformity with our own will. This thesis implies that

> (1) hypothetical imperatives are strongly hypothetical, since these imperatives apply to us only if and because we have both willed them to apply to us, and willed the achievement of some end,

and that

> (2) categorical imperatives are weakly categorical, since these imperatives apply to us only if and because we have willed that to be so.

According to the Heteronomy Thesis, we are subject to certain principles or imperatives, and obligated to act in conformity with them, whether or not we give these imperatives to ourselves as laws. This thesis implies that

> (3) hypothetical imperatives are weakly hypothetical, since these imperatives apply to us only if and because we have willed the achievement of some end,

and that

> (4) categorical imperatives are strongly categorical, since these imperatives apply to us unconditionally, whatever we have willed.

We can now return to Kant's claim that the moral law is a categorical imperative. If Kant means that the moral law is a *strongly* categorical imperative, Kant must reject his Autonomy Thesis. As we have just seen, only *heteronomous* imperatives can be strongly categorical.

Kant may instead mean that the moral law is a *weakly* categorical imperative. But as I shall now argue, we ought to reject this claim, because we ought to reject Kant's Autonomy Thesis.

Kant writes:

> reason commands what ought to happen (G 408).
>
> reason alone. . . gives the law. . . (G 457)
>
> we stand under a discipline of reason, and in all our maxims we must not forget our subjection to it, or. . . detract anything from the authority of the law. . . (CPR 82)

Such remarks conflict with Kant's Autonomy Thesis. If reason alone gives the law, and we are subject to reason's laws, we are not subject only to laws that we give to ourselves.

Kant saw no conflict here. He assumes that, just as each of us has a will, each of us has, or is, *a reason*. He writes, for example, 'one cannot possibly think of a reason that would consciously receive direction from any other quarter with respect to its judgments. . .' (G 448) Kant therefore claims

> The law by virtue of which I regard myself under obligation. . . proceeds from my own pure practical reason, and in being constrained by my own reason, I am also the one constraining myself. (MM 418)

Such claims, I believe, are indefensible. Consider first the laws that govern theoretical reasoning. Such reasoning, it is sometimes said, should obey the laws of logic. But we need a distinction here. Consider, for example, two logical laws:

> *Non-Contradiction*: No proposition can be both true and false.
>
> *Modus Ponens*: If it is true both that *P* and that *If P, then Q*, it must be true that *Q*.

These laws are not normative, nor could our reasoning obey these laws. What we can obey are two closely related epistemic principles or laws. According to

> *the Non-Contradiction Requirement*: We ought not to have contradictory beliefs.

According to

> *the Modus Ponens Requirement*: We ought not to believe both that *P*, and that *If P, then Q,* without also believing *Q*.

Kant claims that, since reason is subject only to laws which it gives to itself, reason must regard itself as the source or author of such requirements. [889] We can accept these metaphorical claims if Kant means only that these laws are rational requirements.

According to Kant's Autonomy Thesis, I am subject to these requirements because I give them to myself as laws. I, Derek Parfit, give myself the law that requires me to avoid contradictory beliefs. Only a madman could think that. Nor would it help to say that it is *my reason* which requires that I avoid such beliefs. Kant's phrase 'my reason' could refer only to my rationality. My epistemic rationality is my ability to be aware of epistemic reasons and requirements, and to respond to both of these in my beliefs. There is no sense in which these abilities could be the source or author of these reasons and requirements. Nor could I or my rationality be the source or author of practical imperatives, such as the moral law.

It may be objected that, in making these remarks, I am not discussing Kant in his own terms. For example, Kant writes:

> to think of a human being who is accused by his conscience as one and the same person as the judge is. . . . absurd. . . a human being's conscience will, accordingly, have to think of *someone other* than himself (i.e. other than the human being as such) as the judge of his actions. . . This requires clarification, if reason is not to fall into self-contradiction. I, the prosecutor and yet the accused as well, am the same *human being* (numerically identical). But the human being as the subject of the moral lawgiving which proceeds from the concept of freedom and in which he is subject to a law that he gives himself (*homo noumenon)* is to be regarded as another (of a different kind) from the human being as a sensorily affected being endowed with reason, though only in a practical respect. . . (MM 438 and note)

In this passage, Kant claims that the human being both *is* and *is not* one and the same person, or human being, as his inner judge and prosecutor, since as a sensorily affected being endowed with reason he both *is* the same as---but ought also to be regarded (though only practically) as being *not* the same as---his noumenal self. A philosopher who could make such claims might seem likely to dismiss as quibbling my claim that I am not pure reason.

Kant, I believe, would not have responded in this way. Kant was rightly proud of having created what he called 'the critical philosophy'; and such philosophy, he writes, 'must proceed as precisely . . . as any geometer in his work' (CPR 92). Given Kant's great originality, and the difficulty of many of the questions which he tried to answer, it is not surprising that he often failed to be precise. And the answers to some of Kant's questions could not be precise. But to take Kant seriously in his own critical terms, we should try to state his ideas, and to assess his arguments, as clearly and carefully as we can.

Kant would not have believed that I, Derek Parfit, am pure reason. So, if pure reason gives me certain laws, I do not give myself these laws. And in being subject to these laws, I am not subject only to laws which I give myself. These truths, which Kant would have accepted, contradict Kant's Autonomy Thesis.

Some writers suggest that, when Kant talks of our *giving* ourselves some law, he uses 'give' in a different sense from that in which he claims that 'reason alone. . . gives the law.' Kant could then without contradiction claim that we give ourselves the laws that, in a different sense, reason alone gives. On the most plausible suggestion of this kind, when Kant talks of our giving ourselves some law, he means only that we *accept* this law, believing it to be a rational or moral requirement. Thomas Hill, for example, writes:

> The sense in which the principles of autonomy are 'imposed on oneself by oneself' is puzzling, but at least it is clear that Kant did not regard this as an arbitrary, optional choice but as a commitment that clear thinking reveals, implicit in all efforts to will rationally, the way one may think that commitment to basic principles of logic is implicit in all efforts to think and understand. . . a will with autonomy accepts for itself rational constraints independently of any desires and other 'alien' influences. [890]

Korsgaard similarly writes:

> you might pay your taxes. . . because you think everyone should pay their share, or because you think that people should obey laws made by popular legislation. These would be, in an ordinary sense, examples of autonomy---of giving the law to yourself because of some commitment to it or belief in it as a law. [891]

On this reading, Kant's Autonomy Thesis could be restated as

> *The Endorsement Thesis*: We are subject only to principles that we ourselves accept.

According to this version of Kant's view, there are some principles which reason gives

to us as laws, in the sense that these principles are rational requirements. But we are *subject* to such principles, and obligated to think and act in conformity with them, only if and because we accept these principles, or believe them to be true.

This version of the Autonomy Thesis, though more modest, has striking implications. On this view, when applied to Korsgaard's example, people ought to pay their share only if they themselves believe that they ought to pay. If we don't accept Kant's Formula of Universal Law, this formula does not apply to us. And if we accepted no moral principles, we would have no obligations, nor could any of our acts be wrong.

These would be unacceptable conclusions. The moral law, Kant claims, is a categorical imperative. I suggested earlier that, if Kant keeps his Autonomy Thesis, he might claim that the moral law is at least *weakly* categorical. We are subject to Kant's Formula, he might say, if we accept this formula. But Kant's Formula would not then be a *categorical* imperative. Moral laws, Kant claims, apply to all rational beings. If Kant's Formula did not apply to those rational beings who don't accept this formula, this formula could not be a moral law.

Kant might reply that everyone accepts his formula. This formula, Kant claims, 'is the sole law which the will of every rational being imposes on itself' (G 444). Since this claim cannot be an empirical generalization, Kant must mean that all rational beings *necessarily* accept this formula.

In what sense might it be necessary that everyone accepts Kant's Formula of Universal Law? At one point, Kant asks

> But why, then, *ought* I to subject myself to this principle? (G 449)

Kant then writes that, unless we can answer this question, we shall not have shown the moral law's 'validity and the practical necessity of subjecting oneself to it'. These remarks suggest that, for Kant's Formula to be valid, it must be *normatively* necessary that we accept this formula.

Given Kant's Autonomy Thesis, this suggestion raises two problems. First, even if we ought to accept Kant's Formula, that does not imply that we *do* accept this formula. And on both readings of the Autonomy Thesis, if we don't accept Kant's Formula, it does not apply to us.

Second, if we don't accept Kant's Formula, Kant's Autonomy Thesis undermines the claim that we *ought* to accept, or are *required* to accept, this formula. According to Kant's thesis, we are required to accept Kant's Formula only if we ourselves accept this requirement. If we do not accept this requirement, it does not apply to us. Nor would it help to claim that we are required to accept this requirement to accept Kant's Formula. That could not be true unless we accept this second requirement, and so on for ever. There is an infinite regress here, of the kind that is vicious rather than

benign.

Given these problems, Kant might appeal instead to some kind of *non-normative* necessity. Return to the principles that govern theoretical reasoning, such as the Non-Contradiction and Modus Ponens Requirements. On Kant's Autonomy Thesis, if we did not accept these requirements, they would not apply to us. But Kant might reject this counterfactual, on the ground that what it requires us to suppose is too deeply impossible. As Hill suggests and Kant might claim, all thinkers necessarily accept these requirements, since their acceptance is necessarily involved in, or in part constitutes, thinking. If we didn't believe that we ought not to believe both *P* and *not P*, we couldn't even count as *believing* P. In believing something, we are committed to disbelieving the negation of our belief. Similarly, if we really believed both *P* and *If, P, then Q,* we couldn't fail to believe that we ought either to believe *Q*, or give up one of these other beliefs.

Kant might make similar claims about the principles that govern instrumental rationality, such as the general Hypothetical Imperative that requires us not to will some end without at the same time willing what we believe to be the necessary means to this end. If we didn't accept this requirement, Korsgaard suggests, we couldn't even count as willing some end. The acceptance of such principles may be necessarily involved in being an agent. [892]

This defence of Kant's Autonomy Thesis would, however, undermine this thesis. According to the rival, Heteronomy Thesis, we are subject to various requirements whether or not we accept these requirements. To use the same examples, we are rationally required to avoid contradictory beliefs, and to take the necessary and acceptable means to our ends, and these requirements do not depend on our acceptance of them. For Kant's view to be different from the Heteronomy Thesis, and to be an assertion of *autonomy*, Kant must claim that these requirements, or their normativity, in some sense derive from or depend on us. He might claim that, if we did not accept these requirements, they would not apply to us. But as I have said, that would be very implausible. On the suggestion we are now considering, we can ignore this possibility, since the acceptance of these requirements is necessarily involved in our even being thinkers and agents. If that is true, however, there is no sense in which these requirements, or their normativity, could be claimed to derive from us.

There is another problem. These claims could not be applied to Kant's Formula of Universal Law. There is no hope of showing that, if we didn't believe that we ought to act only on universalizable maxims, we couldn't be agents, since we would be unable to act. There are many successful agents who have considered and rejected Kant's Formula.

Kant might claim that, even if we reject his formula, and believe it to be false, there is some other sense in which we do accept this formula, and give it to ourselves as a law.

But when applied to us as human beings, this claim would either be false, or would have to be given some sense which made it trivial. Kant might claim instead that we all necessarily accept his formula as noumenal beings in a timeless world. But such a claim would be open to decisive objections. Since Kant cannot defensibly claim that everyone *does* accept his Formula of Universal Law, Kant's claim could at most be that, *if we were fully rational*, we would all accept this formula.

According to Kant's Autonomy Thesis, if we do not accept Kant's Formula, it does not apply to us. To defend his view that his formula applies to all rational beings, Kant must revise his thesis. And as I have just argued, Kant's claim could at most be that we are subject only to those principles or requirements that we either do accept, or would accept if we were fully rational. We would be subject to these requirements even if, because we were not fully rational, we did not accept them.

Kant's Thesis, so revised, would cease to make any distinctive claim. On the rival, Heteronomy Thesis, we are rationally or morally required to have certain beliefs and to act in certain ways, and these requirements apply to us whether or not we accept them. Heteronomists could agree that, if we were fully rational, we would accept these requirements. If we did not accept these requirements, we would be failing to respond to our reasons for accepting them. So the difference between these views would disappear.

There is, I conclude, no defensible and non-trivial version of Kant's Autonomy Thesis. Kant claims, I believe rightly, that there are some categorical imperatives. We are often rationally or morally required to have certain beliefs, or to act in certain ways. And such requirements are unconditional, since they apply to us whether or not we accept them, and whatever we want or will. So we should reject what Kant calls his 'basic principle', according to which morality is grounded in the autonomy of the will.

In arguing against Kant's Autonomy Thesis, I have ignored one complication. In many passages, including some from which I have quoted, Kant uses the word 'heteronomy' in a different sense. When Kant talks of self-*legislation*, he means in part self-*determination*. Reason gives a law, Kant writes, when it determines the will (CPR 31). Since Kant often identifies reason with the will, he often assumes that, when reason determines the will, the will is determining itself. Kant also assumes that, since we are rational beings, it is our reason, or our will, which is our authentic self, or what is most truly us. So Kant believes that *we* are autonomous, or self-determining, when our acts are motivated by our reason, or our will. This can be called *motivational autonomy*.

There is *heteronomy* in this motivational sense when our acts are motivated by something other than our reason, or our will. That is true, Kant claims, when our acts

are motivated merely by some desire. Kant claims that, since our desires are non-voluntary products of our natural constitution, they are alien to our true self. In his words, when we merely try to fulfil some desire,

> the will does not give the law to itself, but an alien impulse gives it by means of the subject's nature (G 444).

When our acts are motivated merely by our desires, rather than by our reason or our will, we can call these acts *motivationally heteronomous.*

Kant's claims about motivational heteronomy contain, I believe, some important truths. But this other use of 'heteronomy' can cause confusion. For example, Kant writes:

> if the will does not give itself the law. . . heteronomy always results. . . only hypothetical imperatives become possible (G 441).

Our will does not give itself some law when our will is subject to some law that is not given by itself. That is so when we are subject to some valid imperative which is strongly categorical. When we act on some moral imperative, Kant claims, our reason can by itself motivate us without the help of any desire, so our act is *motivationally autonomous*. In the sense in which this claim is true, it would apply to our acting on imperatives which are strongly categorical, and in that sense *normatively heteronomous*. When we act on such imperatives, our acts need not be heteronomous in the quite different sense of being motivated by our desires. And when we are subject to strongly categorical heteronomous imperatives, we are not subject only to hypothetical imperatives. So Kant should not claim that, when there is *normative* heteronomy, only hypothetical imperatives are possible. By using the word 'heteronomy' in both normative and motivational senses, which he fails to distinguish, Kant conflates two very different things: motivation by desire, and strongly categorical requirements.

Like many other people, Kant often conflates normative and motivational claims. This has regrettable effects, some of which I discuss in Appendix H.

APPENDIX H KANT'S MOTIVATIONAL ARGUMENT

1

Near the start of *Groundwork 2*, Kant defines imperatives as

> *hypothetical* when they 'represent the practical necessity of a possible act as a means of achieving something else that one wills (or might will)',

and

> *categorical* when they 'represent an act as objectively necessary of itself, without reference to another end' (G 414).

If we claim some act to be necessary as a means of achieving some end, we may mean only that this act is a causally necessary means. And Kant later writes that hypothetical imperatives say 'what one must do in order to attain some end'(G 415). But when Kant defines these imperatives as representing some act's 'practical necessity', this necessity may be partly normative, since Kant may mean that we are rationally required to take the means to our ends. And when Kant defines categorical imperatives as claiming some act to be 'necessary of itself', this necessity seems purely normative. These imperatives, we can assume, are unconditional requirements. Unlike hypothetical imperatives, which apply to us only if and because we will the achievement of some end, categorical imperatives apply to us whatever we want or will.

After defining these two kinds of imperative, Kant asks how such imperatives are possible. Hypothetical imperatives, he answers, need no explanation or defence. If we know some act to be the only means of achieving some end, it is analytically true that we cannot fully will this end without willing this necessary means, 'insofar as reason has decisive influence on us'. Surprisingly, Kant then writes:

> (1) On the other hand, the question of how the imperative of morality is possible is undoubtedly the only one needing a solution. . . It cannot be made out by means of any example, and so empirically, whether there is any such imperative at all, but it is rather to be feared that all imperatives which seem to be categorical may yet be in some hidden way hypothetical. For example, when it is said 'you ought not to promise anything deceitfully', and one assumes that . . . an action of this kind must be regarded as in itself evil and that the imperative of the prohibition is therefore categorical: one still cannot show with certainty in any example that the will is here determined merely through the law, without any other incentive, although it seems to be so; for it is always possible that covert fear of disgrace, perhaps also obscure apprehension of other dangers, may have had an influence

> on the will. . . In such a case. . . the so-called moral imperative, which as such appears to be categorical and unconditional, would in fact be only a pragmatic precept that makes us attentive to our advantage. . . (G 417)

These remarks are puzzling. After asking how there can be categorical imperatives, Kant turns to the prior question of whether there *are* any such imperatives. When Kant writes that this question is not empirical, he might seem to mean that unconditional requirements, since they are normative, are not empirically observable, as detectable features of the world around us. Kant then remarks, however, that 'all imperatives which seem to be categorical may yet be in some hidden way hypothetical.' For example, there may seem to be a categorical imperative which forbids lying. But when someone refrains from lying, Kant points out, we cannot be certain that this person's motives were purely moral. This person's act may have been partly motivated by some self-interested fear or desire. In such a case, Kant concludes, the imperative not to lie, which seemed to be moral and categorical, would really be only pragmatic and hypothetical.

Suppose that, in stating this conclusion, Kant were using 'categorical' in the sense that he has just defined. Kant's claim would then be

> (A) If this person's motive for acting was not purely moral, the imperative not to lie would not here be an unconditional requirement, since this imperative would not apply to this person. Given this person's motives, he was not morally required not to lie.

This cannot be what Kant means. Kant did not have the strange belief that, if we conform to some moral requirement for motives that are not purely moral, this requirement does not apply to us. (A) is both clearly false, and inconsistent with many of Kant's other claims. For example, Kant often claims that we can fulfil duties of justice whatever our motive. He did not mean that, when we fulfil some duty of justice for self-interested motives, this duty did not apply to us. Kant's view is only that, if we do our duty for non-moral motives, our act does not have moral worth.

Since Kant cannot mean (A), he seems to have shifted to other senses of 'hypothetical' and 'categorical'. And Kant does use these words in other senses. In the *Second Critique*, he writes

> Imperatives themselves, when they are conditional---that is, when they do not determine the will simply as will but only with respect to a desired effect, that is, when they are hypothetical. . .(CPR 000)

Imperatives are hypothetical, in the sense Kant here defines, when they determine our will, or motivate us, only with the help of a desire for some effect. Imperatives would be categorical, in a corresponding sense, when they motivate us all by themselves,

without the help of any such desire. As Kant elsewhere writes

> Categorical imperatives differ essentially from [those that are hypothetical] [893], in that the determining ground of the action lies solely in the law of moral freedom, whereas in the others it is the associated ends that bring the action to reality. . .(L 486)

Kant defines a 'determining ground' as 'the motivating cause' of an act (L 493, 268, 582). To express these senses, we can call imperatives

> *motivationally hypothetical* when their acceptance motivates us only with the help of a desire for some end,

and

> *motivationally categorical* when their acceptance motivates us all by itself, or without the help of any such desire.

We can similarly say that, on Kant's other, normative definitions, imperatives are

> *normatively hypothetical* if they require us to act in some way as a means of achieving something that we want or will,

and

> *normatively categorical* if they require us to act in some way unconditionally, or whatever we want or will.

We can now suggest another reading of the end of passage (1). Kant imagines someone who conforms to the moral imperative not to lie, but who acts for some non-moral motive, such as fear of disgrace. Kant then comments that, if

> (B) this person's act was not motivated by his acceptance of this imperative,

it would be true that

> (C) this imperative was not, as it seemed, categorical.

If Kant meant that this imperative would not be *normatively* categorical, or an unconditional requirement, Kant's comment would, as I have said, be baffling. But Kant may mean that this imperative would not be *motivationally* categorical. (C) would then be another way of stating (B).

Though this suggestion would explain this part of passage (1), it would give us another problem. Shortly before this passage, Kant has presented and discussed his normative definitions of 'hypothetical' and 'categorical'. Near the start of (1), Kant asks

> Q1: Are there any categorical imperatives?

On the definition that Kant has just given, this should mean

> Q2: Are there any unconditional requirements? Are we required to act in certain ways, whatever we want or will?

But what Kant then discusses is

> Q3: Are there any requirements whose acceptance motivates us all by itself, or without the help of a self-interested desire?

Why this sudden, unexplained shift?

On what we can call the *conflationist* reading, Kant takes Q3 to be another way of asking Q2. Though Kant uses 'categorical' in both a normative and a motivational sense, he fails to distinguish these senses. Kant assumes that, if some imperative motivates us all by itself, that's what it is for this imperative to be an unconditional normative requirement.

Though there are some passages in which Kant seems not to draw this distinction, it is hard to believe that he was not aware of it. So we might next suggest another, *non-conflationist* reading of passage (1). Kant may assume that

> (D) if no one ever acted for purely moral motives, no one would be subject to categorical moral requirements.

On this view, moral imperatives must have the power to motivate us all by themselves. Passage (1) may be a misleading statement of (D). Kant claims that, if his imagined person did not act for purely moral motives, this person had no duty not to lie. But this may not be what he intended to say. He may have intended to claim that, if *all* cases were of this kind, there would be no categorical imperatives.

When we consider only passage (1), this suggestion seems fairly plausible. A few pages earlier, however, Kant explicitly claims that

> (E) even if no one has ever acted for purely moral motives, obedience to the moral law would still be 'inflexibly commanded by pure reason'. [894]

(D) and (E) cannot both be true.

We might next suggest, however, that (E) is not really Kant's view. Though Kant claims that we can never know that anyone has acted for purely motives, he also writes:

> the pure thought of duty. . . has by way of reason alone. . . an influence on the human heart [that is] much more powerful than all other incentives (G 410-11).

If Kant thought it possible that no one has ever acted for purely moral motives, it is hard to see how he could also believe that the pure thought of duty is much more powerful than all other motives. So Kant may assume that, since we *can* act for purely moral motives, we are subject to categorical requirements.

We have other reasons to believe that Kant assumes (D). There are many passages in which Kant seems to assume that

> (F) we cannot be subject to a categorical imperative unless this imperative motivates us all by itself.

Return for example to Kant's question 'How are all these imperatives possible?' Kant says that he is asking

> (2) how the necessitation of the will, which the imperative expresses. . . can be thought. . . We shall thus have to investigate entirely a priori the possibility of a categorical imperative, since we do not here have the advantage of its reality being given in experience, so that what would be necessary would not be to establish this possibility but merely to explain it. (G 420)

The reality of a categorical imperative, Kant seems here to assume, might have been given in experience, in which case this reality would have needed only to be explained. Kant seems to mean, by this imperative's 'reality', its ability to motivate us all by itself. He goes on to write

> (3). . . how such an absolute command is possible, even if we know its tenor, will still require special and difficult toil, which, however, we postpone to the last section.

In the last section of the *Groundwork*, Kant argues that pure reason can by itself motivate us, and much of Kant's *Second Critique* has the same aim. In passages (2) and (3), Kant seems either to conflate the normative and motivational senses of 'categorical', or to assume that these two senses go together, since an unconditional moral requirement must be able to motivate us all by itself.

In another passage, Kant writes that moral laws

> must hold not only for human being but for *all rational beings as such*, not merely under contingent conditions and with exceptions but with *absolute necessity* (G 408).

Kant here asserts that

> (G) true moral laws must be both universal and normatively categorical, applying to all rational beings whatever they want or will.

Kant continues

> . . . it is clear that no experience could give occasion to infer even the possibility of such laws. For by what right could we bring into unlimited respect, as a universal precept for every rational nature, what is perhaps valid only under the contingent conditions of humanity? And how should laws of the determination of *our* will be taken as laws of the determination of the will of rational beings as such. . . if they were merely empirical and did not have their origin completely a priori in pure but practical reason?

When Kant claims that moral laws must hold for all rational beings, this claim seems normative. But Kant then turns to motivation. If 'the laws of the determination of our will' were *merely empirical*, Kant writes, we could not assume that the same laws would apply to all rational beings. The laws to which Kant here refers cannot be normative requirements, since such requirements are *not* empirical, and we *could* assume that such normative requirements apply to all rational beings. Kant must be referring to laws about how our wills are determined, or how we can be moved to act. Only such laws might be merely empirical in a way that prevents our assuming that they apply to all rational beings. So, in asking whether there are moral laws which hold for all rational beings, Kant takes himself to be asking whether there are necessary truths about what motivates all such beings.

On the non-conflationist reading, Kant here assumes that

> (H) No principle can be a true moral law unless all rational beings would necessarily be motivated to act upon it.

When Kant claims that reason, or the moral law, must *determine* the will of all rational beings, he does not mean that this law must always *move* these beings, guaranteeing that their do their duty. Imperfectly rational beings can fail to do what morality requires. That is why, unlike God or other beings who are wholly good, imperfectly rational beings have duties. But the moral law, Kant may assume, must at least motivate all rational beings in the sense of making them to *some extent* disposed to their duty. We can be motivated to do our duty, even when we are not moved to act in this way. ((H), we can note, allows that we can do our duty for non-moral motives, so (H) does not implausibly imply that, when we act for non-moral motives, we are not subject to the moral law.)

Kant elsewhere writes:

> The question is therefore this: is it a necessary law for all rational beings always to appraise their actions in accordance with such maxims as they themselves could will to serve as universal laws? If there is such a law, then it must already be connected (completely a priori) with the concept of the will of a rational being as

> such. . . since if reason entirely by itself determines conduct (and the possibility of this is just what we want now to investigate), it must necessarily do so a priori. (G 426-7)

When Kant asks whether it is necessary for all rational beings to act only on universalizable maxims, his question again seems to be normative. But Kant then takes his question to be whether reason all by itself can determine conduct. Kant does not say that, to answer his normative question, we must answer another, motivational question. He treats these as a single question. This passage gives some support to the conflationist reading. But Kant may again be assuming here that the moral law cannot be normatively categorical, making unconditional requirements, unless this law is motivationally categorical, motivating us all by itself.

2

In *Groundwork 3* and elsewhere, Kant argues at length that his Formula of Universal Law, which I shall here call Kant's *Formal Principle*, is motivationally categorical. There are two ways to interpret these arguments. On one reading, Kant believes that he has already shown in *Groundwork 2* that, *if* there is a supreme moral principle, this must be Kant's Formal Principle. Kant then assumes that, to show that there is such a supreme principle, we must show that this principle meets one further requirement, by being motivationally categorical.

In many passages, however, Kant seems to suggest a more ambitious argument, which might show in a different way that Kant's Formal Principle is the supreme moral law. Kant seems to argue:

> (G) True moral laws must be both universal and normatively categorical, applying to all rational beings whatever they want or will.
>
> (H) No principle could be such a moral law unless the acceptance of this principle would necessarily motivate all rational beings.
>
> (I) 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.
>
> (J) Only Kant's Formal Principle has such motivating force.
>
> (K) There must be some moral law.

Therefore

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

principle of morality.

We can call this Kant's *Motivational Argument* for his Formal Principle. Premise (I) may explain more fully why Kant assumes that, for some law to be normatively categorical, this law must also be motivationally categorical. Kant seems to assume that, unless some law motivates us all by itself, it could not be necessary that this law would motivate all rational beings, and thereby be able to be a categorical requirement. [895]

One objection to this argument is posed by

> *Moral Belief Internalism* or *MBI*: No one could accept some moral principle without being, to some degree, motivated to act upon it.

If MBI were true, Kant's argument would be undermined, or made trivial. Premise (H) lays down a test that every possible principle would pass. It would be true of every moral principle that its acceptance would necessarily motivate all rational beings. Kant could not then defend premise (J), which claims that only Kant's Formal Principle has such necessary motivating power. Nor would Kant need to argue that his Formal Principle motivates us all by itself.

Suppose next that MBI is false. If we could accept moral principles without always being motivated to act upon them, (H) may seem too strong. As Kant often says, we are not always fully rational. It may seem implausible to claim that, for some principle to be a moral law, there must never be anyone who, even when being irrational, fails to be motivated by their acceptance of this principle. We might suggest that Kant should appeal instead to

> (H2) No principle can be a true moral law unless its acceptance would necessarily motivate all rational beings *insofar as they were rational*.

This is like the claim which, given our imperfect rationality, Kant makes about hypothetical imperatives. If we will some end, Kant writes, we would will what we know to be the necessary means 'insofar as reason has decisive influence' on us (G 417).

If Kant rejects MBI and appeals to (H2), however, his argument would face another, similar objection. On some views, even if we are fully rational, we might fail to be motivated to act on our moral beliefs. But this not Kant's view. Kant clearly assumes that

> (L) if we were fully rational, we would be motivated to do what we believed to be our duty.

Given (L), if Kant appealed to (H2), his argument would again be trivial. All moral principles would motivate all rational beings, insofar as they were rational. So Kant's argument must appeal to the bolder premise (H). That may be in one way an

advantage. Since (H) states a requirement that is harder to meet, there is more hope of defending the claim that only Kant's Formal Principle meets this requirement.

Could Kant defend this claim? Kant assumes that

> (M) all rational beings accept his Formal Principle, and give this principle to themselves as a law.

For example, Kant writes:

> Common human reason . . . always has this principle before its eyes (G 402).
>
> Everyone does in fact appraise actions as morally good or evil by this rule (CPR 69). [896]

If all rational beings necessarily accept Kant's Formal Principle, that would provide one sense in which this is the only principle that necessarily motivates all these beings. That would be true even if, as MBI claims, no one could accept any principle without being motivated to act upon it. (M), however, is clearly false. And Kant could not, I believe, defend (M) without assuming that his Formal Principle is the true moral law. Nor could this assumption be part of an argument that is intended to support this conclusion.

For Kant's argument to be worth giving, he must reject MBI, claiming that we could accept some moral principles without being motivated to act upon them. But Kant might claim that, while we could accept *false* moral principles without being motivated to act upon them, moral *knowledge* necessarily motivates. This defence of (J) would appeal to we can call

> *the Platonic view*: If some moral principle is true, that gives it the power to motivate all rational beings. [897]

If Kant appeals to this view, however, he could not defend (J) except by appealing to his argument's conclusion. If it is a principle's truth which gives it such necessary motivating power, Kant could not show that only his Formal Principle has such power except by showing that only his Formal Principle is true.

There is another way in which Kant's argument might support its conclusion. Rather than assuming that a principle's truth gives it the power to motivate all rational beings, Kant might run this inference the other way. Kant may assume that

> (N) if some principle has the power to motivate all rational beings, that makes this principle true.

If Kant could independently defend (N), he could then conclude that his Formal

Principle is the one true moral law.

Kant, I suggest, did argue in this way. What is most relevant here is Kant's discussion, in the *Second Critique*, of what he calls 'the method of ultimate moral inquiry'. In such inquiry, Kant claims,

> the concept of good and evil must not be determined before the moral law (for which, as it would seem, this concept would have to be made the basis) but only (as was done here) after it and by means of it (CPR 62-3).

Failure to grasp this truth has led, Kant writes, to

> all the errors of philosophers with respect to the supreme principle of morals. . . The ancients revealed this error openly by directing their moral investigation entirely to the determination of the concept of the *highest good*, and so of an object which they intended afterwards to make the determining ground of the will in the moral law. . . they should first have searched for a law that determined the will a priori and directly, and only then determined the object. . .

These claims can be given two readings. On a normative interpretation, Kant's claims are these. When these ancient philosophers asked what was the highest good, they were asking what we had most reason to want, or what was most worth achieving, or something of this kind. Their mistake was to assume that we should first try to decide what is the highest good, and could then conclude that this good end is what we ought to try to achieve. On this reading, Kant claims that we should reverse this procedure. We should start by asking what we ought to do, or what is right, and only then draw conclusions about what is good. In Rawls's phrase, rather than the good's being prior to the right, the right is prior to the good. [898]

What Kant writes, however, is that these philosophers should first have searched for a law that *determined the will*. This seems to mean that, rather than asking

Q4: What is the highest good?

we should ask

Q5: How are rational beings moved to act?

If we can find some law that necessarily determines the will, Kant remark suggests, we could then draw conclusions about both the right and the good. On this reading, rather than morality's being prior to, and thus in one sense determining, the motivation of rational beings, it is the motivation of such beings which is prior to, and determines, morality. The moral law must be founded, not on truths about the highest good, but on truths about motivation.

Kant makes several other claims which seem to express this second view. Thus, after claiming that the concept *good* must not be determined before the moral law, Kant continues:

> That is to say: even if we did not know that the principle of morality is a pure law determining the will a priori, we would at least have to leave it undecided in the beginning whether the will has only empirical or else pure determining grounds a priori. . . since it is contrary to all basic rules of philosophical procedure to assume as already decided the foremost question to be decided. (CPR 63)

The 'foremost question', Kant here assumes, is about motivation. And Kant writes that, on the view that he is rejecting,

> . . . it was thought to be necessary first of all to find an object for the will, the concept of which, as that of a good, would have to constitute the universal though empirical determining ground of the will.

Kant claims that, on this mistaken view, the good is whatever empirically determines the will. On the true view, Kant then writes, the concepts of *good* and *evil* are 'consequences of the a priori determination of the will'. Both views, on Kant's account, describe the good in motivational terms.

Consider next this claim:

> Suppose that we wanted to begin with the concept of the good in order to derive from it laws of the will. . . since this concept had no practical a priori law for its standard, the criterion of good and evil could be placed in nothing other than the agreement of the object with our feeling of pleasure or unpleasure.

Since this claim is about the criterion of good and evil, it may seem to be normative. Kant may seem to mean that, if we start by asking what is good, in the sense of what we have reason to try to achieve, our answer would have to be: only whatever gives us pleasure. But as the context shows, Kant's claim is again about motivation. If we start with the concept of the good, Kant writes,

> then this concept of an object (as a good object) would at the same time supply this as the sole determining ground of the will.

He also writes

> If the concept of the good is not to be derived from an antecedent practical law but, instead, is to serve as its basis, it can only be the concept of something whose existence promises pleasure and thus determines the causality of the subject, that is the faculty of desire, to produce it (CPR 58).

Kant seems here to claim that, if the concept of the good is not derived from the moral law, we would have to regard the good as whatever motivates us, and our answer would have to be: whatever gives us pleasure. On this account, when hedonists say that pleasure is the only good, their claim is psychological.

Kant's account is too narrow, since Greek Hedonism often took a normative form. When Epicurus claimed that what is best is a life without pain, he meant that having such a life is what is most worth achieving. And when other writers claimed that pleasure is not the only good, they did not mean that things other than pleasure can motivate us.

When Kant claims that the concept of the good should be derived from the moral law, he may mean in part that, in Rawls's phrase, the right is prior to the good. But as these other passages suggest, Kant seems to hold another, more radical view. The 'foremost question', Kant claims, is whether there is some law that necessarily determines the will. If there is such a law, Kant seems to assume, this law will tell us both what is right and what is good. When Kant refers to a law 'that determines the will', Rawls takes this to mean that such a law 'determines. . . what we are to do', *i.e.* what we ought to do. [899] But this cannot be all that Kant means. When Kant asks 'whether the will has only empirical or also pure determining grounds' (CPR 63), he is asking what motivates us. And he writes:

> Either a rational principle is. . . in itself the determining ground of the will. . . in which case this principle is a practical law a priori. . . the law determines the will directly and the action is in itself good. . . or else a determining ground of the faculty of desire precedes the maxim of the will. . . in that case such maxims can never be laws (CPR 62).

On Kant's view, these remarks suggest, if there is some principle that necessarily determines the will of all rational beings, this principle's motivating power makes it the true moral law.

3

We can now ask whether Kant's Motivational Argument could succeed. Could Kant show, or give us reason to believe, that only his Formal Principle would necessarily motivate all rational beings?

Kant believed that, when we act on his Formal Principle, our motivation takes a unique form. It is often claimed that, in his account of non-moral motivation, Kant is a psychological hedonist. That claim, however, is misleading. Except when he discusses his Formal Principle, Kant is a hedonist about even moral motivation. Hence Kant's surprising claim that

> all material practical principles. . . are, without exception, of one and the same kind and come under the general principle of self-love or of one's own happiness (CPR 22).

After noting that we can be happy to have done our duty, Kant writes:

> Now a *eudaimonist* says: this delight, this happiness, is really his motive for acting virtuously. The concept of duty does not determine his will *directly*; he is moved to do his duty only *by means* of the happiness he anticipates. (MM 378)

This is just what Kant claims about how we can be moved to act on all material or substantive principles, such as requirements to promote our own perfection or the happiness of others. Kant writes that, even when our will is determined

> by means of reason. . . as with the principle of perfection, the will never determines itself directly, just by the representation of an act, but only by means of an incentive that the anticipated effect of the action has upon the will (G 444).

Though Kant admits that such principles have 'determining grounds' that are 'objective and rational', he claims that such principles

> can become motives of the will only by means of the happiness we expect from them (CPR 41).

We can be moved to act, Kant often says, in only two ways. Either our will is determined by 'the mere lawful form' of our maxim, since we are acting on his Formal Principle,

> or else a determining ground of the faculty of desire precedes the maxim of the will, which presupposes an object of pleasure or displeasure and hence something that *gratifies* or *pains* (CPR 62).

He also writes:

> all determining grounds of the will except the one and only pure practical law of reason (the moral law) are without exception empirical and so, as such, belong to the principle of happiness. . . (CPR 93)

> The direct opposite of the principle of morality is the principle of one's own happiness made the determining ground of the will; and. . . . whatever puts the determining ground that is to serve as a law *anywhere else* than in the lawgiving form of the maxim must be counted in this (CPR 25).

In these and other passages, Kant assumes that

> (O) when we act on Kant's Formal Principle, reason directly and by itself

> motivates us. In all other cases, our motivation takes a hedonistic form.

When Kant claims that 'material principles' are 'quite unfit' to be moral laws, he seems to be appealing to (O). His objection seems to be that, since such principles motivate us in this hedonistic way, they cannot be guaranteed to motivate all rational beings. Even if we all got pleasure from acting---or from the thought of acting---on some material principle, that would be a contingent fact, which depended on our natural constitution. We cannot assume that all rational beings would get similar pleasure, and would thus be motivated to act upon this principle (CPR 34). For some principle to be guaranteed to motivate all rational beings, as is required of any moral law, this principle must motivate us in a different, non-hedonistic way. And that is true, Kant claims, only of his Formal Principle.

Kant did not always assume (O). In one passage in the *Groundwork*, Kant writes:

> In order for a sensibly affected rational being to will that for which reason alone prescribes the 'ought', it is admittedly required that his reason have the capacity to induce a feeling of pleasure or of delight in the fulfilment of duty. . . (G 460)

This remark implies that

> (P) even when we act on Kant's Formal Principle, our motivation must be hedonistic.

Kant seems to be assuming here that, when we accept his Formal Principle, reason always produces in us the needed feeling of pleasure or delight. If we accepted other principles, Kant might claim, reason would not produce in us this feeling. This could be how, compatibly with (P), only Kant's Formal Principle would necessarily motivate all rational beings.

Kant's accounts of motivation are too hedonistic. Even when applied to non-moral motivation, Psychological Hedonism is mistaken. But Kant's distinction could be revised. He might claim that

> (Q) when we accept his Formal Principle, reason always directly motivates us to act upon it. To act on any other principle, we must be motivated by some desire, and we may not have any such desire.

Kant might even allow that all acts are motivated by desires. He could then claim that

> (R) when we accept his Formal Principle, reason always produces in us a desire to act upon it. When we accept other principles, we may not have such a desire.

Since these claims are not hedonistic, they are in one way easier to defend.

Both claims raise the same questions. Does reason by itself motivate us only when we accept Kant's Formal Principle? If so, why is that true?

Kant may be right to claim that, when we act on his Formal Principle, we are motivated by reason, or by our moral beliefs. And he may be right to distinguish between this kind of motivation and some kinds of motivation by desire. But Kant's Motivational Argument requires him to distinguish between two kinds of *moral* motivation. His claim must be that, if we accept his Formal Principle, our moral beliefs motivate us in a special and uniquely reliable way. That would be so if it was only moral knowledge that had such special motivating power, and only Kant's Formal Principle was true. But as I have said, Kant's argument cannot assume that his Formal Principle is true, since that is what this argument is intended to show. For Kant's argument to support his principle, it must be the *content* of Kant's Formal Principle, not its *truth*, which gives this principle its unique motivating power. Kant must claim that, if we believe that we ought to act only on universalizable maxims, this belief necessarily motivates us. If we accept any other moral principle, our moral beliefs would not have such power.

Kant often seems to make this claim. For example, he writes:

> Only a formal law, that is, one that prescribes to reason nothing more than the form of this universal lawgiving as the supreme condition of maxims, can be a priori a determining ground of practical reason (CPR 64).

Kant's defences of this claim are surprisingly oblique. He is more concerned to show that pure reason can be practical, by determining our will. Kant takes it for granted that, *if* pure reason is practical, it moves us to act on his Formal Principle. He even writes:

> pure reason must be practical of itself and alone, that is, it must be able to determine the will by the mere form of a practical rule. . .(CPR 24)

Kant here identifies reason's being practical with its determining the will by a rule's mere form. That is a slip, since reason might move us to act on one or more substantive principles.

As this slip suggests, Kant assumes that his claim is uncontroversial. Thus, when introducing his Formula of Universal Law, Kant writes

> The most ordinary attention to oneself confirms that this idea is really, as it were, the pattern for the determinations of our will (CPR 44).

We can easily be directly aware, this remark implies, that our acceptance of Kant's formula motivates all our moral acts. That is not, however, true.

Kant's claim, as he often says, cannot appeal to empirically established psychological

laws. The Universe may contain non-human rational beings, and we have no evidence about the motivation of such beings. It must be an a priori truth that all rational beings would be motivated by Kant's Formal Principle. And for Kant's argument to succeed, there must be no such truth about any other moral principle.

There are, Kant claims, such a priori truths about the motivating power of the moral law. For example, he writes:

> we can see a priori that the moral law, as the determining ground of the will, must by thwarting all our inclinations produce a feeling that can be called pain. . . (CPR 73)
>
> the moral law. . . inasmuch as it even strikes down self-conceit, that is humiliates it, is an object of the greatest *respect*, and so too the ground of a positive feeling that is not of empirical origin and is cognized a priori. . . .

Similarly, after mentioning our

> boundless esteem for the pure moral law stripped of all advantage. . .

Kant writes

> . . . one can yet see a priori this much: that such a feeling is inseparably connected with the representation of the moral law in every finite rational being (CPR 80).

But Kant does not defend these claims, nor do they imply that the moral law must be his Formal Principle.

There are other features of Kant's view that may have led him to believe that only his Formal Principle necessarily determines the will. He may again be influenced by a failure to distinguish between his uses of the words 'material' and 'formal'. Thus Kant writes:

> all that remains of a law if one separates from it everything material, that is, every object of the will (as its determining ground), is the mere *form* of giving universal law (CPR 27).
>
> If a rational being is to think of his maxims as practical universal laws, he can think of them only as principles that contain the determining ground of the will not by their matter but only by their form.

These remarks seem to assume that, if some principle is not motivationally material, because it can motivate without the help of a desire, this principle must be normatively formal in sense 3, imposing a merely formal constraint. As I have claimed, that does not follow.

Kant may also have assumed that, since pure reason determines our will as noumenal beings in the supersensible timeless world, reason must determine our will with some principle which, because it is merely formal, has the abstract purity of that world. Consider, for example, these remarks:

> The will is thought as independent of empirical conditions and hence, as a pure will, as determined by the mere form of law. . .
>
> It is a question only of the determination of the will. . . whether it is empirical or whether it is a concept of pure reason (of its lawfulness in general) (CPR 31).
>
> Reason takes an immediate interest in an action only when the universal validity of the maxim of the action is a sufficient determining ground of the will. Only such an interest is pure. (G 460 note)

Some passages involve both these assumptions. Thus Kant writes:

> Since the matter of a practical law. . . can never be given otherwise than empirically. . . a free will, as independent of empirical conditions (i.e. conditions belonging to the sensible world). . . must find a determining ground in the law but independently of the matter of the law. . . The lawgiving form. . . is therefore the only thing that can constitute a determining ground of the will (CPR 29).

Kant here argues that, since a moral will must be free from empirical conditions, and cannot be determined by anything material, such a will must be determined by a merely Kant's Formal Principle. As before, that does not follow. Kant was inclined to group together, like opposing armies, several pairs of contrasting concepts and properties:

material	formal
empirical	a priori
pleasure-based	duty-based
heteronomous	autonomous
phenomenal	noumenal
contingent	necessary
conditional	unconditional
impure	pure

The first of these distinctions, however, is not exhaustive. Some substantive principles are not, in the senses Kant intends, either material or formal. And such principles can be a priori, duty-based, necessary, unconditional, and, in the relevant senses, pure.

When Kant rejects all 'material' moral principles, he gives no example of what is claimed by such principles, saying only that they appeal to such things as happiness, perfection, or God's commands. As we have seen, in giving some of the arguments of the Groundwork, Kant seems to overlook those substantive principles that make categorical

requirements. For Kant's Motivational Argument to succeed, however, his claims must apply to all such principles. Kant must claim that his Formal Principle differs from all such 'material' or substantive principles in being the only principle that would necessarily motivate all rational beings.

Kant could not defend this claim. Our moral beliefs do not have special motivating force if and because we derive them from Kant's Formal Principle. Compared with substantive moral beliefs---such as the beliefs that it is wrong to kill, or that we have a duty to care for our children---there is no magic in the thought that we should act only on universalizable maxims.

Kant's Motivational Argument, I conclude, cannot support his principle. Since Kant appeals to this argument so often, he seems to have found it especially convincing. It is not easy to explain why. Of Kant's reasons for believing that his Formal Principle is the supreme moral law, one seems to have been his belief that his Formal Principle has unique motivating force. But Kant, I suspect, had this second belief only because he believed that his Formal Principle is the supreme law.

4

Kant's argument is open, I believe, to other objections. This argument assumes that

> (H) no principle can be a true moral law unless its acceptance would necessarily motivate all rational beings.

As we have seen, there are two ways to defend this claim. On *the Platonic view*, moral knowledge necessarily motivates. If some moral principle is true, that gives it the power to motivate all rational beings. On Kant's view, it seems, this dependence goes the other way. Rather than assuming that a principle's truth gives it such motivating power, Kant seems to assume that

> (S) if some principle has the power to motivate all rational beings, that makes this principle a true moral law.

This view we can now call *Kant's Moral Internalism*. Remember next that, on my proposed revision of Kant's Formal Principle, acts are wrong unless they are permitted by principles whose universal acceptance *everyone* could rationally will. Though Kant appeals only to what we ourselves could rationally will, that is because he assumes that what each of us could rationally will is the same as what everyone could will. And Kant appeals to 'the idea of the will of every rational being as a will giving universal law'(432). So we can assume that Kant would accept

> (T) moral principles are true only if and because these are the principles whose

> universal acceptance everyone could rationally will.

This claim is intuitively plausible. We can see how some principle's truth may depend on its acceptability, which may in turn depend on whether we could rationally will that everyone accept this principle. Kant's Moral Internalism could instead be stated as

> (U) moral principles are true only if and because their acceptance would necessarily motivate all rational beings.

This claim is much less plausible. Why should a principle's truth depend, not on its acceptability, but on its motivating power? Kant himself writes

> Nothing is more reprehensible than to derive the laws prescribing what *ought to be done* from what *is done* (*First Critique*, A/319/B 375).

We can add, 'or from what moves us to do it'. I have rejected Kant's claim that we are autonomous, in the sense of being subject only to requirements that we give ourselves. We are subject, I believe, to several rational and moral requirements, whose truth and normative force do not in any way derive from us. But I believe that, unlike us, morality *is* autonomous in a sense that is close to Kant's. Moral requirements are not determined from outside, or by something other than morality itself. Morality's autonomy is denied by Kant's form of Moral Internalism. Rather than first asking what is good, Kant claims, we should first search for the law which determines the will of all rational beings. We can then derive, from this motivational truth, truths about what ought to be done. This heteronomous account of morality is, I believe, deeply flawed.

One way to bring that out is this. According to what Kant calls the *principle of self-love*, we ought rationally to promote our own happiness. Since Kant believes that all rational beings necessarily want their own happiness, he must agree that this principle would necessarily motivate all these beings. Given Kant's Moral Internalism, he ought to conclude that the principle of self-love is a true moral law.

Perhaps because he sees the problem I have just described, Kant rejects the principle of self-love in a way that is curiously inconsistent with his rejection of other material principles. Kant claims both that

> (V) these other principles cannot be true moral laws because it is *not* a necessary truth that all rational beings would be motivated to act upon them,

and that

> (W) the principle of self-love cannot be a true moral law because it *is* a necessary truth that all rational beings would be motivated to act upon it.

If these objections were both good, we would have to conclude that there cannot be any

true moral laws.

Neither objection, I believe, is good. Unlike (V), which assumes Kant's Moral Internalism, (W) goes to the opposite extreme. (W) assumes that, if some principle would necessarily motivate all rational beings, that *disqualifies* this principle from being a true moral law. In rejecting the principle of self-love on this ground, Kant misapplies another, less implausible view. On that other view, since the concept of *duty* is the concept of a constraint, those who would be certain to act in some way, because they had no contrary temptations, could not have a duty to act in this way. Beings who were wholly good, Kant claims, could not have any duties. This view does not imply, however, that the principle of self-love cannot be a moral law. As Kant himself points out, most of us sometimes fail to act on this principle, as when we fail to resist the temptation of some immediate pleasure, at a foreseen and greater cost to our future happiness. So Kant should not reject this principle on the ground that all rational beings would necessarily have *some* motivation to act upon it. Though Kant seems right to say that the principle of self-love is not a true moral law, he must reject this principle with some claim about its content, rather than its motivating power.

The same applies to other principles. Just as Kant should not reject the principle of self-love on the ground that its acceptance would necessarily motivate all rational beings, he should not reject other principles on the ground that their acceptance would *not* necessarily motivate all such beings.

When we ask which moral principles are true, or what is right and what is good, we should not follow Kant's proposed 'method of ultimate moral inquiry'. We should not search for some law that necessarily determines the will. Perhaps, as Platonists believe, true moral laws would necessarily motivate all rational beings. But if that were so, it would be a consequence of the truth of these moral laws, and the rationality of these beings. If moral knowledge would necessarily motivate all rational beings, that would not be because it is the power to motivate these beings which makes a principle a true moral law. Motivation is not, in that sense, prior to morality.

In some passages, Kant's Moral Internalism seems to take a more extreme, reductive form. He seems to accept

> (X) If some principle would necessarily motivate all rational beings, that does not merely make this principle a true moral law. Having such motivating power is *what it is* to be a true moral law.

This view is suggested by several of the passages quoted above. Thus, after claiming that moral laws

> must hold. . . for all rational beings as such. . .

Kant continues

> how should laws of the determination of our will be taken as laws of the determination of the will of rational beings as such. . . if they were merely empirical and did not have their origin completely a priori in pure but practical reason? (G 408)

Moral laws, Kant here suggests, are not merely the laws *that* necessarily determine the will. They are laws *of* the determination of the will. He also writes:

> the good (the law). . . which objectively, in its ideal conception, is an irresistible incentive. [900]

> . . . So here we lack the ground of duty, moral necessitation; we lack an unconditioned imperative, no coercion can be thought of here that enjoins immediate obligation (L 497).

> Such a being has no need of any imperative, for *ought* indicates that it is not natural to the will, but that the agent has to be coerced (L 605).

Ideal normativity, Kant here assumes, involves an irresistible coercive incentive. Kant similarly writes that, to prove that there are categorical imperatives, we must show

> that there is a practical law which by itself commands absolutely and without all incentives (G 425).

A law commands absolutely, this remark suggests, if this law moves us to act without the aid of other incentives. As Kant also says

> The practical rule, which is here a law, absolutely and directly determines the will objectively, for pure reason, practical in itself, is here directly law-giving (CPR 31).

Reason gives a law, Kant here assumes, by determining the will. Or consider Kant's remark that moral imperatives

> have no regard either for skill, or prudence, or happiness, or any other end that might bring the actions into effect; for the necessitation to act lies purely in the imperative alone (L 487).

Though Kant describes necessitation as the relation which is expressed by 'ought', this remark treats this relation as the bringing about of an act. Consider next Kant's claim that imperatives are categorical when they assert

> the practical necessity of the action in an absolute sense, without the motivating ground being contained in any other end (L 606).

This definition conflates normativity and motivation. Similarly Kant writes:

> Human actions. . . if they are to be moral, have need of practical imperatives, i.e. of practical determinations of the will to an action (L 486).
>
> duty. . . lies. . . in the idea of a reason determining the will by means of a priori grounds (G 408).
>
> Practical good. . . is that which determines the will by means of representations of reason. . (G 413).
>
> The concepts of *good* and *evil*. . . are. . . modi of a single category, namely that of causality. . .(CPR 65).

On such a view, I believe, normativity disappears.

I have been discussing only some of Kant's claims. Kant himself distinguishes between normativity and motivating force, as when he writes:

> Guideline and motive have to be distinguished. The guideline is the principle of appraisal, and the motive that of carrying out the obligation; in that they have been confused, everything in morality has been erroneous. (L 274)

In some passages, Kant seems to forget this warning. But consistency is not, as Kant claimed, a philosopher's greatest duty. It is more important to have ideas that take us closer to the truth.

BIBLIOGRAPHY

Anderson, Elizabeth (1991) 'Mill and Experiments in Living', *Ethics* October.

Aristotle *Nicomachean Ethics.*

Barry, Brian (1989) *Theories of Justice*, *Volume 1* (Harvester-Wheatsheaf).

_____ (1995) *Justice as Impartiality* (Oxford University Press).

Bennett, Jonathan (1974) 'The Conscience of Huckleberry Finn' *Philosophy*, Vol.49, No 188 (April).

Blackburn, Simon (1984) *Spreading the Word* (Oxford University Press).

_____(1993) *Essays in Quasi-Realism* (Oxford University Press).

_____ (1998) *Ruling Passions* (Oxford University Press).

Boyd, Richard (1997) 'How to Be a Moral Realist?', in *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard, and Peter Railton (Oxford University Press).

Brandt, Richard (1992) *Morality, Utilitarianism, and Rights* (Cambridge University Press)

Bratman, Michael.

Brink, David (2001 'Realism, Naturalism, and Moral Semantics' in Moral Knowledge, edited by Ellen Frankel Paul, Fred D Miller and Jeffrey Paul (Cambridge,) 157.

Allison Henry Allison

Broad, C.D. (1959) *Five Types of Ethical Theory* (Littlefield, Adams, and Co).

Chan, David (2004) 'Are There Extrinsic Desires?' *Nous* 38: 2

Copp, David (2001) 'Realist-Expressivism: A Neglected Option for Moral Realism', in *Moral Knowledge*, edited by Ellen Frankel Paul, Fred D Miller, and Jeffrey Paul (Cambridge University Press).

Cullity, Garrett (2004) *The Moral Demands of Affluence* (Oxford University Press).

Cummiskey, David (1996) *Kantian Consequentialism* (Oxford University Press).

Daniels, Norman (1975) *Reading Rawls*, (Blackwell).

Darwall, Stephen (1983) *Impartial Reason* (Cornell University Press).

_____ (1992) 'Internalism and Agency' *Philosophical Perspectives, Vol. 6. Ethics*.

_____ (1992B) Allan Gibbard, and Peter Railton 'Toward Fin de Siecle Ethics: Some Trends, *The Philosophical Review*, January).

_____ (1996) *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (Oxford University Press).

Dean, Richard (2006) *The Value of Humanity in Kant's Moral Theory* (Oxford University Press).

Egan, Andrew (2007) 'Quasi-Realism and Fundamental Moral Error', *Australasian Journal of Philosophy*, 85:2

Engstrom, Stephen (1992) 'The Concept of the Highest Good in Kant's Moral Theory', *Philosophy and Phenomenological Research*.

Enoch, David (2005) 'Why Idealize?' *Ethics* 115 (July).

Falk, W. D. (1950) 'Morality and Nature', *The Australasian Journal of Philosophy*.

_____(1986) *Ought, Reasons, and Morality* (Cornell University Press).

Findlay, John (1961) *Values and Intentions* (George Allen and Unwin).

Frankfurt, Harry (1988) *The Importance of What We Care About* (Cambridge University Press).

Freeman, Samuel (2003) ed.*The Cambridge Companion to Rawls*, (Cambridge University Press).

Gaut, Berys (1997) 'The Normativity of Instrumental Reason', in *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Gaut, (Oxford University Press).

Gauthier, David (1975) 'Reason and Maximization' , *Canadian Journal of Philosophy* 4:

_____ (1984) 'Deterrence, Maximization, and Rationality', *Ethics*, 94: p, 489)

_____ (1985) 'Afterthoughts' in *The Security Gamble*, ed. Douglas MacLean (Totowa, NJ: Rowman & Allanheld).

_____(1986) *Morals by Agreement* (Oxford University Press).

_____(1997) 'Rationality and the Rational Aim' in *Reading Parfit*, edited by Jonathan Dancy, (Blackwell).

Gibbard, Allan (1990) *Wise Choices, Apt Feelings*, (Oxford University Press).

_____(2006) 'Normative Properties' in *Metaethics after Moore*, edited by Terry Horgan and Mark Timmons (Oxford University Press).

_____ (2006) 'The Reasons of a Living Being' in *Foundations of Ethics* edited by Russ Shafer-Landau and Terence Cuneo (Blackwell).

_____ (2003) *Thinking How to Live*, henceforth (Harvard University Press).

Gregor, Mary (1963) *Laws of Freedom*, (Oxford University Press).

Guyer, Paul 1992 *The Cambridge Companion to Kant*, edited by Paul Guyer, (Cambridge University Press).

Guyer, Paul (2006) *Kant and Modern Philosophy* (Cambridge University Press).

Guyer, Paul (2000) *Kant on Freedom, Law, and Happiness* (Cambridge University Press).

Hare, R.M. (1952) *The Language of Morals*, (Oxford University Press).

_____ (1963) *Freedom and Reason* (Oxford University Press).

_____ (1972) 'Nothing Matters', in R.M. Hare *Applications of Moral Philosophy* (Macmillan).

_____(1981) *Moral Thinking* (Oxford University Press).

_____(1997) 'Could Kant Have Been a Utilitarian?', in R.M.Hare *Sorting Out Ethics* (Oxford University Press).

Harman, Gilbert (2000) *Explaining Value* (Oxford University Press).

Herman, Barbara (1993) *The Practice of Moral Judgment* (Harvard University Press).

Hieronymi, Pamela (2005) 'The Wrong Kind of Reason', *The Journal of Philosophy*, 102 no 9 September.

_____(2006) 'Controlling Attitudes', *Pacific Philosophical Quarterly*, 87, no 1 March.

Hill, Thomas E. (1992) *Dignity and Practical Reason* (Cornell University Press).
_____(2000) Respect, Pluralism, and Justice (Oxford University Press).

_____(2002) *Human Welfare and Moral Worth* (Oxford University Press).

Hooker, Bradford (2000) *Ideal Code, Real World* (Oxford University Press).

Hume, David's *Treatise On Human Nature*,

Hume, David *Enquiry* 1999?

Irwin, Terence (1996) 'Kant's Criticisms of Eudaemonism', in *Aristotle, Kant, and the Stoics*, edited by Stephen Engstrom and Jennifer Whiting, (Cambridge University Press).

Jackson, Frank (1992) 'Critical Notice of Hurley' *Australasian Journal of Philosophy* vol. 70.

_____ (1998) *From Metaphysics to Ethics* (Oxford University Press).

Kagan, Shelly (1998) *Normative Ethics* (Westview Press, 1998).

_____(2000) 'Evaluative Focal Points', in *Morality, Rules and Consequences*, edited by Brad Hooker, Elinor Mason, and Dale E. Miller (Edinburgh University Press).

_____(2002) 'Kantianism for Consequentialists', in *Groundwork for the Metaphysics of Morals*, Immanuel Kant, edited and translated by Allen Wood (Yale University Press).

Kahane, Guy

Kamm, Frances (2000) 'Famine Ethics: the Problem of Moral Distance and Singer's Ethical Theory' in *Singer and his Critics*, ed D. Jamieson (Blackwell).

_____ (2004) 'The new problem of distance in morality', in *The Ethics of Assistance*, edited by Deen K. Chatterjee (Cambridge University Press).

_____(2007) *Intricate Ethics* (Oxford University Press).

Kant, Immanual *Declaration concerning Fichte's Wissenschaftslehre*, 7 August 1799, *Correspondence* translated and edited by Arnulf Zweig (Cambridge University Press 1999).

Kant, Immanuel *Metaphysik* L1,28:337, From lectures given around 1778, cited in Paul Guyer *Kant on Freedom, Law, and Happiness* (Cambridge University Press, 2000) 94.

Kant, Immanuel 'What is Orientation in Thinking', VIII, 145

Kant, Immanuel *First Critique.*

Kant, Immanuel *Lectures on Ethics*, edited by Peter Heath and J.B. Schneewind, (Cambridge University Press, 1997

Kant, Immanuel *Religion within the Limits of Reason Alone*, translated by T. Greene and H. Hudson, Harper 1960,

Kant, Immanuel Second Critique.

Kant, Immanuel *The Groundwork*.

Kavka, Gregory (1986) 'The Toxin Puzzle', *Analysis*, 43.

Kemp Smith, Norman (1915) 'Kant's Method of Composing the Critique of Pure Reason', *The Philosophical Review*.

Kerstein, Samuel (2002) *Kant's Search for the Supreme Principle of Morality* (Cambridge University Press).

Kolodny, Niko (2003) 'Love as Valuing a Relationship', *The Philosophical Review*.

_____(2005) 'Why be Rational?', Mind, 2005 Volume 114.

_____(2008) 'Ought: Between Subjective and Objective' co-authored with John MacFarlane (unpublished).

Korsgaard (1986) Christine 'Scepticism about Practical Reason' *The Journal of Philosophy*, January.

_____(1996A) *Creating the Kingdom of Ends* (Cambridge University Press).

_____(1996B) *The Sources of Normativity* (Cambridge University Press).

_____ (1997) 'The Normativity of Instrumental Reason', in *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Gaut, (Oxford University Press).

_____(2003) 'Realism and Constructivism', *Philosophy in America at the Turn of the Century* APA Centennial Supplement, *Journal of Philosophical Research* (2003).

Kuehn, Manfred (2001) *Kant* (Cambridge University Press).

Leibniz (1988) *Political Writings* 2nd edition translated by Patrick Riley (Cambridge University Press, 1988).

Leslie, John (1989) *Value and Existence* (Blackwell).

Lewis, David (1985) in *The Security Gamble*, ed. Douglas MacLean (Totowa, NJ: Rowman & Allanheld).

Mackie, John (1982) *The Miracle of Theism* (Oxford University Press).

McClennen, Edward (1988) 'Constrained Maximization and Resolute Choice', *Social AlliPhilosophy and Public Policy*, *5.*

Moore, G E (1903) *Principia Ethica* (Cambridge University Press).

Mulgan, Timothy (2001) *The Demands of Consequentialism* (Oxford University Press).

Murphy, Liam (2000) *Moral Demands in Nonideal Theory* (Oxford University Press).

Nagel, Thomas (1970) *The Possibility of Altruism* (Oxford University Press).

_____(1973) in his 'Rawls on Justice', *Philosophical Review* April.

_____(1979) *Mortal Questions* (Cambridge University Press).

_____(1986) *The View from Nowhere* (Oxford University Press).

_____(1991) *Equality and Partiality* (Oxford University Press).

_____(1995) *Other Minds* (Oxford University Press).

_____(1997) *The Last Word* (Oxford University Press).

Newman, Cardinal John Henry (1901) *Certain Difficulties Felt by Anglicans in Catholic Teaching* (Longman).

Nowell-Smith, Patrick (1954) *Ethics* (Penguin).

Nozick, Robert (1974) *Anarchy, State and Utopia* (Blackwell).

_____ (1981) *Philosophical Explanations* (Oxford).

_____(1993) *The Nature of Rationality* (Princeton).

O'Neill, Onora (1975) *Acting on Principle* (Columbia University Press).

_____(1989) *Constructions of Reason* (Cambridge University Press).

_____(1996) *Towards Justice and Virtue*, 59 (Cambridge University Press),

Parfit, Derek (1986) 'Comments' in *Ethics*, Summer.

_____(2006) 'Normativity' *Oxford Studies in Metaethics Vol 1*, edited by Russ Shafer-Landau (Oxford University Press).

_____(1997) 'Reasons and Motivation', *Proceedings of the Aristotelian Society*, *Supplementary Volume*.

_____(1984-7) *Reasons and Persons* (Oxford University Press).

Pereboom, Derk (2001) *Living Without Free Will* (Cambridge University Press).

Pogge, Thomas (1998) 'The Categorical Imperative', in Kant's *Groundwork of the Metaphysics of Morals: Critical Essays*, edited by Paul Guyer (Rowman and Littlefield

_____(2002) World Poverty and Human Rights (Polity).

_____(2004) 'Parfit on What's Wrong', the *Harvard Review of Philosophy*, Spring).

Potter, Nelson (1998) 'The Argument of Kant's Groundwork', in *Kant's Groundwork of the Metaphysics of Ethics, Critical Essays*, edited by Paul Guyer, (Rowman and Littlefield).

Prichard, H.A. (1949) *Moral Obligation* (Oxford University Press).

Rachels, Stuart (2000) '"Is Unpleasantness Intrinsic to Unpleasant Experiences?" *Philosophical Studies*, Vol. 99, No. 2, May (II).

Railton, Peter (2003) *Facts, Values, and Norms* (Cambridge University Press).

Raphael, D. D. (2001) *Concepts of Justice* (Oxford University Press).

Rashdall, Hastings (1892) a review of Sidgwick's *Elements of Politics*, in the *Economic Review* 2.

Rawls, John (1971) *A Theory of Justice*, (Harvard University Press).

_____ (1989) 'Themes in Kant's Moral Philosophy', in E. Foerster (ed.) Kant's Transcendental Deductions (Stanford University Press).

_____ (1999) *Collected Papers* edited by Samuel Freeman (Harvard University Press).

_____(1996) *Political Liberalism*, (Columbia University Press).

_____(2000) Rawls *Lectures on the History of Moral Philosophy*, edited by Barbara Herman (Harvard University Press).

_____ *Justice as Fairness* (2001), Harvard University Press 106).

Raz, Joseph (2000)*Engaging Reason* (Oxford University Press).

Reid, Thomas (1983)*The Works of Thomas Reid* (Georg Olms Verlag).

Ridge, Michael (2006) 'Introducing Variable Rate Rule-Utilitarianism'**,** *The Philosophical Quarterly* (April).

Ross, Sir David (2001), *Foundations of Ethics* (Oxford University Press).

Ruskin, John (1903) *The Works*, edited by E.T.Cook and Alexander Wedderburn (London).

Scanlon, T.M. (1997) 'Contractualism and Utilitarianism', *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (Oxford University Press).

_____ (1998) *What We Owe to Each Other*, (Harvard University Press).

_____(2003) 'Rawls on Justification', in *The Cambridge Companion to Rawls*, edited by Samuel Freeman (Cambridge University Press).

_____(2003B) 'Value, Desire, and the Quality of Life', in *The Difficulty of Tolerance* (Cambridge University Press).

_____(2007) *Common Minds* edited by Geoffrey Brennan, Robert Goodin, and Michael Smith.

_____(2007B) 'Wrongness and Reasons', in *Oxford Studies of Metaethics Volume 2* edited by Russ Shafer-Landau.

_____(2001) 'Thomson on Self-Defence' in *Fact and Value*, edited by Alexy Byrne, Robert Stalnaker, and Ralph Wedgwood (The MIT Press).

Schneewind, Jerome (1977) *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford University Press).

_____ (1992) *The Cambridge Companion to Kant*, edited by Paul Guyer, (Cambridge University Press,)

_____ (1998) (Kant's *Groundwork of the Metaphysics of Morals: Critical Essays*, edited by Paul Guyer (Rowman and Littlefield).

Schroeder, Mark (2007) *Slaves of the Passions* (Oxford University Press).

Schultz, Bart (2004) *Henry Sidgwick: Eye of the Universe, an Intellectual Biography* (Cambridge University Press).

Shafer-Landau, Russ (2003) *Moral Realism* (Oxford University Press).

Shaver, Robert

Sidgwick, Henry *The Methods of Ethics* (Macmillan and Hackett various dates*).*

_____(1906) *Henry Sidgwick A Memoir*, by A.S. and E. M. S (Macmillan

_____ (2000) *Essays on Ethics and Method* edited by Marcus George Singer (Oxford University Press).

Smith, Michael (1994) *The Moral Problem* (Blackwell).

_____ (2004) *Ethics and the A Priori* (Cambridge University Press).

_____(2008) 'Desires, Values, Reasons, and the Dualism of Practical Reason' in [Ratio Volume, edited by John Cottingham].

Stratton-Lake Philip (2004) *On What We Owe To Each Other* (Blackwell).

Strawson, Galen (1994) 'The Impossibility of Moral Responsibility' *Phil.Studies* 75, 5-24.

_____ (1998) 'Free Will' in the *Routledge Encyclopaedia of Philosophy*, edited by E.Craig (London, Routledge).

Sturgeon, Nicholas (2006) 'Ethical Naturalism', in *The Oxford Handbook of Ethical Theory*, edited by David Copp (Oxford University Press) 117 note 27.

Sugden, Robert (1990) 'Contractarianism and Norms', *Ethics* 100.

Swinburne, Richard (1979) *The Existence of God* (Oxford University Press).

Temkin, Larry Temkin

Thomson, Judith (1990) *The Realm of Rights* (Harvard University Press).

_____(2003) *Goodness and Advice* (Princeton University Press) 19.

Timmons, Mark (1998) *Morality Without Foundations* (Oxford University Press,

Unger, Peter (1984) 'Minimizing Arbitrariness: Toward a Metaphysics of Infinitely Many Isolated Concrete Worlds', *Midwest Studies in Philosophy* IX reprinted in Peter Unger *Philosophical Papers Volume 1*.

Williams, Bernard (1981) 'Internal and External Reasons', in *Moral Luck* (Cambridge University Press).

Williams, Bernard (1981B) *Moral Luck* (Cambridge University Press).

_____(1985) *Ethics and the Limits of Philosophy* (Fontana).

_____(1995) 'Internal Reasons and the Obscurity of Blame' in his *Making Sense of Humanity* (Cambridge University Press).

_____(1995B) *Making Sense of Humanity* (Cambridge University Press).

_____ (1995B) *World, Mind, and Ethics*, edited by J.E.J.Altham and Ross Harrison (Cambridge).

_____(2003) *The Sense of the Past* (Princeton University Press).

_____(2006) *Philosophy as Humanistic Discipline* (Princeton University Press).

Williams, T. C. (1968)*The Concept of the Categorical Imperative* (Oxford University Press).

Wodehouse, P.G. (1952) *Pigs Have Wings* (Ballentine Books).

Wood, Allen (1999) *Kant's Ethical Thought* (Cambridge University Press).

Wood, Allen (2002) 'What is Kantian Ethics?' in *Groundwork for the Metaphysics of Morals*, Immanuel Kant, edited and translated by Allen Wood (Yale University Press).

_____ (2006) 'The Supreme Principle of Morality', in *The Cambridge Companion to Kant and Modern Philosophy*, edited by Paul Guyer (Cambridge University Press).

_____(2002) *Groundwork for the Metaphysics of Morals*, translated by Allen Wood, (Yale University Press).

_____(2008) *Kantian Ethics* (Cambridge University Press).

340,202 without notes, 375,000 with notes.

---

[1] In these opinions I follow Broad (1959) 143-4.

[2] *Declaration concerning Fichte's Wissenschaftslehre*, 7 August 1799, Immanuel Kant, *Correspondence* translated and edited by Arnulf Zweig (Cambridge University Press 1999) 560.

[3] Sidgwick (1906) 284.

[4] Sidgwick is referring here to another of his books, but he would have applied this claim, I believe, to his *Methods*.

[5] I suggest that we can ignore Book 1, Chapter II, Book II chapter VI, and Book III, Chapter XII.

[6] For discussions of Sidgwick, however, see Jerome Schneewind's outstanding *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford University Press, 1977). See also and Bart Schultz's fascinating *Henry Sidgwick: Eye of the Universe, an Intellectual Biography* (Cambridge University Press, 2004).

[7] Sidgwick (2000) xxviii.

[8] Sidgwick (1906) 396.

[9] Sidgwick (1906) 92.

[10] Sidgwick (1906) 170-1.

[11] For example, in the first edition, Sidgwick writes: 'A and B are supposed to see that the happiness of a community will be enhanced . . by a little of what is commonly blamed as vice, along with a great deal of what is commonly recommended as virtue: and convinced that others will supply the virtue, A and B think themselves justified, on Utilitarian grounds, in supplying the vice' (ME First Edition 451). In later editions, 'vice' became 'irregularity'.

[12] ME 298-9 (my italics).

[13] ME First Edition (1874) 473. When a friend remarked that Sidgwick should be proud of his great book, Sidgwick replied 'The first word in my book is "Ethics" and the last is "failure".'

[14] ME 507 note.

[15] ME 501. Characteristically, Sidgwick adds this note: 'I do not think, however, that we are justified in stating as *universally* true what has been admitted in the previous paragraph. Some few thoroughly selfish persons appear at least to be happier than most of the unselfish; and there are other exceptional natures whose chief happiness seems to be derived from activity, disinterested indeed, but directed towards other ends than human happiness.'

[16] Sidgwick (2000) 118.

[17] ME 295.

[18] ME 437.

[19] ME 248 note.

[20] ME 284.

[21] ME 490.

[22] *Mind*, 1877 125-6, quoted in Schultz, 349.

[23] Sidgwick (1906) 74.

[24] Bernard Williams, *The Sense of the Past* (Princeton University Press, 2003) 283. This sentence continues 'which is no doubt part of what Bloomsbury found oppressive and stuffy'.

[25] ME 359. The end of this passage reads: 'And if we consider the matter in its relation to the individual's perfection, it is certainly clear that he misses the highest and best development of his emotional nature, if his sexual relations are of a merely sensual kind: but we can hardly

know a priori that this kind of relation interferes with the development of the higher (nor indeed does experience seems to show that this is universally the case). And this latter line of argument has a further difficulty. For the common opinion that we have to justify does not merely condemn the lower kind of development in comparison with the higher, but in comparison with none at all. Since we do not positively blame a man for remaining celibate (though we perhaps despite him somewhat unless the celibacy is adopted as a means to a noble end): it is difficult to show why we should condemn----in its bearing on the individual's emotional perfection only---the imperfect development afforded by merely sensual relations.'

[26] M 421.

[27] Broad (1959) 144.

[28] Rashdall (1892).

[29] Broad (1959) 14.

[30] He should not have claimed, I believe, that there is a single fundamental normative concept (ME 32), since he takes to be equally fundamental the concepts of what we ought morally to do, and of what we have most reason to do, both from an impartial point of view, and from our own point of view. When he was a student, Sidgwick wrote 'I will not stir a finger to compress the world into a system' (M 108). But he later came too close to doing that. While defending Hedonism, Sidgwick writes: ' If we are not to systematise human activities by taking Universal Happiness as their common end, on what other principles are we to systematise them?'(ME 406). He should not have assumed that we *are* to systematise these activities. Sidgwick is mistaken, I believe, to reject all but one trivial principle of distributive justice (ME 416-7). And he makes some claims, that are simply false, as when he writes, 'I think that a "plain man", in a modern civilized society, if his conscience were fairly brought to consider the hypothetical question, whether it would be morally right for him to seek his own happiness on any occasion if it involved a certain sacrifice of the greater happiness of some other human being----without any counterbalancing gain to anyone else---would unhesitatingly answer in the negative' (ME 382).

[31] O'Neill *(*1989) 126.

[32] Kemp Smith (1915) 531.

[33] *Religion within the Bounds of Bare Reason*, 72.

[34] Quoted in John Rawls *Lectures on the History of Moral Philosophy* (Harvard University Press, 2000) xvii. Rawls charmingly adds: 'I always took for granted that the writers we were studying were much smarter than I was. If they were not, why was I wasting my time and the students' time by studying them?' (xvi). Since philosophy makes progress, we can now see 'plain mistakes' made by people who were very much smarter than us.

[35] Kemp Smith (1915) 527. Though this is a remark about Kant's *First Critique*, it also applies, I believe, to Kant's books on ethics.

[36] CPR 24.

[37] *Lectures* 127 and 148.

[38] *Lectures* 369.

[39] *Lectures* 44.

[40] Christine Korsgaard, in Korsgaard (1996A) 126.

[41] MM 429-30.

[42] Sidgwick (1906) 177.

[43] We should ignore the outbursts of some other great, passionate writers, such as Ruskin's contemptuous remarks about Palladio's Venetian churches. The *Redentore* Ruskin calls 'a mean, contemptible suburban church'. Discussing *San Giorgio*, he writes, 'It is impossible to conceive a design more gross, more barbarous, more childish in its conception, more servile in plagiarism, more insipid in result, more contemptible under every point of rational regard'. Ruskin (1903) xi. 381.

[44] Sidgwick (1906) 151. Sidgwick's remark is about Kant's terminology. But he continues 'we must go back to Kant and begin again from him. Not that I feel prepared to call myself a Kantian, but I shall always look on him as one of my teachers'.

[45] *Lectures* 18.

[46] See Nagel's wonderful *The View from Nowhere* (Oxford University Press, 1986), especially Chapter VIII, and his *The Last Word* (Oxford University Press, 1997).

[47] If we ask 'Is that true?', either answer would be astonishing. There are not many questions of which that could be claimed.

[48] I follow T. M. Scanlon, *What We Owe to Each Other*, (Scanlon 1998) Chapter 1. Reasons can be claimed to be provided, not only by facts, but also by things in other categories, such as mental states, or properties. Two examples are the claims that our desires give us reasons, and that an act's wrongness gives us a reason not to do it. But all such reasons could be redescribed as being provided by certain facts, such as certain facts about our desires, or about the wrongness of some act.

[49] When we claim that we have *more reason*, or *most reason* to act in some way, we use the word 'reason', not as a *count noun*, like 'lake' or 'cow', which refers to particular reasons, but as a *mass term*, like 'water' or 'beef', which refers to some reason or set of reasons without distinguishing between these reasons. Similar remarks apply to the claim that we have *sufficient reason* or *decisive reason* to act in some way.

[50] Like the concept of *a reason*, the concepts *should*, and *ought* cannot, I believe, be helpfully defined. It might be suggested that, when we say that we *should* or *ought* to do something in the decisive-reason-implying senses, we *mean* that we have decisive reasons, or most reason, to act in this way. This definition is fairly plausible. But when I claim that we ought to do what we have decisive reasons to do, my use of 'ought' seems to be adding something. Some writers suggest that we can explain the concept of *a reason* by appealing to the decisive-reason-implying concept *ought*. I doubt whether this explanation would succeed. But even if these concepts are both indefinable, they are very closely related, in ways that do something to explain them both. We can partly *identify* these versions of the concepts *should* and *ought* by saying that these concepts apply to some act *just when, and because*, we have decisive reasons, or most reason, to act in this way.

[51] The word 'rational' can also be used more thinly, to mean 'not irrational or open to any rational criticism'. Some act of ours might be in this sense rational, though we have no beliefs whose truth would give us reasons to act in this way, if we also have no beliefs whose truth would give us reasons *not* to act in this way.

[52] See Kolodny (2005).

[53] Motivating reasons can be acceptably regarded in two ways. On the psychological account, motivating reasons are beliefs. On the non-psychological account, these reasons are *what* we believe. When we truly believe that we have some reason, and we act for this reason, the non-psychological account is more natural. In my example, if I were asked why I don't eat walnuts, it would be more natural to reply 'Because they would kill me'. But if I later learnt that my doctor was mistaken, since walnuts wouldn't kill me, this reply would be misleading, so I would instead say 'Because I believed that they would kill me'. We might also describe some motivating reason either as what we wanted to achieve, or as our desire or aim. If asked why I don't eat walnuts, I might say either 'To avoid killing myself', or 'Because I want to stay alive'.

We need not choose, I believe, between the psychological and non-psychological accounts, since we can use them both. The acceptability of both accounts can, however, cause confusion. On one account, motivating reasons are the true or apparent normative reasons which are *what* we believe when these beliefs explain our decisions and our acts. On the other account, motivating reasons are motivating states of mind. Since motivating reasons

can thus be regarded both as normative reasons and as motivating states, that may suggest that normative reasons are motivating states. That, I believe, is a grave mistake.

[54] Some object that this definition is too wide. Michael Slote, for example, said: 'If I am looking for examples of bad books for a bad book hall of fame, I am going to reject a good book. . . In that case, won't you have to call it bad too?' But this objection is not, I believe, good. A good book would a bad example of a bad book, and a bad choice for this hall of fame.

[55] Wodehouse (1952) 93.

[56] Scanlon (1998) 97. Scanlon calls this the *buck-passing view*.

[57] Though Scanlon claims that goodness and badness are not reason-giving properties, he sometimes mentions what I call *derivative* reasons. Scanlon writes, for example, 'There can be more than one reason to respond to a human being who is in pain: his pain is bad, and we may owe it to him to help him relieve it' (1998) 181). The source of this reason, he would agree, are the features which make this pain bad.

[58] We can also note that an *agent's* point of view is not, in my sense, impartial. Even when my acts would affect people who are all strangers to me, my acts would involve *me*, and I am not a stranger to myself. Most of us believe that certain acts would be wrong even if they would make things go best in the impartial-reason-implying sense. We might believe, for example, that it would be wrong for me to violate one person's rights, even if I would thereby prevent several other people from acting wrongly, by violating several other people's rights. On these assumptions, we would all have reasons, from an impartial point of view, to want me to act in this way, since fewer people would then wrongfully violate other people's rights. But we might believe that, though I would have impartial reasons to want myself to act in this way, I would have stronger *person-relative* reasons to want *not* to act in this way, and to refrain from doing so. These other reasons would be given either by this act's wrongness or by the facts that make it wrong, or both.

[59] There are other kinds of desire, as when we want some x-ray to be clear because of what that would signify.

[60] This phrase is, in a way, misleading, since these reasons are provided, not by the value of these outcomes, but by the facts that make them good or bad. But it would be pedantic to use some more accurate phrase, such as 'value-corresponding', this phrase would not apply well to objective theories, and it is better to use a phrase that covers those objective theories which do claim that these reasons are provided by the value of these outcomes.

[61] Korsgaard (1996A) 225.

[62] We are here appealing to the normative reasons which have become our motivating or belief-producing reasons. I follow Scanlon (1998) 18-22.

[63] Kant should not have claimed that our responses to such reasons must be voluntary and free (G 448).

[64] These state-given reasons to have desires are, we should note, quite different from desire-based subjective reasons. For example, like object-given reasons to have desires, but unlike desire-based reasons, these state-given reasons to have desires are value-based. (We might be claimed to have desire-based state-given reasons to have some desire when, and because, we want to have it.)

[65] In these remarks I partly follow Korsgaard(1996A) Chapter 8.

[66] For a different view, see Rachels (2000).

[67] Korsgaard (1996A) 262.

[68] Korsgaard (1996A) 284.

[69] I discuss these and other attitudes to time in Sections 62-70 of Parfit (1984-7). (In that rather tortuous discussion I failed to make it clear that, in my view, the most rational attitude is temporal neutrality.)

[70] Though we can truly claim that I want to climb this ladder, some people claim that it would be better to drop the concept of an instrumental desire. See, for example, Chan (2004).

[71] There is another way to describe such cases. Rather than saying that we have no *reason* to fulfil such desires or aims, Subjectivists might claim that we *could not* fulfil them. If you do not deserve to suffer, my hurting you would not give you what you deserve, and if you have not injured me, I have nothing to avenge. All telic desires, Subjectivists might say, should be regarded as implicitly taking a conditional form. What we really want is that something will happen *if* certain facts are as we believe them to be. If these facts are *not* as we believe, such desires could not be fulfilled, and this could be why they provide no reasons for acting.

[72] Korsgaard (1996A) 317.

[73] Williams (1985) 19.

[74] As before, I follow Scanlon (1988) Chapter 1. See also Raz (2000) Chapter 2.

[75] As Scanlon writes: 'if having a desire involves seeing something other than that desire as providing a reason, this may explain the plausibility of the idea that desires provide reasons. It is true that having a desire involves taking oneself to have a reason. The mistake lies in confusing the reason with one's taking it to be a reason.' Reference, and Stephen Schiffer [reference].

[76] I follow Scanlon (2003B).

Similar claims apply to cost-benefit analyses. These calculations can rightly appeal to people's preferences, without thereby assuming a desire-based theory about reasons. See Scanlon's remarks in his

[77] Nozick (1993) 176.

[78] Reference to Bratman.

[79] Kant's *Lectures*, 58-9 (Prussian edition 27: 264-5), and Sidgwick's ME 74-5.

[80] For an excellent account of such reason-giving facts, see Kolodny (2003).

[81] Williams (2006) 111.

[82] Williams draws this distinction in his 'Internal Reasons and the Obscurity of Blame' in Williams (1995) 36-7. This distinction uses 'substantive' in a sense that differs from the sense I defined above. When these Subjectivists make claims about *procedural* rationality, these claims are substantive in that other sense.

[83] Rawls (1996) 49.

[84] Michael Smith 'Desires, Values, Reasons, and the Dualism of Practical Reason', in . . . edited by John Cottingham. . . . I here summarize the principle that Smith calls 'R2'.

[85] op. cit. note 4.

[86] Smith (2004) 269-70.

[87] This claim is made, for example, by Richard Hare and Richard Brandt.

[88] See Guy Kahane, [reference to his thesis].

[89] Our reason to *have* this desire would be a reason of the *state-given* kind whose claim to be reasons I question in Section 4 and Appendix B. If Subjectivists agree that there are no such reasons, they could still claim that we might have desire-based subject-given reasons to want to have, and to cause ourselves to have or to keep, some desire.

[90] Even Hume claimed only that such desires or preferences would not be *contrary* to reason.

[91] There are other objections to these theories. Consider my *whimsical Despot* in Appendix B, who threatens that I shall be tortured unless, at noon tomorrow, I have the aim of being tortured. According to aim-based subjective theories, since I now have the aim of *not* being tortured, I would have an aim-based reason to achieve this aim by causing myself to have the

aim of *being* tortured. But if I succeed in causing myself to have *this* aim, that would give me an aim-based reason to achieve this aim by causing myself to have the aim of *not* being tortured. And if I succeed in causing myself to have *this* aim, that would give me an aim-based reason to achieve this aim by causing myself to have the aim of *being* tortured, and so on for ever. We have no reason to believe in this unending spiral of aim-based reasons.

Objective theories, in contrast. . .

[92] For similar objections to theories of this kind, see Enoch (2005).

[93] Korsgaard (1996A) 261 (the words I omit are 'apparently ontological', since that is not the issue here). If we don't have to assess the things we are choosing, it is not clear that our choices deserve to be called rational.

[94] We ought, I have argued, to reject all subjective theories. We can next briefly consider a *hybrid* theory. On this view, for us to have a reason to try to fulfil some desire, we must have some value-based object-given reason to have this desire. What we want must be in some way good, or worth achieving. But when our desires are in this way rational, our having these desires would give us further reasons to try to fulfil these desires. And when we must choose between equally good possible aims, our desires or preferences can break ties, by giving us reasons to adopt one of these aims. If that were true, our desires or preferences would often break ties, since there are often no precise truths about the relative strength of conflicting object-given reasons.

I believe, though not very strongly, that we ought to reject even this hybrid theory. When we have certain desires, this fact may make it true that we have further reasons to try to fulfil these desires. But these further reasons would be provided, I believe, not by the fact that we would be fulfilling these desires, but by various other facts which causally depend on our having these desires. I described some such facts near the end of Section 7; and there are others. Though I believe that we should reject this hybrid theory, my arguments against pure subjective theories may not show that we should reject this theory. This question would then remain open. But this question does not, I believe, have much importance, since this hybrid theory is fundamentally objective and value-based.

[95] As these remarks imply, this impartial-reason-implying sense of 'best' has no connection with 'impartial observer' accounts either of the goodness of outcomes, or of morality. These accounts define what is right, or what is impersonally best, as what an impartial observer *would in fact* prefer. Such accounts seem to me worthless. If we claim only that this observer has an impartial point of view, we cannot assume that all such observers would have the same preferences. If we add psychological assumptions, we may be able to work out what this observer would prefer, but our conclusions would have no importance.

[96] Rawls (1971) 395.

[97] Rawls (1971) 417.

[98] As Sidgwick notes in *The Methods of Ethics*, henceforth *ME* 112. Rawls claims that, in giving this definition, he is following Sidgwick. But though Sidgwick suggests a similar definition, and claims that it has some merits, he then rejects it, in part because it isn't normative. Sidgwick defines his good as 'what I should practically desire if my desires were in harmony with reason, assuming my own existence alone to be considered'. (In an earlier edition, Sidgwick refers to 'the ultimate end or ends *prescribed* by reason as what *ought* to be sought or aimed at' (ME 5th edition 112) my italics.) Reference to Crisp and Shaver?

[99] Rawls (1971) 408 (my italics).

[100] Rawls (1971) 184-5, RE 161.

[101] Rawls (1971) 432.

[102] Rawls (1971) RE 491.

[103] Rawls (1971) 401. Cf 111 and 451.

[104] In Harry Frankfurt's words, we 'need to understand what is *important*, or, rather, what is *important to us*' Frankfurt (1988) 81, and 91 note 3.

[105] That is true even of some writers who claim to be questioning desire-based subjective theories. Robert Nozick, for example, makes twenty three proposals about how we should go beyond a purely instrumental, desire-based account of rationality (Nozick (1993)) Chapter V.) None of these proposals include the idea that we might have reasons to have our desires that are given by the intrinsic features of their objects, or what we want.

[106] This argument is suggested, for example, by Williams's remarks in 'Internal and External Reasons' (Williams (1981) 102 and 106-7, and in 'Internal Reasons and the Obscurity of Blame'(Williams (1995) 39. For a longer discussion of such arguments, see my 'Reasons and Motivation' (Parfit (1997).

[107] A reference here to Nyholm's ideas?

[108] Darwall (1992) 168. (Darwall's sentence continues 'perhaps when the agent's deliberative thinking is maximally improved by natural knowledge.') Darwall's claim seems an overstatement, since these Metaphysical Naturalists might describe some kinds of normativity in rule-involving or attitudinal terms. But Darwall may be right to assume that, when these people discuss reasons, their most plausible move is to identify normative and motivating force.

[109] Stephen Darwall, Allan Gibbard, and Peter Railton, in Darwall (1992B) and Darwall (1996) 176-7.

[110] There are also *non*-reductive desire-based subjective theories about reasons. These are the theories that I have mainly been discussing. But many people accept desire-based theories, I suggest, because that allows them to regard normativity in a reductive, naturalist way as some kind of motivating force. (There are also some naturalists who reject desire-based theories about reasons. Some of these people might claim to be describing value-based objective reasons. But these theories are not in my sense value-based, since these people deny that there are irreducibly normative truths. There are some other naturalists who agree that natural facts could not be normative. These *non-cognitivists* believe that, to preserve the normativity of our normative claims, we should not regard these claims as intended to be true.

[111] These people would reject this description, since they would claim that normative reasons *are* certain causes of behaviour. Reductive views are hard to describe in a neutral way.

[112] In these remarks, I follow Nagel (1986) and (1997).

[113] For such desires to be justified by such beliefs, they must also be caused by these beliefs in the right way: one that does not involve *deviant causal chains*. We need not here discuss what such deviance would involve.

[114] I am distinguishing here between some desire itself and someone's having this desire. If you and I both want Venice to be saved from the rising sea, we have the same desire, but my having this desire is not the same as your having it. In this example, we both want the same event. When we want different events, we may still have what is in a wider sense the same desire. That would be true, for example, if we are playing chess, and we both want to win. There is a similar distinction between some belief itself and someone's having this belief. The words 'desire' and 'belief' are ambiguous, since they can refer either to some desire or belief itself, or to someone's having this desire or belief. I shall sometimes say which of these I mean. But I shall often mean both, and these slippery distinctions can often be ignored.

[115] Hume writes that though desires cannot be strictly 'contrary to reason', they are, in a loose sense, 'unreasonable' when they are 'founded on false suppositions'. Hume's *Treatise*, Book II, Part III, Section III.

[116] These claims do not apply to at least one important, partly normative belief. For many of our acts to be rational, we must believe or assume that there are unlikely to be facts unknown to us that give us decisive reasons not to act in these ways. In many cases, it is irrelevant whether this belief is true or rational. That depends on *why* we believe that there are

unlikely to be such unknown reason-giving facts. (There are, I assume, other exceptions to these claims.)

[117] As Scanlon argues (1998) Chapter 1.

[118] Scanlon (1998) 29-31.

[119] This view is not implausible because we can have other reasons for having such a *discount rate*, caring less about events that are more remote. Our beliefs about such events are often less likely to be true, so that such predicted events are less likely to occur. It is often less urgent to try to produce or prevent such more remote events. And it may not be irrational to have a discount rate, not with respect to distance in time, but with respect to the degree of psychological connectedness between ourselves as we are now and ourselves at different future times. None of these, however, is a discount rate with respect to time itself. For a further discussion my discussion in Parfit (1984-7), sections 63 to 70, 102 to 105, and Appendix F.

[120] Scanlon (1998) 25-30. Add comments on Broome.

[121] For some examples, see Appendices B and C.

[122] Explain the word 'transitive'.

[123] Comment on contrary claims and arguments made by Temkin and Rachels.

[124] Though Sidgwick calls Egoism one of 'the *Methods of Ethics*', he is discussing a view about what he calls 'the *rational* end of conduct for each individual' (ME xxviii, my italics).

[125] ME, Concluding Chapter. This is only part of Sidgwick's view. Sidgwick makes other claims, to which I shall turn in Section 16.

[126] In Sidgwick's words, 'It would be contrary to Common Sense to deny that the distinction between any one individual and any other is real and fundamental, and that consequently 'I' am concerned with the quality of my existence as an individual in a sense, fundamentally important, in which I am not concerned with the quality of the existence of other individuals: and this being so, I do not see how it can be proved that this distinction is not to be taken as fundamental in determining the ultimate end of rational action for an individual' (ME 498).

[127] Findlay (1961) p 294. Compare Rawls's claim: 'Utilitarianism does not take seriously the distinction between persons' (Rawls (1971) 27). This fact also gives us reasons to accept principles of distributive justice. Given Sidgwick's belief that the distinction between persons is fundamental and of great normative significance, it is somewhat surprising that he gave so little weight to principles of distributive justice, allowing the principle of equality only to break ties.

[128] In Nagel (1986) especially chapters VIII and IX, and Nagel (1991) Chapter 2.

[129] For example, Sidgwick writes of 'the inevitable twofold conception of a human individual as a whole in himself, and a part of a larger whole. There is something that it is reasonable for him to desire, when he considers himself as an independent unit, and something again which he must recognize as reasonably to be desired, when he takes the point of view of a larger whole' (Third Edition of ME, p 402, quoted in Schneewind (1977) 369.) Nagel calls 'the transcendence of one's own point of view. . . the most important creative force in ethics' (Nagel (1986), 8).

[130] In Sidgwick's words, 'the good of any one individual is of no more importance, from the point of view. . . of the Universe, than the good of any other. . . And. . . as a rational being I am bound to aim at good generally. . . not merely at a particular part of it' (ME 382).

[131] ME 508.

[132] For a discussion of these reasons, see Kolodny (2003).

[133] Nagel (1986) 160.

[134] Some people would add 'unless this person deserves to be in pain'.

[135] There are other, more complicated cases. Suppose that we could either (1) save some stranger from ten hours of pain, or (2) save ourselves from two hours of pain, or (3) do what would both save the stranger from five hours of pain and save ourselves from one hour of pain. Though (3) would be neither impartially best nor best for ourselves, wide value-based views would imply that, as a compromise, we could rationally do (3).

[136] Nagel (1986) 161.

[137] It might be objected that, if I am moved not only by concern for this stranger's well-being but also in part by the fact that my act would be generous and fine, my motivation is not ideal. In Williams's phrase, I would be like someone who is moved, not by his great love for Isolde, but by his conception of himself as a great Tristan (Williams (1981B) 45). But if some act is generous and fine, that gives us *some* reason to act in this way.

[138] Jefferson McMahan suggests that, if my act would be generous and fine, this act would make things go impartially better, not causally, but by being in itself good. If that is true, we could suppose that, since I am younger than this stranger, my death would be a greater loss, so that, on balance, I would not have stronger impartial reasons to save this stranger.

[139] ME 386 note 4.

[140] Reid (1983) 598. Reid may not be committed to this view, since he believed that we did not face this dilemma.

[141] ME 508. According to what I earlier called Sidgwick's *Dualism of Practical Reason*, we could rationally do either what would be impartially best or what would be best for ourselves. Sidgwick does not distinguish these versions of his 'Dualism', because he believes that our duty is always to do what would be impartially best. My description of Sidgwick's view goes beyond what he actually writes. Sidgwick makes some remarks which suggest that, in cases in which duty and self-interest conflict, reason would tell us nothing. But suppose that, in such a case, there was some third possible act, which would both be wrong and be bad for ourselves. Sidgwick would surely have believed that reason would tell us not to act in this third way. His view would be only that, though this third act would be irrational, we could rationally act in either of the other ways.

[142] ME First Edition (1874) 473. Since Sidgwick cut this passage from later editions, it is worth quoting in full: 'But the fundamental opposition between the principle of Rational Egoism and that on which such a system of duty is constructed, only comes out more sharp and clear after the reconciliation between the other methods. The old immoral paradox, 'that my performance of Social Duty is good not for me but for others', cannot be completely refuted by empirical arguments: nay, the more we study these arguments the more we are forced to admit that, if we have these alone to rely on, there must be some cases in which the paradox is true. And yet we cannot but admit with Butler that it is ultimately reasonable to seek one's own happiness. Hence the whole system of our beliefs as to the intrinsic reasonableness of conduct must fall, without a hypothesis unverifiable by experience reconciling the Individual with the Universal Reason, without a belief, in some form or other, that the moral order which we see imperfectly realized in this actual world is yet actually perfect. If we reject this belief, we may perhaps still find in the non-moral universe an adequate object for the Speculative Reason, capable of being in some sense ultimately understood. But the Cosmos of Duty is thus really reduced to a Chaos: and the prolonged effort of the human intellect to frame a perfect ideal of rational conduct is seen to have been foredoomed to inevitable failure'.

[143] Rawls (1971) 575.

[144] This is forcefully argued, for example, by Kolodny (2005), Scanlon (2007), and Broome

[145] This sense could have two versions, one appealing to the evidence of which we are actually aware, the other to the evidence that is available in the sense that we could have made ourselves aware of this evidence.

[146] Some Act Consequentialists, for example, claim that (1) it is always wrong to fail to do what would *in fact* make things go best. Others claim that (2) it is always wrong to fail to do what we *believe* would make things go best. Of these claims, (1) assumes that the ordinary sense

of 'wrong' is the fact-relative sense, and (2) assumes that this ordinary sense is the belief-relative sense.

[147] See, for example, Nagel's 'Moral Luck' in Nagel (1979).

[148] There is a different way in which we can be more blameworthy if our act, though not wrong in the belief-relative sense, is wrong in the evidence-relative sense. We might be blameworthy for our negligence in failing to look at the available evidence.

[149] See Bennett (1974) 123-134.

[150] Rather than talking of the expectable goodness of these outcomes, many people talk of their *expected* goodness. That word seems misleading (as it would be to replace 'desirable' with 'desired').

[151] Even in cases of the simplest kinds, Expectabilists need not assume that the expectable goodness of outcomes depends only on the expectable sum of benefits. As Broome and Kamm suggest, it may also matter, for example, how these benefits, or people's chances of getting these benefits, are distributed between different people. In life-saving cases that involve many people, for example, it might be better if everyone were given equal chances of being saved, even if slightly fewer people would then be saved. And there may be cases in which we should be risk-averse, giving greater weight to avoiding the worst outcomes.

[152] ME 207-8.

[153] This sense of 'mustn't-be-done' differs from the non-moral decisive-reason-implying sense of 'mustn't', as used in the claim that you mustn't touch some live electric wire. (There might, we can add, be more than one indefinable sense of 'wrong', with different such senses being used by different people. This possibility I shall here ignore.)

[154] Though Sidgwick called Egoism one of the 'Methods of Ethics', he was using 'Ethics' in a wider sense, which covers all claims about what we have reason to do.

[155] ME 382-3.

[156] ME 200, 403.

[157] Rather than claiming that we ought to maximize happiness, these Utilitarians might claim that we ought to minimize suffering, or more precisely to minimize the sum of suffering minus happiness. These ways of acting are the same, just as minimizing net losses is the same as maximizing net profits. But by telling us to minimize suffering, these Utilitarians would remind us of the most effective way of trying to make the lives of sentient beings go better. And this statement of their view better expresses what makes it plausible. On this view's Buddhist version, the two great virtues are insight and compassion.

[158] Sidgwick, for example, writes: 'there are several different ways in which a Utilitarian system of morality may be used. . . (1) it may be presented as practical guidance to all who choose 'general good' as their ultimate end, whether they do so on religious grounds, or through the predominance in their minds of impartial sympathy, or because their conscience acts in harmony with Utilitarian principles, or for any combination of these or any other reasons; or (2) it may be offered as a code to be obeyed not absolutely, but only so far as the coincidence of private and general interest may in any case be judged to extend; or again (3) it may be proposed as a standard by which men may reasonably agree to praise and blame the conduct of others, even though they may not always think fit to act upon it' (Sidgwick (2000) 607. In this passage, Sidgwick's (1) seems to suggest Impartial-Reason Consequentialism.

[159] Even if 'right' and 'wrong' had only one meaning, we should accept the distinction I have drawn between the fact-relative, evidence-relative, belief-relative, and moral-belief-relative senses of these words. (These senses can be claimed to be different applications of the same meaning, as is shown by the fact that my definitions of these senses all use the word 'wrong' in the same sense.)

[160] Nor would this question have much theoretical importance. Suppose that, by acting in some way, I could save someone's life, or relieve someone's pain. These facts would give me reasons to act in these ways. But it is not worth asking, I believe, whether these would be *moral* reasons. If I had no reasons *not* to act in these ways, these facts might make it true that I ought morally to act in these ways. But that is not enough to show that we ought to think of these facts as giving me moral reasons. (Here is a similar point. Some people claim that, if some earthquake killed many people, this event would be morally bad. But there is no need to make that claim. We can believe that this event is bad in the impartial-reason-implying sense. From an impartial point of view, we all have reasons to want people not to be killed in such natural disasters. We can regard such events as in this sense bad, without considering any distinctively moral questions.)

[161] Scanlon (1998), 97.

[162] Rawls (1971) 52, Nagel (1995) 182.

[163] *The Groundwork*, henceforth *G*, 392. Page references are to the page numbers of the Prussian Academy edition, which are given in most English translations.

[164] In Kant's words: 'the human being and in general every rational being exists as an end in itself, not merely as a means to be used by this or that will at its discretion; instead he must in all his actions, whether directed to himself or also other rational beings, always be regarded at the same time as an end' (*G* 428-9).

[165] G 430.

[166] Korsgaard (1996A) 139.

[167] O'Neill (1989) 111.

[168] Korsgaard (1996A) 140.

[169] I here follow Korsgaard (1996A) 295-6. (Korsgaard does not herself believe that deception is always wrong.)

[170] After saying that the person whom he deceives 'cannot possibly consent to my way of treating him', Kant refers to this remark as having introduced what he calls 'the principle of other human beings' (G 430). (A) is the simplest statement of this principle.

[171] O'Neill (1989) 110.

[172] Korsgaard writes: 'the other person is unable to hold the end of the very same action because the way you act prevents her from choosing whether to contribute to the realization of that end' (Korsgaard (1996A) 138-9).

[173] Other writers have assumed or claimed that this is what Kant means. See, for example, Hill (1992) 45.

174 That seems often true, for example, when Kant claims that we could not will that some maxim be a universal law.

[175] G 429-30, my italics.

[176] Rawls (2000) 100-91. A similar claim is made in Hill (1992) 45.

[177] G 436.

[178] Rawls (200) 191 and 182-3.

[179] CPR, note on p.8.

[180] Herman (1993) vii.

[181] The Consent Principle would also imply that it would be wrong for me not to save White's life, since White could not rationally consent to my failing to save her life. So this principle would mistakenly imply that I cannot avoid acting wrongly.

[182] Things might be different if White was old and Grey was a young professional dancer. White might then have sufficient reasons to consent to our saving Grey's leg rather than White's life, since White's loss might here be no greater than, or even much less than, Grey's.

This is the kind of morally relevant further fact that, in considering my examples, we should suppose would not obtain.

[183] This argument was suggested to me by Ingmar Persson.

[184] In some cases, it is not enough to appeal to the claim that someone does in fact consent, since people can be under various pressures which remove the legitimating force from their consent. And even when that is not true, it may be important whether this consent be fully informed and procedurally rational, and sufficiently stable. That is the kind of consent that is rightly required, for example, by those laws which permit doctors to help their patients to commit suicide.

[185] The Consent Principle appeals to what we could rationally choose, if we knew the relevant facts. In this example, these facts would include the wrongness of this way of saving Blue's life. In asking whether Blue could rationally consent to my failing to act in this way, we need not know whether Blue believes that this act would be wrong.

[186] I owe this example to Garrett Cullity.

[187] There are, of course, other alternatives. This person would have sufficient reasons to consent to my giving this money to some other aid agency, which would use my gift to save someone else from some similarly great harm. This point does not affect my claim that, in such cases, the Consent Principle requires me to make such a gift.

[188] *Metaphysics of Morals*, henceforth *MM* 454. See Wood (1999) 5-8, from whom I take this and the next quotation.

[189] *Lectures on Ethics*, edited by Peter Heath and J.B. Schneewind, henceforth *Lectures* (Cambridge University Press, 1997) 179 (Prussian Edition, 27: 416).

[190] This may be the most important moral question that most rich people face. For three excellent discussions, see Murphy (2000), Mulgan (2001) Cullity (2004) and Pogge (2002).

[191] G 392.

[192] Kant writes, 'all rational beings stand under the law that each of them is to treat himself and all others never merely as means but always at the same time as ends-in-themselves' (G 433). It is sometimes said that we can ignore Kant's claim that we must never treat people merely as a means, since it is enough to know what Kant means by treating people as ends. If we treat someone as an end, that ensures that we are not treating this person merely as a means. [References] But treating people as ends, Kant claims, consists in part in not treating them merely as a means, so we should ask what that involves.

[193] Kamm gave me this objection in discussion. In Kamm (2007) 12-13 and her notes to these pages, Kamm gives an account of treating merely as a means which is very different from mine. On Kamm's account, whether we are treating someone merely as a means does not depend on our attitude to this person. And we might be treating someone *merely* as a means even if we are not treating this person *as a means*, or are sacrificing our life for this person's sake. Though Kamm makes several plausible moral claims, she is not, I believe, describing the ordinary meaning of the phrase 'treat merely as a means'.

[194] G 423. Kant discusses someone for whom 'things are going well', and who 'contributes nothing' to those who are in need.

[195] MM 443. But Kant also praises Leibniz for taking the trouble to place a worm back on its leaf after examining it under a microscope (CPR 5:160).

[196] As this example also suggests, the moral belief mentioned in condition (1) need not be true. I am not proposing (B) as a *criterion* that might help us to decide whether someone is treating someone else merely as a means, or is close to doing that. My aim is only to describe two of the ways in which we might plausibly deny that some act is of this kind. We cannot object to (B) by claiming that, even if (3) our treatment of someone is governed in sufficiently important ways by some relevant belief or concern, it might still be true that (4) we are treating this person merely as a means. If (4) were true, (3) would not be true, since our treatment of this person would not be governed in *sufficiently* important ways, or this governing belief or concern would not be *relevant*.

[197] G 429.

[198] For a further defence of these claims, see pages 000 below.

[199] This is claimed, for example, by Nozick (1974) 31, and Kamm (2007) 000.

[200] For example, 'rational beings. . . are always to be valued at the same time as ends, that is, only as beings who must be able to contain in themselves the end of the very same action' (G 429-30, my italics).

[201] See page 00 above.

[202] O'Neill (1989) 111 and 114.

[203] Korsgaard (1996A) 347. Korsgaard may be intending only to describe Kant's view.

[204] O'Neill (1989) 138.

[205] Korsgaard (1996A) 142.

[206] Korsgaard (1996A) 93.

[207] Rawls (1971), 111. and 184

[208] Rawls (1999) 355.

[209] Since Rawls makes no use of these proposed senses of 'right' and 'true', my remarks are no objection to his moral theory.

[210] As when he claims that, if someone kills himself to avoid suffering, or gives himself sexual pleasure, this person thereby treat himself merely as a means (G 429, and MM, 425).

[211] It might be suggested that, when this Egoist saves this child, what he is doing is not wrong, but his doing of it is. For a comment on this suggestion, see pages 000 below.

[212] Thomson (1990) 166-168. Thomson adds: 'Where the numbers get very large, however, some people start to feel nervous. Hundreds! Billions! The whole population of Asia!'

[213] Thomson (1990) 153. Thomson's claim is about an act that would save four people's lives; but she would apply it, I believe, to the saving of a single life.

[214] We might claim, however, that it would be wrong for this gangster to save his *own* life in this way.

[215] G 428.

[216] Wood (1999) 152-5.

[217] Wood (1999) 117.

[218] MM 462-8.

[219] Wood (1999) 141, and Kagan (2002) 000.

[220] Wood (1999) 155.

[221] Wood (1999) 139. (This book is *the Metaphysics of Morals*.)

[222] Wood (2006) 346.

[223] MM 444 and 392.

[224] MM 423-5.

[225] MM 429-30.

[226] Wood (1999) 154, and 371, note 32.

[227] Rawls (1971) 31, note 16.

[228] Wood (1999) 141.

[229] Herman (1993) 208, 153.

[230] I here follow Scanlon (1998) Chapters 1 and 2.

[231] Some writers claim that events are good as a means, or instrumentally good, only when and because these events would be an effective means to some *good* end. On this account, giving someone a lethal poison would not be good as a means of killing this person unless this person's death would be good. It seems clearer to claim that (1) some event would be good as a means if this event would be an effective means to some end, but that (2) we have no reason to want some event that would be good as a means to some end unless this end is good, or is an end that we have reasons to want to achieve. There are other distinctions that are worth drawing. Of the things that are good as ends, for example, some are good by *contributing* to some wider good end.

[232] Moore (1903) 171. (At the end of this paragraph he seems to contradict this claim.)

[233] Scanlon (1998) 99.

[234] It may be objected that since these things have features that give us reasons to treat them in certain positive ways, they are good in the reason-involving sense that I defined in Section 2. But that definition does not imply that things are good whenever they have such features. This objection shows, however, that my definition is incomplete. We must say more to explain which are the reason-giving features that can make something good. (There are other kinds of value which are not kinds of goodness. One example is economic value. Some bad paintings are very valuable. But such value is irrelevant here.)

[235] Scanlon (1998) 104.

[236] Scanlon (1998) 105.

[237] It is a different question whether assisting suicide should be a crime. Even when some kind of act is not wrong, it may be justifiable for such acts to be treated as crimes, since that may be the best way to prevent various bad effects.

[238] G 435-6.

[239] Herman writes, 'the domain of the good is rational activity and agency, that is willing' (Herman (1993) 213).

[240] G 396-7.

[241] G 433 and 438. If everyone had good wills and always acted rightly, that would produce the Realm of Ends not by *causing* but by *constituting* this ideal state of affairs.

[242] Kant's phrase is 'das höchste Gut', which literally means 'the Highest Good'. But Kant's phrase is misleading. As Kant himself points out, what he calls 'das höchste Gut' does not have a goodness that is *higher* than the goodness of a good will, but only the goodness that is most complete (CPR, 111). The phrase 'the Greatest Good' better suggests what Kant means, since this good is the greatest, not by being the highest, but by being the most complete.

[243] For references, see the notes near the start of Section 32.

[244] CPR 119.

[245] G 428.

[246] *The Critique of Judgment* 442-3.

[247] Herman (1993) 238.

[248] Wood (1999) 133.

[249] Herman (1993) 238. Wood writes: 'Kant, however, proposes to ground categorical imperatives on the worth of any being having humanity, that is, the capacity to set ends from reason, irrespective of whether its will is good or evil' (Wood (1999) 120-1). Kant sometimes remarks that, by acting wrongly in certain ways, we would throw away our dignity, so that we had even less worth than a mere thing. But that is not really Kant's view.

[250] Herman (1993) 213.

[251] Herman (1993) 121. Thomas Hill similarly writes that, when Kant claims that persons are ends-in-themselves, that is a short way of saying that rationality in persons is such an end ((1992) 392).

[252] G 435.

[253] Newman (1901) Vol I, 204. [Ross, with less excuse, makes a similar claim.]

[254] Hill (1992) 50-57.

[255] G 435.

[256] Reference. [Kemp Smith? Beck? Allison?] As one example, we can note how Kant misdescribes his view. Kant claims that humanity is an end-in-itself, which has dignity in

the sense of supreme and unconditional value. But Kant also claims that only good wills have such value. These claims do not conflict, Korsgaard suggests, because Kant uses 'humanity' to refer to 'the power of rational choice', and this power is 'fully realized' only in people whose wills are good, because it is only these people whose choices are fully rational (Korsgaard (1996A) 123-4). This suggestion has some plausibility. But Kant also uses 'humanity' to refer to rational beings, which he claims to be ends-in-themselves, with supreme value. We could not similarly claim that rational beings are the same as good wills because such beings are fully realized only when they have good wills. Nor could we claim that rational beings are the same as the Realm of Ends, or the Greatest Good: the world of universal virtue and deserved happiness. Though Kant claims that only good wills have dignity, we should admit that, on Kant's view, there are several kinds of thing that have such supreme or unsurpassed value. For a discussion of these questions, see Dean (2006).

[257] MM 427.

[258] Herman (1993) 215.

[259] Herman (1993) 210.

[260] See, for example, CPR 20.

[261] As I have said, there are other kinds of value which are not kinds of goodness, such as economic value. That is irrelevant here.

[262] Herman (1993) 129.

[263] For example, Kant writes 'the greatest good of the world, the *Summum Bonum*, or morality coupled with happiness to the maximum possible degree' (Lectures 440 (27: 717). (See note 220 above on why I translate such claims with the word 'greatest'.)

[264] CPR 125, Kant writes 'We', but he means 'all of us' or 'everyone'. He also writes, 'The production of the Greatest Good in the world is the necessary object of a will determinable by the moral law' (CPR 122), and 'it is our duty to realize the Greatest Good to the utmost of our capacity' (CPR 143 note).

[265] CPR 129.

[266] I am here following Kant, who writes, 'By this they meant the highest good attainable in the world, to which we must nevertheless approach, even if we cannot reach it, and must therefore approximate to it by fulfilment of the means' (Lectures 253 (27:482). He also writes: 'This Summum Bonum I call an ideal, that is, the maximum case conceivable, whereby everything is determined and measure. In all instances we must first conceive a pattern by which everything can be judged ' (Lectures 44 (27:247).

[267] For example, Stephen Engstrom writes that, on Kant's view, the achievement of such proportionality would be 'the next best thing' (Engstrom (1992) 769).

[268] Kant for example writes that 'a rational and impartial spectator can never be pleased' at the sight of the happiness of a will lacking any trace of virtue, and that when such happiness is removed 'everyone approves and considers it as good in itself'. And he writes, , 'if someone who likes to vex and disturb peace-loving people finally gets a sound thrashing for one of his provocations. . . everyone would approve of it and take it as good in itself even if nothing further resulted from it' (CPR 61).

[269] In Moore's words '"right". . . does and can mean nothing but "cause of a good result"(Moore (1903) 196). Moore must mean 'cause of the best result'. Characteristically, Moore adds, 'it is important to insist that this fundamental point is demonstrably certain'. (When Moore's clouds, for many decades, hid the light from Sidgwick's sun, that was in part because, unlike the judicious Sidgwick, Moore writes with the extremism that makes Kant's texts so compelling. With the exception of the 'doctrine of organic unities', every interesting claim in Moore's *Principia* is either taken from Sidgwick or obviously false. (This remark of mine is an overstatement of the Moorean kind.)

[270] It is surprising that Moore makes this mistake, since he devotes an entire chapter to condemning such mistakes, which he calls 'the Naturalistic Fallacy' (though it is neither naturalistic nor a fallacy). Sidgwick more accurately describes this mistake in two sentences (ME 26 note 1, and 109).

[271] CPR, 63-4.

[272] G 413. Explain why the word 'ought' is a mistranslation.

[273] G 412.

[274] In Kant's words, 'It is impossible to think of anything at all in the world. . . that could be considered good without limitation except a good will.' He goes on to say that this goodness is unsurpassed, and absolute.

[275] CPR 64.

[276] *Religion* 72.

[277] Lectures 440-1 (27:717). This 'highest end' is the Greatest Good.

[278] *Religion Within the Boundaries of Mere Reason*, 6: 8.

[279] Provided, Moore adds, that these rules are both 'generally useful and generally practiced' (Moore (1903) 211-13). Moore denied that it would be best if there was most happiness; but this point is irrelevant here.

[280] *Enquiry* Appendix III, 256 (my emphasis). He also writes 'The result of the individual acts is here, in many instances, directly opposite to that of the whole system of actions; and the former may be extremely hurtful, while the latter is, to the highest degree, advantageous.' In the *Treatise* Hume writes: 'however single acts of justice may be contrary, either to public or private interest, 'tis certain, that the whole plan or scheme is highly conducive, or indeed absolutely requisite, both to the support of society, and to the well-being of every individual. 'Tis impossible to separate the good from the ill'. Book III, Section 2, 497 in Selby-Bigge.

[281] 'On a supposed right to lie from philanthropy' (8: 425-30).

[282] 8: 426.

[283] *Lectures* 388 (27:651).

[284] *Metaphysik* L1,28:337, From lectures given around 1778, cited in Guyer (2000) 94.

[285] MM 385-388. Our duty to promote our own virtue is the most important part of a wider duty to promote our own perfection, which includes our other abilities as rational beings.

[286] See, for example, the quotations in note 188 above.

[287] As Rawls writes: 'There is nothing in the CI-procedure that can generate precepts requiring us to proportion happiness to virtue' (Rawls (2000) 316.)

[288] *First Critique* 640. He also writes: 'there is in the idea of a practical reason something further that accompanies the transgression of a moral law, namely its deserving punishment' (CPR 37).

[289] In Kant's words, 'he must also assume freedom of the will in acting, without which there would be no morals.'

[290] This argument is valid, and Kant claims that we know that its premises are true. Surprisingly, however, Kant denies that we can know this argument's conclusion to be true. More exactly, Kant claims that we know this conclusion only from a practical point of view. This claim seems a mistake. If we know that both this argument's premises are true, as Kant claims, we have decisive epistemic reasons to believe this argument's conclusion. So we know this conclusion to be true from a theoretical point of view. Kant's claim can at most be that we have no theoretical knowledge about *how* this conclusion can be true, since we cannot understand the atemporal noumenal world.

[291] For example, Kant writes 'he *ought* to remain true to his resolve, and from this he rightly concludes that he must *be able* to do it' (Religion 50).

[292] This may be one of Schultz's points in the book that Kant reviews. Schultz writes: 'Remorse is merely a misunderstood representation of how one could act better in the future.'

[293] 'Review of Schultz', 8: 13.

[294] Also in his 'Review of Schultz', 8:13

[295] *The* CPR, 5:95.

[296] *Religion Within the Boundaries of Mere Reason*, 6: 44.

[297] *Nicomachean Ethics*, 1114a19; cf.1114b30 *seq*.

[298] See Nagel (1986) Chapter 7. In my statement of this argument, I partly follow Galen Strawson, who gives excellent versions of this argument in Strawson (1994) and Strawson (1998).

[299] For discussions of the many questions raised by the belief that no one can deserve to suffer, see Sidgwick, ME, Chapter V, especially section 4, and Pereboom (2001) Chapters 5 to 7.

[300] Herman (1993) vii.

[301] CPR 27.

[302] CPR, 19.

[303] G 423.

[304] *The* CPR 34.

[305] G 424 and surrounding text. As Kant elsewhere says, 'An action is morally impossible if its maxim cannot function as a universal law. . .' Lectures 000. Kant also writes 'Some actions are so constituted that their maxim cannot even be thought without contradiction as a universal law. . .' Following O'Neill, several writers call this formula the 'contradiction in conception test'. When we have decided what it would be for some maxim to be a universal law in Kant's intended sense, we may find that it would be logically impossible, and in that way a contradiction, to suppose that certain maxims are such laws. This claim applies to some of the maxims that I shall discuss. But when Kant claims that certain maxims could not be universal laws, he appeals to empirical impossibilities, which rest on assumptions about human nature. By adding such assumptions to our description of some possibility, we might be able to produce some kind of contradiction. But the idea of a contradiction

would not here do useful work. So I shall ask whether certain maxims could not be universal laws, in whatever is Kant's intended sense, without restricting the kind of impossibility that would be involved.

[306] MM 453. Kant also refers to the universality of a law that everyone *could* act in certain ways (G 422, my emphasis).

[307] To apply (A), we must know in what sense we could not all be permitted to act on some maxim. That would be in one sense true if, in a world in which we all acted on this maxim, at least some of us would be acting wrongly. But (A) would not help us to decide whether, in such a world, some of us would be acting wrongly. There is, I believe, no other helpful sense in which it might be claimed to be wrong to act on some maxim if it could not *be* true that we are all permitted to act upon it. Kant elsewhere claims it to be wrong to act on some maxim if we could not rationally *will* it to be true that we are all permitted to act upon it. That is a more plausible claim, to which I shall return.

[308] See, for example, O'Neill (1989), 157. (O'Neill's view has since changed. See, for example O'Neill (1996) 59.)

[309] This is most clearly shown in Kant's discussion of lying promises in G 422.

[310] Herman (1993) 118-119.

[311] Herman (1993) 119.

[312] Korsgaard (1996A) 136.

[313] *Lectures* 232-3 (29:609).

[314] CPR, 19.

[315] G 402-3, and 422.

[316] G 422.

[317] Rawls *Lectures* 169.

[318] We can add that, in believing that such lying promises were permissible, these people would have lost the concept of a moral, trust-involving promise. (There might still be a practice that was like the practice of such promises, except that it took a non-moral form. Such promises would be like threats. Just as we could have reasons to fulfil our threats to preserve our reputation as a threat-fulfiller, we could have reasons to keep such promises to preserve our reputation as a promise-keeper.)

[319] 'On a supposed right to lie from philanthropy' *8*, 425-30.

[320] These imagined cases might be claimed to be unrealistic, because in the real world the facts would not have been as simple as I have asked us to suppose. But these cases are plausible enough to provide good tests of the acceptability of (G). We could not defend this formula by saying that these examples are too bizarre, or fantastic. Moral principles ought to succeed when applied to somewhat simplified imagined cases of this kind. And Kant's claims about a lying promise are similarly simplified.

[321] Korsgaard (1996A) 95.

[322] Given these people's motives, they may not be truly *generous*. But they might still be admired by themselves and others for what was mistakenly believed to be their generosity.

[323] O'Neill (1989), 133 and 215 and elsewhere.

[324] O'Neill (1989) 138-9.

[325] O'Neill (1989) 215-6.

[326] O'Neill (1989) 102-3.

[327] Korsgaard (1996A) 92-3.

[328] This is Herman's example (Herman (1993) 138-9).

[329] I take this example from Blackburn (1998) 218.

[330] Herman again (Herman (1993) 141).

[331] Of Kant's many versions of this formula, most take the form of commands, so that they could not be either true or false. But when Kant first proposes this formula, he writes 'I ought never to act except in such a way that I could also will that my maxim would become a universal law' (G 402).

[332] Herman (1993) 123.

[333] He writes, for example, 'Maxims must be chosen as if they were to hold as universal laws of nature' (G 436). See also G 421, and CPR 69-70.

[334] For example, Kant writes 'could I indeed say to myself that everyone *may* make a false promise when he finds himself in a difficulty?' (G 403), and he refers to 'the universality of a law that everyone. . . *could* promise whatever he pleases with the intention of not keeping it' (G422,). Similarly Kant refers elsewhere to 'the law that everyone *may* deny a deposit which no one can prove has been made' (CPR 27). And as I have said, Kant writes of a maxim's being 'a universal permissive law' (MM 453). (In all these quotations the emphases are

mine.) This permissibility version of Kant's formula was suggested by Scanlon in unpublished lectures in 1983. See also Pogge (1998) Wood (1999) 80, and Herman (1993) 120-1.

[335] Kant does not explicitly appeal to this formula. But he is reported to have said, in lectures, 'you are so to act that the maxim of your action shall become a universal law, i.e. would have to be universally *acknowledged* as such' (*Lectures* 264 (27: 495-6). And Kant also writes: 'if everyone . . *considered* himself authorized to shorten his life as soon as he was thoroughly weary of it' (CPR 69). (As before, the emphases are mine.)

[336] Suppose we appealed only to the Permissibility Formula. We would then ask whether we could rationally will it to be true that everyone is permitted to act on some maxim, even though this would make no difference to anyone's moral beliefs, or to anyone's acts. This would not be a helpful question. First, it is hard to imagine that we could will it to be true that certain acts are permitted, or are wrong. As Kant himself claims, and many other people have believed, not even God could have willed that certain kinds of wrong act be morally permitted. And if the fact that certain acts are permitted would make no difference to what anyone believes or does, it is unclear what reasons we could have for willing that these acts be permitted, other than the fact that, as we believe, these acts really are permitted. But whether that belief is true is what Kant's formula is intended to help us to decide.

[337] G 403.

[338] Rawls (2000) 166-70, who attributes this point to Herman.

[339] I am here assuming that, unlike Kant's Consent Principle, Kant's Formula of Universal Law is intended to be the only moral principle we need, so that when some version of this formula does not imply that some act is wrong, this formula thereby implies that this act is morally permitted.

[340] O'Neill (1989) 85.

[341] See Wood's excellent discussion in Wood (1999) 103-5.

[342] Lectures, 187

[343] MM 455-7.

[344] *The* CPR 34.

[345] If Kant accepted the Whole Scheme View, as I suggest on page 000, it might not have been irrational for him to will that no one ever tells a lie. But the Whole Scheme View is false, and when we apply Kant's formula, we should ask what people could rationally will if they had no false beliefs.

[346] Wood (2006) 345, and Wood (2002) 172.

[347] Herman (1993) 104, 132.

[348] O'Neill (1975) 129, 125. See also O'Neill (1989) 130.

[349] Hill (2002) 122.

[350] Herman (1993) 117.

[351] O'Neill (1989) 86, 98, 103.

[352] G 403.

[353] G 422.

[354] G 423.

[355] G 421-2.

[356] CPR, 8 note. Kant also writes: 'all imperatives of duty can be derived from this single imperative', and 'These are a few of the many actual duties. . .whose derivation from the one principle is clear.'

[357] G 404, 424.

[358] O'Neill , Herman, Pogge, and Shelly Kagan all make or discuss proposals of this kind (O'Neill (1989) 87, 130-1; Herman (1993) 147-8; Pogge (2004) 56-58; and Kagan (2002) 122-127.

[359] ME 202 note. Sidgwick claims that, though this revolutionary's intention was to kill the Czar, it would be false to say that he did not intend to kill the other people. It is better to say, I believe, that what he was intentionally doing was acting in a way that he knew would kill many people.

[360] There are some exceptions. We might claim, for example, that in driving recklessly, someone caused an accident and thereby killed some some other people.

[361] In Kant's longer statement, this maxim is: 'from self-love I make it my principle to shorten my life when its longer duration threatens more troubles than it promises agreeableness' (G 422). This maxim might be a policy, since we can often shorten our lives. Smokers might do that every time they smoke. But Kant is here discussing a single act of suicide.

[362] O'Neill (1975) 112.

[363] G 424. O'Neill herself later writes 'this is not to say that in the actual world there is some contradiction in the thinker of each deceiver' (O'Neill (1989) 132).

[364] O'Neill (1989) 87.

[365] O'Neill (1975) 112-117, and 124-143, and O'Neill (1989) 130. Herman makes similar claims in Herman (1993) Chapters 4 and 10.

[366] As is suggested by his remarks about his self-reliant man whose maxim is 'Don't help others, but don't cheat them either' (G 423) . Kant claims that, if everyone acted on this man's maxim, this world would be better than the actual world in which many people help others, and many people cheat. But Kant also claims that we could not rationally will it to be true that everyone acts on this man's maxim. Kant's implied comparison must here be with a world in which no one acts on this man's maxim.

[367] There are also probabilistic each-we dilemmas, which appeal to the likely effects of different acts, or to what would be expectably-best for people. I discuss these cases in Chapter 2 to 5 of Parfit (1984-7), and Parfit (1986).

[368] In the simplest cases (1) each of us can *often* either benefit herself or give a greater benefit to others, and (2) because the number of people involved is fairly small, what each does may affect what, in later situations, other people do. In a two person-case, for example, if I give you the greater benefit, you may reciprocate, and give me the greater benefit. If I switch to giving myself the lesser benefit, you may retaliate, and give yourself the lesser benefit. Though these are called 'repeated prisoner's dilemmas', they are *not* prisoner's dilemmas, or each-we dilemmas. In such cases, it is not true that, if each rather than no one does what is certain to be better for herself, that would be worse for all of us. These cases are theoretically much less interesting, and fundamental, since they are merely one of the many kinds of case in which it is unclear which way of acting would be best for ourselves. Such cases are also practically much less important, since they are much less common. They are, however, important to evolutionary psychologists who are trying to explain various features of animal behaviour, and human psychology, and to historians who are discussing the small communities in which, in earlier centuries, most people have lived.

[369] It is worth mentioning one kind of case that shows the significance of numbers. We can call these *Samaritan's dilemmas*. Each of us can sometimes help some needy stranger, at some small but real cost or burden to ourselves. That might be true, for example, when we could help someone who has had an accident, or we could return lost property of great personal value. If all of us always gave such help to strangers, that might be better for all of us than if none of us ever gave such help. But if we live in large cities, as more than half of the world's population now do, it might also be better for each person if he never gave such help. This person would then avoid the costs to himself. And whether he received such help would very

seldom depend on whether he gave such help to others. The strangers whom each of us failed to help would hardly ever be the same people as the strangers who could later help us. So our failure to help others would hardly ever lead others, bearing a grudge, to deny us help. But if no one helps others, though *each* of us would be doing what would be better for herself, *we* would be doing what would be worse for all of us.

[370] There is a further distinction between those goods which in fact benefit even those people who do not help to produce them, and those which are bound to do that, since there is no feasible way to prevent non-contributors from getting these benefits. Clean water may often be in the first category, and clean air in the second.

[371] There is also a way in which, in such cases, common sense morality itself implies that we ought to cease to give priority to our M-related people. If we and the other members of the relevant group could all communicate, and we all knew each other to be trustworthy, we would all be rationally or morally required to make a joint conditional promise that we shall always act differently, by giving the greater benefits to others. If this joint promise would become binding only if everyone makes it, this fact would, when we are deciding whether to make this promise, *tie our acts together*. In making such a promise, each of us would be doing what would be best for himself or his M-related people, since he would be helping to bring it about that everyone rather than no one did what would be better for him or for these other people. Since this promise requires unanimity, each person would know that, if he did not make this promise, the whole scheme would fail. So common sense morality would itself tell us all both to make and to keep this promise. This solution, however, could seldom be achieved, since we are not all trustworthy, and, even if we were, it would often be too difficult to arrange and achieve such a joint conditional agreement. If we were all sufficiently conscientious Kantians, we would avoid this problem.

[372] MM 393.

[373] In a different way, however, this solution may be indirectly collectively self-defeating. See page 000 below.

[374] We might, however, draw a distinction here. It is clear that, in each-we dilemmas, what we *should all ideally do* is to give the greater benefits to others. If all rather than none of us acted in these ways, that would be better for everyone. But Kant's formula requires such acts even when most other people are *not* acting in these ways. In such cases, by acting in these ways, we would lose the lesser benefits that we could give ourselves without receiving the greater benefits from others. This requirement may sometimes be too demanding. It might also be unfair. In unsolved Parent's Dilemmas, for example, it may be unfair to our children if we give the greater benefits to other people's children, when other people are not giving such greater benefits to our children. In at least some of these cases, we might justifiably believe that it makes a difference how many other people are doing what we should all ideally do. We might be required to give the greater benefits to others only when *enough* other people are

acting in this way. In other cases, we might be permitted, as a defensive second-best, to give the lesser benefits to ourselves, our children, or our other M-related people. For a suggestion about what would count as *enough*, see my (*1984-7)* 100-1. [Transfer that to here?] [Refer also to Murphy.]

[375] I take this example from Pogge (1998) 190.

[376] It may be objected that two of these are incomplete maxims, since they don't tell us the agent's purpose or aim. But it would be tedious and unnecessary always to describe such a purpose. Kant often doesn't do that. We can often assume that some maxim's aim is to benefit the agent. And in many cases, the points we are making are not affected by the agent's aim.

[377] Pogge (1998) 190. Pogge is here following an unpublished lecture given by Scanlon in 1983.

[378] In his biography of Kant, however, Manfred Kuehn writes: 'Kant formulated the maxim: 'One mustn't get married'. In fact, whenever Kant wanted to indicate that a certain, very rare, exception to a maxim might be acceptable, he would say: 'The rule stands: "One shouldn't marry! But let's make an exception for this worthy pair" (Kuehn (2001) 169).

[379] We should suppose that you and I are the only people who could act on some maxim by doing A. As elsewhere, 'everyone' refers to all of the people to whom some maxim applies. So, in willing that both you and I act on this maxim, I would be willing that everyone acts upon it.

[380] Korsgaard (1996A) 149. Korsgaard makes this claim not about Kant's Law of Nature Formula but about his Formula of Humanity. But this difference is irrelevant here.

[381] Hill (2000) 66.

[382] Similar but more complicated claims would apply to other cases: those in which it would be best, not if everyone acted in the very same way, but if everyone played his or her part in the best possible pattern of acts.

[383] This maxim needs some qualifications to pass Kant's test, since there are some cases in which we ought to break some promise, or fail to help someone in need. But this complication does not affect my argument.

[384] This version of RC is open to another objection which I discuss on p 000. But this objection is irrelevant here.

[385] This rule would not in fact be ideal, for reasons that I describe on p 000, but this point is irrelevant here.

[386] For the best recent statement and defence of Rule Consequentialism that is known to me, see Hooker (2000).

[387] I am partly following some of Kagan's suggestions in Kagan (2002), and Kagan (1998) 231-5.

[388] As before, similar claims apply to those versions of RC which appeal to the rules whose being *accepted* by people would make things go best. My proposed revision applies much more easily to these *acceptance-versions* of Rule Consequentialism, because the optimific rules would take much simpler forms. (As Michael Ridge has pointed out, even if such rules took conditional forms, there may be no set of rules whose acceptance would make things go best at *each* level of acceptance. But there would be sets of rules whose acceptance at different levels would, on balance or on the whole, make things go best. For a partly similar discussion of these questions, see Ridge (2006) 242-253.)

[389] See Hooker's discussion of this question Hooker (2000).

[390] As Herman notes, Herman (1993) Chapter 7.

[391] If people have conflicting beliefs, for example, these beliefs cannot all be true; and we can plausibly assume that everyone ought to have, or try to have, true moral beliefs.

[392] We might be able to defend a moral theory that is partly self-effacing, because it implies that we should not all accept this theory. But such theories need to be defended. For some discussion, see Chapter 1 of my *Reasons and Persons*.

[393] MM 451. I have changed 'benevolent' to 'beneficent', since that must be what Kant means.

[394] The ancient Near East, India, and China. Add references.

[395] G 430 note.

[396] G. 423 (my italics).

[397] Nagel (1970) 000, and (1991) 000-000.

[398] Hare (1963) 000.

[399] Rawls (1971), *passim*.

[400] As Leibniz pointed out. See Leibniz (1988) 56. (I owe this reference to Raphael (2001) 84-5.)

[401] MM 450-1.

[402] Kant similarly writes: 'since all others with the exception of myself would not be all, so that the maxim would not have within it the universality of a law. . . the law making benevolence a duty will include myself, as an object of benevolence, in the command of practical reason' (MM 450).

[403] O'Neill (1989) 94.

[404] Rawls (1971), section 30.

[405] G 424.

[406] See Wood (1999) 3 and 7.

[407] See, for example, G422.

[408] Korsgaard (1996A) 101.

[409] Nagel (1991) 42-3.

[410] Kant does write 'every rational being. . . must always take his maxims from the point of view of himself, and likewise every rational being' (G 438). But this remark comes in Kant's discussion, not of his Formula of Universal Law, but of his Formula of the Realm of Ends. And if Kant had intended that we should imagine others doing to us what we do to them, he would not have so contemptuously dismissed the Golden Rule.

[411] G 423, (my emphases).

[412] Rawls writes: 'I believe that Kant may have assumed that [our] decision. . . is subject to at least two kinds of limit on information. That some limits are necessary seems evident. . .' (Rawls (2000) 175.)

[413] Quote and discuss the passage from the CPR to which Rawls appeals.

[414] Williams (1968) 123- 131.

[415] Scanlon (1998) 170-1, and in unpublished summaries of lectures.

[416] G 402.

[417] G 432. And he refers to 'the concept of every rational being as one who must regard himself as giving universal law. . .' But Kant never explicitly appeals to what everyone could rationally will. The phrase just quoted, for example, ends 'through all the maxims of his will' (G 434). If each person regards himself as giving laws through the maxims of *his* will, he is not asking which laws everyone could will. At several other points, when Kant

seems about to appeal to what everyone could will, he returns to his Formula of Universal Law, telling us to appeal to the laws that we ourselves could will.

[418] This move from Kant's original formula to Scanlon's revised version is, however, a move to a significantly different view. Scanlon describes this difference in some lecture notes from which, because they are unpublished, I shall quote at length. Discussing the Formulas of Universal Law and of the Realm of Ends, Scanlon writes:

'My own view is that [these] formulas, when generously interpreted, may be extensionally equivalent, but that their apparent rationales---and the reasons why they have appealed so strongly to so many people over the years---are in fact quite distinct. Roughly speaking, these three successive formulations of the moral law represent a slide from a view of morality as grounded in the requirements of freedom understood as independence from inclination to a view (to me much more plausible and appealing) of morality as based in a kind of ideal agreement.

This difference is shown in the fact that while the question asked by the Universal Law form of the Categorical Imperative is whether I (the agent) could will a maxim to be a universal law, the formula of the Kingdom of Ends makes explicit the idea of a harmony of different wills, each legislating in such a way as to recognize the status of all as ends-in-themselves. The aim of objective self-consistency and the aim of harmony with other wills may, if Kant is correct, have many of the same consequences, but they reach these consequences in quite different ways.

The test posed by the Universal Law form is, on its face, a test of what an agent can will, and its authority derives from the conditions under which the agent can conceive of him or herself as free. So neither in its application nor in its derivation does this formula depend essentially on the agent's relation to others.'

[419] If these formulas sometimes had conflicting implications, we would have to choose between them. These formulas might conflict in cases in which (1) we could not rationally will it to be true that everyone acts in some way, but (2) we *could* rationally will it to be true that everyone believes such acts to be morally permitted, because we know that, if everyone had these beliefs, there wouldn't be too many people who would choose to act in this way. If these formulas did conflict, when applied to such cases, MB5 would be clearly better. To avoid such conflicts, we might move from LN5 to

> LN6: It is wrong to act in some way unless everyone could rationally will it to be true that everyone acts in this way, when they know that there won't be too many other people who would choose to act in this way.

But this formula is too similar to MB5 for it to be worth discussing both formulas. And MB5 is, I believe, both closer to Kant's view, and clearly better. LN6 is too simple, since it makes

a difference *why* there won't be too many people who would choose to act in some way. It makes a difference, for example, whether some people are refraining from acting in some way because they believe that, given the number of people who are already acting in this way, further acts of this kind would be wrong. When that is true, those who act in this way may be unfairly benefitting from the conscientious self-restraint of others. Rather than including such details into our descriptions of how people are acting, as the Law of Nature Formula requires, we do better to include such details in the content of the beliefs to which the Moral Belief Formula refers. As I have already said, while it is only in certain cases that we can usefully ask 'What if everyone did that?', it is always relevant to ask 'What if everyone thought like you?'

[420] Scanlon (1998) 171.

[421] Wood (1999) 172, Herman (1993) 104,132, O'Neill (1975) 125, 129.

[422] or something similar, such as steadily increasing penalties for failure to agree.

[423] Gauthier (1986) 133.

[424] See, for example, Sugden (1990).

[425] See Brian Barry,

[426] Gauthier (1986) 18 note 30, and 268.

[427] Gauthier (1986) 269. Gauthier also argues that, if we accept his Contractualist theory and his minimal version of morality, he can show that, even in self-interested terms, it cannot be rational to act wrongly. No other moral theory, Gauthier claims, can achieve this aim. Gauthier's argument, I believe, fails in ways that I describe in Appendix C.

[428] Quoted in Gauthier (1986) 17.

[429] Rawls (1971) sections 18-9.

[430] One example is Rawls's appeal to the arbitrariness of the natural lottery. I am here following several writers, especially Thomas Nagel, in Nagel (1973) reprinted in Daniels (1975), and Barry (1989) and (1995).

[431] Rawls (1971) 569, RE 498.

[432] Rawls (1971) 575, RE 503-4.

[433] Rawls (1996) 49.

[434] Rawls (1971) 184-5, RE 161. Compare his claim 'in order that the parties can choose at all, they are assumed to have a desire for primary goods' in Rawls (1999) 266.

[435] In appealing to his formula, Rawls writes, 'we have substituted for an ethical judgment a judgment about rational prudence' (Rawls (1971) 44). When we are behind the veil of ignorance, we are 'assumed to take no interest in one another's interests' (Rawls (1971)) 147. The people behind the veil of ignorance, he also writes, 'are prompted by their rational assessment of which alternative is most likely to advance their interests' (1999) 312). Rawls does *not* assume that, in the actual world, everyone is self-interested.

[436] Rawls (1971) 142.

[437] Rawls (1971) 140.

[438] As Rawls writes, 'The combination of mutual disinterest and the veil of ignorance achieves the same purpose as benevolence. For this combination of conditions forces each person in the original position to take the good of others into account' (Rawls (1971) 148). Rawls's comparison here is with *impartial* benevolence, and, as he points out, the veil of ignorance makes *partiality* impossible.

[439] Rawls (1971) 22.

[440] Add references to Brian Barry.

[441] He writes, for example, 'the Utilitarian extends to society the principle of choice for one man'(Rawls (1971) 28).

[442] Rawls (1971) 165-6, RE, 143-4.

[443] Rawls (1971) 168, RE 145.

[444] Rawls (1971) 122 and 121, RE 105.

[445] John Rawls, (1999) 335-6. See also Rawls (1971) Section 40.

[446] As Rawls claims, Rawls (1971) 397

[447] Add some remarks about G. A. Cohen's discussion of this question.

[448] Rawls (1999) 265.

[449] Rawls (1971) 166, RE, 143.

[450] This objection to Rawls's argument I take from Nagel (1973) 11.

[451] Even when applied to the basic structure of society, the Maximin Argument may have implications that are much too extreme. Rawls sometimes defines the worst off group in broad terms, so that this group includes many people who are better off than some other people. On one suggestion, for example, the worst off people are those whose income is below the average income of unskilled workers (Rawls (1971) 98, RE, 84.) But if the Maximin Argument were sound, it would require a much narrower definition of this group. On this argument, each person ought to try to make her own worst possible outcome as good as possible. On Rawls's suggested broader definitions, we ought to choose policies that would make the representative or average member of the worst off group better off, even when that would be worse for the worst off people in this group. That is precisely what, when applied to society as a whole, Rawls's argument is claimed to oppose. When defending his broad definitions of the worst-off group, Rawls writes: 'we are entitled at some point to plead practical considerations, for sooner or later the capacity of philosophical or other arguments to make finer discriminations must run out' (84). But there is no difficulty in describing the worst off group as those who are equally worst-off, since these people are not better off than anyone else.

[452] Rawls (1971) 584, RE 512.

[453] As before, I am discussing only one part of Rawls's view. Though Rawls writes that his imagined contractors 'decide solely on the basis of what best seems calculated to further their interests so far as they can ascertain them,' he makes various other conflicting claims, as when he appeals to what he calls the *strains of commitment*.

[454] Rawls (1971) 29, RE 25-6.

[455] Rawls (1999), 174.

[456] In his last book, Rawls expresses doubts about his stipulation that, behind the veil of ignorance, we would 'have no basis for estimating probabilities'. He writes 'Eventually more must be said to justify this stipulation' (Rawls (2001) 106). But nothing more is said.

Rawls adds some other stipulations which allow him to put less weight on his claims about probabilities. He tells us to suppose that, by choosing his principles of justice, we would guarantee for ourselves a level of well-being that would be 'satisfactory', so that we would 'care little' about reaching an even higher level. We should also suppose that, if we chose any other principles, we would risk being much worse off. On these assumptions, Rawls argues, it would be rational for us to choose his principles of justice. Rawls then considers the objection that, by adding these assumptions, he makes his theory coincide with one version of rule Utilitarianism, since his principles would be the ones whose acceptance would make the average person as well off as possible. Rawls replies that, on his definition, rule Utilitarians are not Utilitarians (Rawls (1971) 181-2 and note 31, RE 158-9 note 32. This reply

is disappointing. Rawls earlier described his aim as being to provide an alternative to all forms of Utilitarianism. We do not provide an alternative to some view if we accept this view, but give it a different name.

[457] Rawls (1971) 4, RE 3.

[458] As I have said, it might be rational to choose principles which guaranteed that everyone would get some minimum level of primary goods, or which gave greater weight to avoiding what would be the worst outcomes for ourselves. But these principles would be fairly close to rule Utilitarian principles. [Quote some of Scanlon's remarks.]

[459] Explain why we can appeal here to altruistic reasons, to which I said we could not appeal when applying Kant's Formula of Universal Law. The difference is that we would there be appealing to rational requirements.

[460] See Scanlon (1998) 333-342.

[461] Scanlon (1998) 4-5 (and elsewhere).

[462] Scanlon (1998), 191-7. Scanlon does not assume that, when two people disagree, at least one of these people must be being unreasonable. There can be reasonable mistakes. But if neither person is being unreasonable in rejecting the other's principle, there may be no relevant principle that could not be reasonably rejected, with the result that Scanlon's Formula would fail. So, when Scanlon claims that no one could reasonably reject some principle, he should be taken to mean that anyone who rejected this principle would be making a moral mistake, by failing to recognize or give enough weight to other people's moral claims, even if this might be a not unreasonable mistake.

[463] Scanlon appeals to this restriction (though not with this name) on Scanlon (1998) 4-5, 194, and 213-6.

[464] Scanlon (1997) 272.

[465] Nor can we reject principles with claims that implicitly appeal to our deontic beliefs. Grey might claim that she could reasonably reject the Greater Burden Principle because it is *her* leg that would be being sacrificed to save Blue's life, and we can all reasonably insist that we have a veto over what other people do to our bodies. Grey would here be implicitly appealing to what some call the rights of self-ownership, or to the claim that it is wrong for other people to injure us without our consent. Scanlon's version of the Deontic Beliefs Restriction would exclude such appeals.

[466] Scanlon (1998) 215.

[467] Scanlon (1997) 267.

[468] This example was first suggested, I believe, by James Rachels in "Political Assassination," which originally appeared in *Assassination*, edited by Harold Zellner (Cambridge, Mass: Schenkman, 1974), 9-21 and is reprinted in James Rachels, *The Legacy of Socrates: Essays in Moral Philosophy*, edited by Stuart Rachels (New York: Columbia University Press, 2007), 99-111. See also the discussion by Judith Thomson in. . .

[469] This anxiety might not be rational, but that does not undermine these claims.

[470] In giving this argument, I am ignoring one feature of Scanlon's view. Scanlon claims that, in rejecting principles, we cannot appeal to the benefits or burdens that groups of people would *together* bear. If we follow this *Individualist Restriction*, we cannot oppose the Act Utilitarian view about *Transplant* by appealing to the anxiety and mistrust argument, since this argument appeals to the bad effects on many people of such anxiety and mistrust. Scanlon ought, I believe, to drop the Individualist Restriction, as I argue in my Response to Scanlon's Commentary below.

[471] These emergencies do not include intended threats to people's lives, such as threats by terrorists. Such cases have special features, such as our reasons not to act in ways that would encourage later threats of the same kind, and must therefore be covered by some other principle.

[472] Rawls (1999) 344.

[473] Rawls writes: 'the idea of approximating to moral truth has no place in a constructivist doctrine: . . . there are no such moral facts to which the principles adopted could approximate' (1999, 353.) It is Constructivists, we can add, who draw these distinctions, and who claim that, according to intuitionists, there are such independent normative truths. Some intuitionists would reject, or question, some of these meta-ethical claims.

[474] (1999) 351.

[475] Scanlon (2003) 149.

[476] I discuss this distinction further near the start of Part Five.

[477] See note 634 below.

[478] When we claim that someone could justifiably reject some formula, we do not imply that this formula is false, or should be rejected. People can justifiably have some false beliefs.

[479] As when he writes, 'Besides good and evil, or in other words, pain and pleasure. . . ' 439.

[480] The CPR, 60. Kant also claims that the principle of prudence, or self-love, is a hypothetical imperative, which applies to us only because we want future happiness. This claim assumes a desire-based view, ignoring our reasons to want our future happiness.

[481] On one interpretation, the Stoics were making the interesting claim that pain is not bad even in this non-moral sense. See for example, Irwin (1996) 80. According to some other writers, the Stoics *were* merely claiming, like Kant, that pain is not morally bad.

[482] Ross (2001) 272-284. (Though Ross makes these claims about pleasure, he intends them to apply to pain.)

[483] Nagel (1986) 161.

[484] Judith Thomson, for example, writes: 'Suppose someone asks whether [something] would be a good event. We should reply 'How do you mean? Do you mean "Would it be good *for* somebody?"'. We had better be told whether that is what is meant, or whether something else is meant. . . Consequentialism, then, has to go' (Thomson (2003) 19. In making this last claim, Thomson assumes too quickly that her question can't be answered.

[485] When there are no such precise truths about the relative goodness of outcomes, 'not worse than' should not be taken to mean 'at least as good as'.

[486] It could not be true both that

> certain acts are wrong because it would be bad if we acted in these ways,

and that

> it would be bad if we acted in these ways because such acts are wrong.

Wrong acts must have some other feature that makes them either bad or wrong. Nor could it be true that

> certain acts are wrong because such acts are disallowed by the best principles,
>
> such acts are disallowed by these principles because it would be worse if we acted in these ways,

and that

> it would be worse if we acted in these ways because such acts are wrong.

These claims go round in a circle, telling us nothing. Just as Contractualists must claim that, when we apply their formulas, we cannot appeal to the *deontic reasons* that might be provided

by the wrongness of certain acts, Consequentialists must claim that, when we apply their principles, we cannot appeal to the *deontic goodness* or *badness* of right or wrong acts. When Consequentialists make claims about how the rightness of our acts depends on facts about what would be best, these claims should use the word 'best' in what we can call its *deontic-value-ignoring* sense.

Similar claims apply to Non-Consequentialists. We reject Act Consequentialism if we believe that

> (A) certain acts are wrong even when it would be better if people acted in these ways.

To illustrate (A), we might appeal to a case like *Bridge*, claiming that

> (B) it would be wrong for you to save the five by causing me to fall in front of the runaway train, thereby killing me.

If we believe (B), would we also believe that it would be *better* if you acted wrongly in this way? Most of us would answer No. But *Bridge* would not then illustrate (A), since we would not believe that your act would be wrong even though it would be better if you acted in this way. So if we reject Act Consequentialism by making claims like (A), our claims must use the word 'better' in its deontic-value-ignoring sense.

If acts can be deontically good or bad, as some Consequentialists believe, we may object that Consequentialist theories should not tell us to ignore the value of such acts. But like the Deontic Beliefs Restriction, this *Deontic Values Restriction* applies to only some of our moral thinking. Consequentialists make various claims about how the rightness of our acts depends on how it would be best for things to go. It is only *while* we apply these claims that we cannot appeal to our beliefs about deontic values. At other times we can use words like 'good' or 'bad' in their ordinary, all-inclusive senses. In what follows in my text, I shall often such words in their narrower, deontic-value-ignoring senses. But in most cases this distinction makes no difference.

[487] In the sense explained in Section 21 above.

[488] On one version of Motive Consequentialism, the best motives for each person to have are the motives whose being had by *this* person would make things go best. The standard terminology, we can note, is in one way misleading. When Direct Consequentialists apply the Consequentialist Criterion to acts, I have said, these people are Act Consequentialists. But there could be Act Consequentialists who were Indirect Consequentialists, because they applied the Consequentialist Criterion directly to acts, and only indirectly to other things, such as rules or motives. On this view, though the best or right acts are the ones that would make things go best, the best rules are not the rules whose acceptance would make things go best, but the rule 'Always do what would make things go best', and the best motives would

not be the motives whose being had would make things go best, but the motives of an Act Consequentialist.

These various possibilities are very well discussed in Kagan (2000) and Kagan (1998) Chapters 6 and 7.

[489] See, for example, Rawls (2000) 173-6 and 232-4.

[490] If these people themselves accept some desire-based subjective theory about reasons, they would not have the concept of how it would be best for things to go in the impartial-reason-implying sense. But they might want things to go in the ways that would in fact be best in this sense.

[491] When we ask how we would have most reason to want things to go, from an impartial point of view, we may find it hard to decide how strong our reasons are for wanting people not to act wrongly. Would we have stronger reasons to want one person not to be murdered or to want two people not to be accidentally killed? If one person's acting wrongly would prevent several others from acting wrongly, would we have most reason to want, or hope, that the first person acts wrongly? In assessing premise (D), however, we can ignore these questions. When we apply the Kantian Contractualist Formula, or any other such formula, we must set aside our beliefs about which acts are wrong. I shall return to this point below.

[492] For a partial defence of such a principle, see Kamm (2000) and (2004).

[493] We should not assume that, if everyone accepted some moral principle, everyone would always act upon it. But in this imagined case you should assume that, if I accept the Numbers Principle, I would save the five rather than you. I would have no reason not to act on this principle.

[494] What I am rejecting is the view that, in deciding how to act in particular cases, we are rationally required to give equal weight to everyone's well-being. Things are different when we are giving arguments for or against moral principles. When giving such arguments, we ought to give no priority to our own well-being. We can be strong impartialists at this higher level, while rejecting strong impartialism as a view about how we should act. See Barry (1995) Chapters 1, 8, and 9.

[495] In some other imaginable cases, the stakes would be even higher. You might have to choose between saving either yourself or several strangers from many years of unrelieved suffering, in lives that would be worse than nothing. Here too, I believe, you could rationally choose to bear this great burden, if you could thereby save others from such burdens. Such a heroic, noble act would be fully rational.

[496] Bernard Williams (1981B) 18.

[497] Scanlon (1998) 125.

[498] It is sometimes claimed that we could not have impartial reasons to want some people to act wrongly. But that is not so. Our claim should at most be that we always have impartial reasons to want *no one* to act wrongly. On the view just described, it would be best, in *Lesser Evil*, if no one killed anyone as a means, even though the five would then die. But we are supposing that at least one person will act in this way. Though it would be bad if you acted wrongly, by killing me as a means, it would clearly be even worse if Grey and Green both acted wrongly, by each killing two other people as a means. If these are the only possibilities, other people would have more reason to hope that you will act wrongly, since that would be the lesser of two evils. Fewer people would then be killed wrongly as a means. So if other people learn that it is you, rather than Grey and Green, who have acted wrongly, they should regard that as good news. This view also implies that *you* would have impartial reasons both to want yourself to act wrongly, and to act wrongly, by killing me as a means. But these impartial reasons, we might coherently believe, would be decisively outweighed by your other, *person-relative* deontic reasons *not* to act wrongly. If that were true, you would have decisive reasons, all things considered, *not* to do what you had these impartial reasons to do.

[499] I discuss some possible exceptions in Section 81.

[500] ME, Book IV, Chapters III to V.

[501] Kagan suggests a similar argument in Kagan (2002) 128, and 147-150. It is a surprising fact that, though many writers claim that Kant's formula does not support Consequentialism, Kagan is (as far as I know) the first person to ask whether we could rationally will it to be true that the Act Consequentialist maxim be a universal law. (Sidgwick however writes: 'I could certainly will it to be a universal law that men should act in such a way as to promote universal happiness; in fact it was the only law that it was perfectly clear to me that I could thus decisively will, from a universal point of view' (ME xxii).)

Kagan claims that we could rationally will 'a universal law that everyone is to act in such a way as to maximize the overall good', because we would thereby be willing a world in which everyone 'complies with this maxim' by doing what would maximize the good. In arguing that we could rationally will this world, Kagan appeals to claims about instrumental or self-interested reasons. He notes that, in such a world, we might be required to make significant sacrifices for the good of others. Despite this fact, he claims, it would be rational in self-interested terms to will this world, given the 'logical possibility' that we might be in anyone's position. This amounts to assuming a veil of ignorance, as in Rawls's version of Contractualism. Richard Hare gives a similar argument in Hare (1997). These arguments differ in several ways from the arguments that we have been discussing. For another, even

more different argument, see Cummiskey (1996). Kant's texts are inexhaustibly fertile, provoking in different people very different thoughts.

[502] It is easy to overlook our reasons to consider these other effects. Kagan may have thought it enough to claim that AC is the maxim *whose being universally followed* would make things go best. But we should not consider only the effects of this maxim's being *followed*, since we would then take into account only the effects of people's acts, and we would thereby ignore some other important effects, such as the effects of people's being disposed to follow these principles. This point does not apply when we ask which are the maxims or principles whose universal *acceptance* would make things go best.

[503] This would not always be true. As Allan Gibbard, Gerald Barnes, and Donald Regan have argued, AC is sometimes indeterminate, since each of us might be following AC even though we are not together doing what would make things go best. It may be true of each member of some group that, if she alone had acted differently, that would have made things go worse, but that, if everyone had acted differently, things would have gone better. [References.] This complication does not undermine the claims in my text.

[504] That is mainly because, in asking which are the principles whose being universally followed would make th the various ways in which, when people try to make things go best, they often go astray, through miscalculatio like. We can also note that, on some versions of Rule Consequentialism, we appeal to the principles that are community, or during some period. Kantian Contractualism might take this form. If we ask which princip optimific in the 20th and 21st Centuries, these principles would be even closer to AC than they would be in mo extreme inequalities in wealth and power, and great advances in technology, many Act Consequentialists wou centuries, to do far more good than ever before. So AC might now be UF-Optimific. But AC was not optim we could hope that things will change, so that AC would cease to be optimific in future centuries.

[505] I discuss some of these questions in Sections 37 to 43 of Parfit (1984-7). And see again Kagan (2000) and (1998) Chapters 6 and 7.

[506] In Section 34.

[507] It is easy to go astray here. Some writers claim that, if we had to choose between doing our duty and promoting happiness, we ought always to do our duty. But this claim is another trivial truth. We could accept this claim even if we believed that we would never have to make this choice, since our only duty is to promote happiness.

[508] *The First Critique*, A 851 B 879.

[509] *Metaphysik* L1,28:337, cited in Guyer (2000) 94.

[510] These claims, we can note, cannot be put the other way round. We could not defensibly claim that, if everyone could rationally will that some principle be universally accepted, that makes this principle optimific, by making it one of the principles whose universal acceptance

would make things go best. The effects of some principle's acceptance do not depend only on whether this principle's acceptance could be rationally willed. Nor could we claim that (L2) if some principle is the only relevant principle that no one could reasonably reject, that would make it the only relevant principle whose universal acceptance everyone could rationally will. My argument for (L) consists in claims (A) to (I) above, and there is no similar argument, I believe, for (L2).

[511] Scanlon (1998) 11.

[512] As I have said, in claiming that we could justifiably reject some theory, or belief, I do not imply that this theory or belief is false. We can justifiably have some false beliefs.

[513] Though Kantian Rule Consequentialism has different versions, which may conflict, these conflicts are not between the Kantian and Rule Consequentialist parts of this view.

[514] According to Kantian Rule Consequentialists, we ought to follow the optimific principles because these are the only principles whose being universal laws everyone could rationally will. This version of Rule Consequentialism is, in this sense, founded on Kantian Contractualism. As I have also claimed, however, it is because these principles are optimific that these are the principles whose being universal everyone could rationally will. In this other sense, Rule Consequentialism is more fundamental. But there is no contradiction here, since these are two different kinds of dependence.

We can also note that, though Kantian Contractualism provides this firmer foundation for Rule Consequentialism, it is only Rule Consequentialism that could be accepted on its own. It would be only the Rule Consequentialist principles whose being universal laws everyone could rationally will, so Kantian Contractualism succeeds, or is acceptable, only if Rule Consequentialism succeeds.

Note to the Copy editor:

Endnotes 514 to 529 should become footnotes for Allen Wood's Commentary 'HUMANITY AS AN END IN ITSELF' above.

[515] 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.

[516] Henry Sidgwick, *The Methods of Ethics* (Indianapolis: Hackett, 1981), 373-374.

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

[518] 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 Edition. Indianapolis: Hackett, 2001 24-26); (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 distnct stage of thinking that cannot be made a matter of rules or principles.

[519] Sidgwick *The Methods of Ethics*, 359-361.

[520] 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.

[521] 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 the 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.

[522] Hill, Thomas E.*Dignity and Practical Reason* (Cornell University Press, 1992) 56-57.

[523] 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'.

[524] Sidgwick, *Methods of Ethics*, 96-103, 374, 421-422.

[525] 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 which a great many people can find morally acceptable the monstrous conduct, domestic as well as foreign, of the utterly evil regime presently ruling this country.

[526] 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.

[527] 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 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.

[528] A notable exception is Judith Thomson, *The Realm of Rights* (Cambridge, Gauthier (1986): Harvard University Press, 1990), Chapter 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 Derek Parfit for bringing this reference to my attention.

[529] 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 of 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.

[530] Other fans of trolley problems (a different kind of fan of them), admit that they do not elicit moral intuitions that would be of much use in real life, but these fans are struck by the degree of convergence among different people's intuitions about some trolley problems, since this suggests to them that the degree of agreement among people about even such weird examples that are so different from our real-life moral judgments is itself a significant datum that is of psychological interest and requires theoretical explanation. I remain skeptical that convergence among responses to trolley are interesting data of any sort, or that they prove anything at all, except perhaps the very general point, which seems to me to cast serious doubt on a lot of what passes for psychological and sociological research -- namely that people can easily be misled in all kinds of surveys by superficial features of the way questions are posed to them. This suggestion has been made to me by John Mikhail and Marc Hauser, who both think that the convergence of responses to some trolley problems, even across differences in age, gender and culture, constitute evidence for the existence of an innate moral faculty, analogous to the Chomskian innate linguistic faculty, and further that studying responses to trolley problems can help us determine the contents of this faculty.

End of Allen Wood's notes

[531] Wolf adds that, even when we ought to treat people in ways to which they do not consent, such acts are 'always to be regretted'. I agree that I should have made that further claim.

[532] I have added the reference to harming people, which I assume that Wolf intends.

[533] G 432.

[534] ME, viii.

[535] Add some remarks about Thomson?

[536] Wood (2008) 00. See also his discussion in Wood (1999) 000.

[537] Wood (2006) 372 note 2. These defenders of Kant are 'self-appointed', Wood writes, 'because Kant never tries to use the universalizability test as a general moral criterion in the way they are trying to defend.' That, I believe, is not true, given the passages I cite on p 000 above.

[538] Wood (2008) 00.

[539] Wood (2008) preface.

[540] G 431.

[541] Wood (2008) 000.

[542] Wood (2008) 000. Wood is quoting Kant's claim that 'one does better in moral judging always to proceed in accordance with the strict method and take as ground the universal formula of the categorical imperative: *Act in accordance with that maxim which can at the same time make itself into a universal law'*. Most commentators assume that Kant is referring here to his Formula of Universal Law. Wood argues that Kant is referring to his Formula of Autonomy. For an earlier defence of this claim, see Wood (1999) 187-190.

[543] What FA gives us, Wood writes, is only 'a spirit in which to think about how to act. . . not a procedure for deducing. . . principles to act on' (Wood (2008) 00).

[544] My version of this formula appeals, not to what it *would be rational* for everyone to choose, but to what everyone *could rationally* choose. It would be much harder to defend the claim that there is some set of principles that everyone would be rationally required to choose.

[545] G 436. Though this claim may be true of Kant's Formulas of Autonomy and of Universal Law, it cannot cover Kant's Formula of Humanity.

[546] Reference.

[547] Commentary 000.

[548] Wood (1999) 121.

[549] Herman (1993) 210, 212.

[550] Just before this definition, Kant refers to 'the dignity of a rational being' (G 434).

[551] Herman (1993) 238.

[552] Wood (1999) 130.

[553] Kerstein (2002)182.

[554] Korsgaard (1996A) 125.

[555] And Wood earlier wrote: 'humanity, or 'the human being and every rational being in general' is the end-in-itself. . . an ultimate end or value. . the goodness of the end he is seeking is indemonstrable. Hence the argument that humanity is such an end.' [555]

[556] Reference to Wood's Commentary above 000.

[557] Wood (2008) 000.

[558] Wood (2008) 000.

[559] Reference to Korsgaard?

[560] Wood (1999) 127.

[561] Wood (1999) 129

[562] Commentary 000.

[563] G 435.

[564] Commentary 000.

[565] As Richard Dean writes, 'There is an inherent conceptual difficulty in claiming that a capacity has incomparably high value. . . to attribute some value to a mere capacity implies an even greater value for the realized capacity' Dean (2006) 86.

[566] G 406.

[567] Wood (1999) 120.

[568] Wood gives another objection to this view, claiming that we are not morally required to try to act as often as possible out of duty. But Wood answers this objection, claiming that we can act with good wills, and in a way that has moral worth, even when we are not acting out of duty. As Wood notes (In the piece in Schoenecker) 'Even if we doubt, on the grounds I have suggested earlier, that Kant is right that the good will is good without limitation, simply recognizing that the good will is an important good is enough to give us reason to attend to the importance of acting on moral principles.' 2006

[569] CPR 125 and 129.

[570] Commentary 000.

[571] Wood (2008) 000. See also Wood (1999) 000.

[572] Wood (2008) 000.

[573] Wood (2008) 000.

[574] Herman (1993) 213.

[575] Herman (1993) 214.

[576] Commentary 000.

[577] Herman (1993) 124.

[578] Herman (1993) 124.

[579] Herman (1993) 129. This sentence continues '(self) by another'. But I am not my agency.

[580] Wood (1999) 116-7.

[581] Wood (1999) 144.

[582] CPR 78.

[583] This formula, she wrote, can give us 'predeliberative moral knowledge', by showing that there is a moral presumption against acting in certain ways for certain reasons. This task is 'the only one it can perform' (Herman (1993) 147). See also Herman (1993) 112 and 146.

[584] Herman (1993) 104, 132.

[585] She writes: 'It may be that Kant's theory cannot realize its ambitions, but as I hope to show later on in this paper, I don't think the best interpretation of Kant has yet reached that stage of the dialectic.'

[586] Reference.

[587] Cite passages.

[588] MM 219.

[589] Kant's Formula, she elsewhere writes, may be intended only to show that there is a 'deliberative presumption' against acting in certain ways for certain reasons. In this commentary, Herman may be making a different, stronger claim. Kant may intend his formula to give us a criterion of when some act is wrong in the motive-dependent sense, even though such acts may *not* be wrong in the sense of being morally impermissible and contrary to duty.

[590] G 403.

[591] G 404, 424.

[592] Herman herself elsewhere writes 'On a Kantian account, we say that an action is contrary to duty when its maxim cannot be willed to be a universal law' (Herman (1993) 89).

[593] Herman (1993) 34.

[594] G 402.

[595] G 432. And he refers to 'the concept of every rational being as one who must regard himself as giving universal law. . .' But Kant never explicitly appeals to what everyone could rationally will. The phrase just quoted, for example, ends 'through all the maxims of his will' (G 434). If each person regards himself as giving laws through the maxims of *his* will, he is not asking which laws everyone could will. At several other points, when Kant seems about to appeal to what everyone could will, he returns to his Formula of Universal Law, telling us to appeal to the laws that we ourselves could will.

[596] Herman (1993) 95.

[597] Herman (1993) 94.

[598] Herman (1993) 99.

[599] Herman (1993) 118.

[600] Herman (1993) 120 (my italics).

[601] She writes, for example, 'Desires do not give reasons for action: they may explain why such and such is a reason for action. . . but the desire itself is not a reason.' Herman (1993) 194-5.

[602] 8 426 (my italics).

[603] G 423.

[604] Herman (1993) 49.

[605] Nor, we can add, would it be enough to appeal to what people *prefer.*

[606] Herman (1993) 52.

[607] Herman (1993) 54 note 12.

[608] MM 453.

[609] Lectures 233

[610] Herman (1993) 155, 153, 238.

[611] MM 457.

[612] I am very grateful to Herman for correcting several mistakes I made when interpreting her claims. Herman's commentary makes several other very interesting, subtle, and plausible claims. I do not attempt to discuss these claims, in part because they are not directly relevant to my claims and arguments.

[613] This claim is not strictly accurate, since Grey would not be required to make this gift, on Scanlon's Formula, if every principle that required this act could be reasonably rejected by *someone*. This person would not have to be Grey. But given Scanlon's other assumptions, if Grey could not reasonably reject any such principle, nor could anyone else.

[614] Scanlon (1997) page 272.

[615] Scanlon's account of the problem raised by *Case One* is somewhat different from mine. Scanlon suggests that, just as White could reasonably reject any principle that permitted Grey not to give his organ to White, Grey could reasonably reject any principle that required him to make this gift. If that were true, Scanlon suggests, there would be 'a moral standoff', in which there was 'no right answer' to the question of what Grey ought to do. We would 'solve this problem', Scanlon writes, if we claimed that no one could reasonably reject any

optimific principle, so that Grey could not reasonably reject the optimific principles that would require Grey to give his organ to White. But this solution, Scanlon remarks, would have 'a cost', since it is intuitively implausible to claim that Grey is required to make this gift. I shall ask here whether we could solve this problem in a way that avoids this 'cost', by defending the claim that White could *not* reasonably reject some principle that permits Grey not to make this gift.

[616] Scanlon (1998) 229.

[617] Scanlon (1998) 212, and elsewhere. [Scanlon's claims about fairness do in a less direct way appeal to claims about well-being.]

[618] Scanlon (1997) 267. He also says that he is one of those 'who look to Contractualism specifically as a way of avoiding Utilitarianism' (1998) 215.

[619] Scanlon (1998) 235.

[620] Scanlon (1998) 241.

[621] Scanlon (1998) 230.

[622] Strictly, when applying Scanlon's Formula, we consider the objections to such principles that would be had, not by two particular people, but by any of the people who, in cases of this kind, would be in positions that are relevantly similarly to the positions of these two people. This complication does not affect my claims.

[623] Scanlon (1998) 240.

[624] According to Telic Egalitarians, inequality is in itself bad. When benefits come to people who are worse off, that is in one way better because it reduces the inequality between different people. [624] This view is open to the *Levelling Down Objection*. Suppose that those who are better off all suffer some misfortune, and become as badly off as everyone else. Telic Egalitarians must admit that, on their view, these events would be in one way a change for the better, because there would no longer be any inequality, even though these events would be worse for some people and better for no one. Many people would find these claims hard to accept. The Priority View avoids this objection. Because this view does not assume that inequality is in itself bad, this view does not imply that it would be in any way better if those who are better off became as badly off as everyone else.

[625] Scanlon writes: 'where the base line is equal, benefiting only Blue seems objectionable, because all have the same claim to some benefit' (Stratton-Lake (2004) 131).

[626] These claims apply only to those cases in which both (1) the baseline is equal and (2) we can give much greater benefits to some people than to others. If we could give equal benefits to each person, as is often true, no one could reasonably reject a principle requiring us to give everyone such benefits. But cases in which (1) and (2) are true, though they are much less common, help us to see more clearly what is distinctive in the version of Scanlon's view that includes his Individualist Restriction.

[627] Scanlon (1998) 229.

[628] Scanlon (1997) 123.

[629] Scanlon imagines a case in we have to choose between these outcomes:

	Future months of pain	
	for A	for B
(A)	61	0
(B)	60	2

He then writes: 'the way in which A's situation is worse strengthens her claim to have *something* done about her pain, even if it is less than could be done for someone else' (Scanlon (1998) 227). Since Scanlon refrains from saying that we ought to give A her lesser benefit, though A's situation is *much* worse than B's, Scanlon here gives very little weight to distributive principles.

[630] Scanlon (1998) 239-40.

[631] Scanlon (2001) 200.

[632] It might be suggested that the burden of acting wrongly, if we were in Grey's position, would outweigh the burden of not receiving the many more years of life if we were in White's position. But this principle would not impose on us the burden of acting wrongly. We could avoid that burden by giving away our organ, and thereby losing a few years of life. And that would be a smaller burden than losing many years of life.

[633] Refer to Nagel, *Equality and Partiality*,

[634] references.

[635] He writes: 'I should have avoided describing Contractualism as an account of the property of moral wrongness. . . This claim. . . can be dropped from my account without affecting the

other claims I make for Contractualism' (Stratton-Lake (2004) 137). He also writes:'The fact that an action would cause harm may make it reasonable to reject a principle that would permit that action, and thus make that action wrong in the Contractualist sense I am describing. It is also true that an action's being wrong in this sense makes it morally wrong in the. . . general sense of that term' (Stratton-Lake (2004) 136). For a longer discussion, see Scanlon (2007B).

[636] Scanlon (1998) 222.

[637] Scanlon (1998) 182.

[638] Scanlon (1998) 219.

[639] Comment on Scanlon's discussion of this claim.

[640] Scanlon (1998) 186.

[641] Scanlon (1998) 168.

[642] Parfit (1984-7) Chapter 16.

[643] There is another view that should be mentioned here. We might claim that, if some act would indirectly cause someone to exist who would have a life worth living, this act would thereby benefit this person. According to what we can call

> *the Wide Person-Affecting View*: Other things being equal, one of two acts would be wrong if it would benefit people less.

If causing to exist can benefit, this view rightly implies that, in *Case Four*, our three possible acts are morally equivalent. The benefits to Tom and Dick of our doing A would be equal to the benefits to Tom and Harry of our doing B, which would be equal to the benefits to Dick and Harry of our doing C.

Though the Wide Person-Affecting View provides one fairly plausible answer to the Non-Identity Problem, this view is irrelevant here. First, this view does not revise the Two-Tier View. In the cases that we are considering, this view coincides with the No Difference View. And we have other reasons not to appeal to this view. [More to be added.]

[644] In the simplest case, that would be true of three people whose preferences were these:

> White prefers B to A, and C to B
>
> Grey prefers C to B and A to C
>
> Black prefers A to C and B to A.

White and Black prefer B to A, White and Grey prefer C to B, and Grey and Black prefer A to C.

[645] This case counts against some other widely accepted principles. For example, according to

> *the Pareto Principle*: One of two outcomes would be worse, and one of two acts would be wrong, if this outcome or act would be worse for some particular people and better for no one.

This principle unequivocally implies that, since A would be worse than B, which would be worse than C, and so on down the series, the best outcome would be G, and G would be what we ought to do.

[646] Stratton-Lake (2004) 128.

[647] Stratton-Lake (2004) 128.

[648] This problem would not be solved if Scanlon appealed instead to the non-comparative account of benefits and burdens. On this account, A and B would be morally equivalent, since cancelling either program would impose on equal numbers of people the burden of living for only 40 years. Nor would it help to appeal to people's rights.

[649] Scanlon (1998) 219.

[650] This point is even clearer when we turn to cases in which different numbers of people might exist. Scanlon includes, among the acts that his formula condemns, irresponsible procreation. He may be thinking only of cases like that of Jane, who has a child when she is too young to give this child a good start in life. But he may also have in mind those who have very many children, with the result that their children are very poor, though having lives that are worth living. We may believe that it would be better if, instead of having ten children, this couple have only two or three children. But, if this couple have ten children, we should not claim that it would have been, in the relevant sense, better for these ten children if there had been only two of them.

[651] Scanlon (1998) 186-7

[652] Parfit (1984-7) Sections 124-6.

[653] Reference.

[654] Stratton-Lake (2004) 133.

[655] Rawls (1971) 25.

[656] Refer to Nagel's claims in *Equality and Partiality*.

[657] It might be claimed that, in *Case Four*, the Kantian Formula supports a principle that is not optimific, since we would not make things go better if we gave each of these people an equal chance of being saved. But this can be denied. It may be better, because fairer, if people are given such chances. And even if this act would not make things go better, the Kantian Formula would here merely be supporting one of a set of principles which are all optimific.

[658] Another example is provided by Liam Murphy's claims about the demandingness of morality. MORE.

[659] The word 'property' can also be used more narrowly, so that it refers only to instantiated properties or to properties that can have causes and effects.

[660] This claim uses the word 'happiness' in some naturalistic sense which involves no value judgment, such as the judgment that egoists or sadists cannot be truly happy.

[661] There are other claims which use normative concepts, but are not in this sense normative. One example is the claim that acts are right if they are not wrong. This claim merely states how these concepts are related, and neither states nor implies that anyone has any reason to act in some way. Though in one sense normative, this is not a *substantive* normative claim.

[662] Korsgaard (1996B) 85. Korsgaard continues: 'What the argument. . . actually seems to do is to prove that if there were any Utilitarians then their morality would be normative for them'. Korsgaard seems here to mean 'would motivate them'.

[663] Anderson (1991) 21. Anderson also writes 'These agents do not find the perspective of quantitative Hedonism to have normative force: upon reflection, they are unwilling to sacrifice the higher pleasures for any of the lower. No [such] agent, on Mill's view, can be moved by quantitative Hedonism'.

[664] Though Moore himself did not distinguish between concepts and the properties to which they refer. Moore was following Sidgwick, who made the relevant claims more accurately.

[665] As Williams writes, 'I think that the sense of a statement of the form 'A has a reason to do X' is given by the internalist model' ('Internal Reasons and the Obscurity of Blame', in Williams (1995) 40. (I have substituted 'do X' for 'phi'.) See also 'Internal and External Reasons', in Williams (1981). These articles contain many similar remarks. In some passages quoted below, Williams discusses how we should define the term 'reason' and what claims about reasons mean. He also writes: 'What are we saying when we say that someone has a reason to do something? . . . we do have to say that in the internal sense he indeed has no reason to pursue these things. . . . if we become clear that we have no such thought, and persist in saying that the person has this reason, then we must be speaking in another sense,

and this is the external sense. . . What is that sense?. . . In considering what the external reason statement might mean. . . . '.

[666] Darwall similarly writes that, on his view, 'the content of the judgment that there is reason for one to do X is simply that were one rationally to consider facts relevant to doing X, then one would be moved to prefer doing X. (Darwall (1983) 128 (with 'A' replaced by 'X').

[667] Williams calls this decisive-reason-implying sense of 'ought' 'the practical or deliberative sense', and he writes: 'Since "A ought to do X" in the practical sense is relativised to the agent's set of aims, projects, objectives, etc. . . it follows that if a given claim of this kind is based on the assumption that A had a certain objective which he does not have, and if there is no sound deliberative route to that objective from objectives that he does have, then the claim is wrong' (1981 120). Williams also writes, 'If A tells B that he ought to do a certain thing, but A is under a misapprehension about what B basically wants or is aiming at, then A's statement, if intended in this sense, must be withdrawn' (*1981* 124). Falk discusses these senses of 'ought' and 'should' in many of the articles reprinted in Falk (1986).

[668] This formulation is intended to cover Williams's remark that, when we say that someone has a reason to do X, we mean something like 'A could reach the conclusion that he should do X (or a conclusion to do X) by a sound deliberative route from the motivations he has in his actual motivational set' (1995 35). Though Williams writes only that A 'has a reason to do X', his later use of 'should do X' shows that he is discussing a decisive reason, and what he calls the 'practical' sense of 'should' and 'ought'. We need not here discuss Williams's claim that A's motivations must *already* be in A's actual 'motivational set', rather than being motivations that A might acquire while deliberating on the relevant facts.

[669] If we use 'external' merely to mean 'not internal', there might be other external senses of the phrase 'has a reason'. Some of these might be naturalistic senses. According to a hedonistic naturalistic form of Rational Egoism, for example, the claim that we have decisive reasons to act in some way might be held to mean that this act would maximize our own happiness. But though there is conceptual space for such naturalistic external senses of the word 'reason', such senses are seldom proposed, and have little importance.

[670] Williams (1995) 104. Williams claims only that you need to take this medicine to preserve your health. I have added that, if you don't preserve your health, you will lose many years of happy life. That further assumption would not alter Williams's view about this example.

[671] Williams gives some other arguments, which I discuss briefly near the end of Chapter 4, and in Parfit (1997). Some of these arguments are aimed at some proposals about what it might mean to claim that someone has an external reason. But these proposals do not describe the indefinable irreducibly normative sense of the phrase 'has a reason' that Scanlon, I, and others believe that we use. If we can use the phrase 'has a reason' in this external sense, our claims about such reasons are untouched by these arguments.

[672] For example, Williams considers someone who maltreats his wife, and whose attitudes and acts would not be altered by informed and rational deliberation. Externalists, Williams writes, will want us to say that this man has a reason to treat his wife better. 'Or rather, the external reasons theorist *may* want me to say this: one of the mysterious things about the denial of internalism lies precisely in the fact that it leaves it quite obscure when this form of words is thought to be appropriate. . . What is the difference supposed to be between saying that the agent has a reason to act more considerately, and saying one of the many other things we can say to people whose behaviour does not accord with what we think it should be? As, for instance, that it would be better if they acted otherwise. I do not believe, then, that the sense of external reason statements is in the least clear. . . ' (1995) 39-40. And Williams writes elsewhere that externalists do not 'offer any *content* for external reasons statements' (1995B) 191 (my italics).

[673] Darwall (1983) 210-11.

[674] Darwall (1983) 128. In these quoted passages, Darwall is not describing his own view, which I shall discuss later.

[675] (1995) 36.

[676] I do not mean to imply that only natural facts can give us reasons. Some normative facts can also give us reasons. But my distinction still applies.

[677] I follow Scanlon (1998) 20.

[678] Falk (1986) 35, 184.

[679] Darwall (1983) 134

[680] Darwall (1983) 128.

[681] Darwall (1983) 86. As we shall see, however, Darwall's final version of Analytical Internalism is not a form of Analytical Naturalism.

[682] Darwall, for example, makes such claims (in discussion). It is unclear whether Williams would make these claims. He comes closest to doing that in WME, but..

[683] Falk (1950) 80.

[684] Falk (1986) 48, 62-3.

[685] Falk (1986) 65.

[686] Falk (1986) 66. When Falk discusses a case like *Revenge*, he writes: 'That "causing you hurt will revenge me" may prove a strongly persuasive consideration. . . But this need still not be more than a 'bad' or 'insufficient' reason for doing what this consideration is tempting me to do. For it may still be that, if I still made way in my thoughts for a more faithful and less passion-distorted view of the act. . . I would cease to find it choice-influencing altogether. The consideration would be a "bad" reason and an inferior guide for lack of "true" power of influence' (Falk (1986) 93).

[687] Falk (1986) 34.

[688] (1995B) 16.

[689] Williams similarly writes that certain reasons 'are not, as it turns out, the strongest reasons for me, now: the strongest reason is that I desire very much to do something else' (1985) 19. External reasons cannot *turn out* to be strong or weak.

[690] Darwall (1983) 80. He also writes: 'If something's being a reason is simply a non-natural property of it of which we take notice in judging the consideration to be a reason, then the desire to act for reasons is in no sense integral to the self. It is a fascination with a nonnatural property that one may have or lack without any change in the self. So understood, the desire to act for reasons is unintelligible' (Darwall (1983) 57).

[691] Smith (1994) 57.

[692] Boyd (1997) 119.

[693] The word 'property', we can note, is here used broadly, so that it can be used in describing all normative facts. When someone ought to act in some way, for example, we could say either that this act has the property of being what this person ought to do, or that this person has the property of being someone who ought to act in this way. We can similarly say that some fact has the property of giving someone a reason.

[694] For one version of this argument, see Jackson (1998) 122-129.

[695] When Jackson gives this argument, he appeals to the claim that, since triangles are *equilateral* just when they are *equiangular*, these concepts refer to the same property. When applied to this example, this view has some plausibility. These triangles have a single shape that can be described in these two ways. No such claim applies to the concepts of *being the only even prime number* and *being the positive square root of 4*.

[696] The concept of *the property that makes acts right* is irreducibly normative because this concept contains the concept *right*. If this more complex concept were not normative, (F) would not be a normative claim, as its restatability as (G) shows it to be.

[697] When certain natural properties of acts would make these acts right, the rightness of these acts is often claimed to *supervene* on these natural properties. Mental states, it is similarly claimed, supervene on states of the brain. Though these two kinds of supervenience are in some ways similar, they also differ greatly, I believe, in other ways. Normative supervenience should be considered on its own.

[698] For two such arguments, see Smith (1994) and Boyd (1997).

[699] Schroeder (2007) 75-8. Say why the analogy is only partial.

[700] Sturgeon (2006) . . To put this distinction in a different way: While Sturgeon claims that normative facts may be natural facts even if we *cannot* be confident that we shall *ever* be able to restate these facts in non-normative terms, this definition implies that normative facts are not natural if we *can* be confident that we shall *never* be able to restate these facts in these terms. These claims do not conflict.

[701] Like many Naturalists, Sturgeon seems here to ignore the difference between rightness and the property that makes acts right. To illustrate how Moral Naturalism might be true, it is not enough to suppose that acts are right just when they maximize pleasure. What we are supposing might be true because, when acts maximize pleasure, that makes them have the different property of being right. That would not help to show how rightness might be a natural property.

[702] Sturgeon makes some relevant remarks, which I shall discuss in the unwritten Section 8.

[703] Sturgeon writes: ' if ethical naturalism is defended by the [causal] argument I have considered, it can remain neutral on the question of whether we can ever find reductive naturalistic definitions for ethical terms.' Sturgeon here concedes that, if his theory's claim to be Naturalist cannot be defended by appeal to the Causal Criterion, his theory could not remain neutral about the possibility of giving reductive definitions. As Sturgeon also writes 'Perhaps ethics could then be plausibly required to earn its place [within a Naturalist view] by another route'.

[704] Refer to Sturgeon, and to Cuneo's remarks about virtues.

[705] (As I argue in Appendix A, this kind of explanation would not be wholly different from our most fundamental naturalistic explanations.)

[706] But they chose the right number, as when we speak of a *square deal*. It would have been less plausible to claim that Justice was the number 13.

[707] Gibbard (2006) 323.

[708] Some claims do, in one sense, use this normative phrase or concept, without being normative. One example would be the claim that Sidgwick believed that maximizing happiness was the property that makes acts right. This claim is not normative, since it is a merely natural fact that Sidgwick had this normative belief. But this claim, we can say, merely *mentions* this phrase, and the property to which this phrase refers, without claiming that anything *has* this normative property. Sentences like (F), in contrast, *use* this phrase, by claiming that some acts have this property.

[709] In (Q) the phrase 'the writer of *Hamlet*' is a *rigid designator*, meaning 'the actual writer of Hamlet'. This phrase would refer to Shakespeare even in the possible worlds in which he didn't write *Hamlet*. It might be claimed that, even in the referential sense, (P) and (Q) do not state the same fact, since only (Q) ascribes to Shakespeare the property of being the writer of Hamlet. On this criterion for the identity of facts, however, we regard the phrase 'the writer of Hamlet' as merely one way of referring to Shakespeare, and we ignore the other information that this phrase gives us.

[710] Williams, for example, seems to accept this view, and Darwall explicitly accepts it (in conversation) .

[711] Gibbard (2006) 329.

[712] Moore *Principia Ethica* edited by Thomas Baldwin, (Cambridge University Press, 1993) 64.

[713] Gibbard (2006) 328.

[714] This analogy is less close than the analogy with the discovery that heat is molecular kinetic energy. Water isn't a property, but a stuff, and 'water' could be used as a name, to refer to 'that stuff there, in that lake, or the stuff that falls from clouds as rain'. We already know that this stuff has various properties, so when we learn that water is H2O, that tells us indirectly about the relations between having this molecular composition and these various other properties.

[715] Some of Gibbard's claims suggest this other view. Though Gibbard writes, 'no explanatory purpose would be served by supposing an extra property', he also writes 'the identity claim itself works as the *start* of the explanation' (op.cit. 329, my italics). Gibbard might say that, to complete this explanation, the Utilitarian Naturalist could claim that acts that maximize happiness have some other normative property which is different from the property of being what we ought to do. But this Naturalist would then have to identify this other normative property with some natural property, and my objections would apply to this further claim.

[716] Some Naturalists might say that the property of being what we ought to do is not the same as having one of these natural properties, but *consists* in this set of properties. If some act

has one of these natural properties, that would *constitute* this act's being what we ought to do. That is like the way in which, if I have only one child, who is a girl, my being a parent *consists* in my having a daughter. These properties are not the same, because I could also be a parent by having a son. My argument could be restated to apply to such views.

[717] This argument may seem to assume that, for some claim to be substantive, it must tell us about the the relation between different properties. That is not so. Some substantive claims tell us only that there are some things that have a certain property. Two examples are:

(4) There are some acts that are forbidden by God,

(5) There are some acts that are wrong.

These claims are substantive, as is shown by the fact that they would be rejected by some people. Atheists would reject (4). Moral Nihilists would reject (G). Another such claim is

(6) There are some acts that are disallowed by the only set of principles whose universal acceptance everyone could rationally choose.

Some people would reject (6) because they believe that there are no such principles. Another such claim is

(7) There are some acts that would maximize happiness.

Some people would reject (7) because they believe that interpersonal comparisons of hedonic well-being make no sense. Since these people deny that some people can be happier than others, they believe that there could not be any truths about which acts would maximize the sum of happiness that would be had by different people. [717]

Consider next

(8) Wrong acts are wrong.

This claim, I earlier wrote, is not substantive, but trivial. But if (8) were taken to imply that some acts are wrong, this claim would be in one way substantive. (8) is wholly trivial only if (8) means

(9) If certain acts are wrong, these acts would be wrong.

Though Nihilists deny that any acts are wrong, they would accept (9).

Return now to the Utilitarian Naturalist claim that

(B) when some act would maximize happiness, that is the same as this act's being what we ought to do.

If (B) is intended to imply that there are some acts that would maximize happiness, this claim is in one way substantive. As I have just said, some people deny that any acts could have this property. But this disagreement is irrelevant here. Of those who are neither Utilitarians nor Naturalists, many believe that some acts would maximize happiness. We are asking whether, if we already have that belief, (B) might give us further information, thereby stating a substantive normative view.

[718] This is implied by what Schroeder calls 'Biconditional' Schroeder (2007) 57. Schroeder adds many qualifications to this claim, but these are irrelevant here.

[719] Schroeder (2007) v, 65, and 86-7.

[720] This is the claim that Schroeder calls 'Reason' in Schroeder (2007) 59.

[721] Schroeder (2007) 95-6.

[722] Schroeder (2007) 60.

[723] Darwall (1992) 168. (Darwall's sentence continues 'perhaps when the agent's deliberative thinking is maximally improved by natural knowledge.') Darwall's claim seems an overstatement, since these Metaphysical Naturalists might describe some kinds of normativity in rule-involving or attitudinal terms. But Darwall may be right to assume that, when these people discuss reasons, their most plausible move is to identify normative and motivating force.

[724] Schroeder's Chapter 4 makes these points well.

[725] Sturgeon, in .

[726] Jackson (1998) 124-5. Jackson also writes: 'all there is to tell about moral nature can be told in naturalistic terms' (1992) Section 4).

[727] Railton (2003) xvii-xviii.

[728] Discuss the relation between Hard Naturalism and Analytical Naturalism.

[729] Soft Naturalists might retreat to the view that, though there are some irreducibly normative facts, these facts are also, in some wider sense, natural facts. As I have argued, however, this form of Naturalism is not worth discussing, since such Naturalists would accept the main claims of Non-Naturalist Cognitivism.

[730] I take this example from Gibbard (1990).

[731] Brandt (1992) 35-6

[732] Brandt (1992), 29.

[733] Jackson (1998) 127.

[734] Jackson (1998) 142.

[735] As Hume, for example, writes: 'when you pronounce any action or character to be vicious, you mean nothing but that. . . you have a feeling or sentiment of blame' (David Hume, *A Treatise of Human Nature*, Book III Section I, 15). (Hume may not have meant this literally.)

[736] As Hume also writes, 'Vice and virtue, therefore, may be compared to sounds, colours, heat and cold, which. . . are not qualities in objects, but perceptions in the mind.'

[737] As Hume writes: 'Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason' (*A Treatise*, Book III, Section I, 3).

[738] Nagel (1970).

[739] For a fuller, partly similar response, see Copp (2001).

[740] Another such writer is R. M. Hare, whose *Universal Prescriptivism* is inspired by Kant. On Hare's view, moral claims are like universal imperatives or commands, which tell everyone to keep their promises and not to lie. In his final statements of his theory, Hare argues that, if we ask which universal commands we can honestly accept, we would all reach the same Utilitarian answers. Since we would reach the same answers, we can claim these answers to be true. Hare's theory can thus be regarded as a version, not of Non-Cognitivism, but of *Kantian Constructivism*. See Hare (1981) and (1997).

[741] Gibbard (2003) 194. Moore's main contribution, Gibbard also writes, was to ask 'What. . is at issue in moral disputes? What does the disagreement consist in?'

[742] Blackburn (1998) 49, 275, 90.

[743] Blackburn (1998) 69.

[744] Gibbard (2003) 74.

[745] Gibbard (2003) 184.

[746] Gibbard (2003) ix-x.

[747] Gibbard (2003) x.

[748] Gibbard (2003) 10.

[749] Gibbard (2003) 254.

[750] Gibbard 2003, 9.

[751] From the *Phil Phen* Symposium.

[752] Gibbard (2003) 9.

[753] Gibbard (2003) 270.

[754] Gibbard (2003) 273, 271.

[755] Gibbard (2003) 54.

[756] Gibbard (2003) 268-74.

[757] Gibbard might reply that, when we are tempted not to do what we have planned, we shall be more likely to act on our plan if we believe that this is what we ought to do. But this reply would not help Gibbard to explain the concept *ought* by appealing to the idea of adopting plans.

[758] Gibbard (2006) 77.

[759] As Gibbard himself writes: 'For anything I've claimed, a convenient interpretation might be no more than a convenient fiction---like the stupidities we attribe to the computers on our desks.' Though convenient fictions can have some uses, they are not relevant here.

[760] Gibbard (2003) 17, x.

[761] Blackburn (1984) 197. Blackburn's quasi-realism, he also writes, attempts to practise alchemy by transmuting 'the base metal of desire into the gold of values' (*Phil.Phen* July 2002).

[762] Gibbard (1990) 287.

[763] Blackburn (1998) 309.

[764] Blackburn (1998) 118 note 36.

[765] Blackburn (1998) 318 (my italics).

[766] Gibbard (2003) 65.

[767] Blackburn (1993) 20.

[768] Blackburn (1998) 318.

[769] Blackburn (1998) 117. He elsewhere writes, surprisingly, 'I think this view is confirmed if we ask: could one not work oneself into a state of doubting whether the capacities generating moral attitudes are themselves so very admirable? The answer is that one could, but that then the natural thing to say is that morality is all bunk and that there is no pressure toward objectivity for the quasi-realist to explain' (Blackburn (1993) 20).

[770] I follow Shafer-Landau (2003) 28-9.

[771] Blackburn (1998) 313.

[772] Egan (2007).

[773] Blackburn (1998) 318.

[774] As he writes: 'No, no no, I do not say that we can talk as if kicking dogs were wrong, when 'really' it isn't wrong. I say that it is wrong (so it is true that it is wrong, so it is really true that it is wrong, so this is an example of a moral truth)' Blackburn (1998) 319).

[775] As Blackburn writes: 'The projectivist can say this vital thing: that it is not because of our responses. . that cruelty is wrong' (Blackburn (1993) 172). He also writes: 'One ought to look after one's young children, whether one wants to or not. But that is because we insist on some responses from others, and it it is sometimes part of good moralizing to do so' (Blackburn (1993) 177). But he could withdraw this claim.

[776] Blackburn (1993) 129.

[777] *Inquiry* 1999?

[778] Blackburn (1993) 173.

[779] Blackburn (1998) 50.

[780] In some passages, I believe, Gibbard, also fails to distinguish correctly between internal moral claims and external meta-ethical claims. For example, he writes: 'Are oughts, then, matters of fact? In a minimalist sense of the term 'fact', there are of course facts of what a person ought to do.' (*Phil. Phen. Symposium*). If I claim that we ought to keep our promises,

Gibbard could say that what I claim is true, or is a fact, since Gibbard's minimalist use of the terms 'true' and 'fact' would here express agreement with my moral claim. But when Gibbard claims that there are facts about what people ought to do, that claim is not moral, but meta-ethical. On Gibbard's meta-ethical view, I believe, there are no facts or truths about what people ought to do. In defending his partly similar version of Non-Cognitivism, Mark Timmons, I believe, correctly describes the relation between these moral and meta-ethical (or, as he calls them, 'metaphysical') claims. Timmons writes: 'the two most obvious perspectives from which to judge the correct assertibility of moral statements are what we can call the *detached* perspective and the *engaged* perspective. . . Given my irrealist story about moral discourse, when one judges from a morally detached perspective, and thus simply in the light of semantic norms, moral statements are neither correctly assertible nor correctly deniable, and so they are neither true nor false' Timmons (1998).

[781] From the reply to Egan. Ask Simon for permission to quote from this.

[782] Blackburn (1993) 4, 20.

[783] Gibbard (1990) 8.

[784] Gibbard (1990) 70, 46.

[785] Gibbard (2003) 9-10. More exactly, Gibbard says that 'ought' here adds nothing.

[786] Gibbard (1990) 68-76.

[787] Gibbard (1990) vii.

[788] Gibbard (1990) 7.

[789] Gibbard (1990) 9.

[790] Gibbard (1990) 153.

[791] Gibbard (1990) 172

[792] Gibbard (1990) 173.

[793] Gibbard (1990) 175.

[794] Gibbard (1990) 177.

[795] Gibbard (1990) 33.

[796] Gibbard (1990) 8.

[797] Gibbard (2003) 17, x.

[798] Hare (1972) 33-4.

[799] Hare (1972) 40.

[800] Hare (1952) 195.

[801] Blackburn (1998) 70. Blackburn writes 'for any fact'; but, since he is defending Expressivism about normative claims, he must intend his remark to apply to what Realists claim to be normative facts.

[802] Gibbard (2003) 98.

[803] Gibbard (2003) 15.

[804] Gibbard (2003) 16. Blackburn makes similar claims. See, for example, Blackburn (1998) 87.

[805] Nowell-Smith (1954) 319-20.

[806] Nowell-Smith (1954) 61.

[807] Williams (1981B) 122. (I have expanded some abbreviations.)

[808] Hare (1981) 217.

[809] Korsgaard (2003) 112.

[810] What Korsgaard calls *normative realism* differs from Non-Platonic Intuitionism, of the kind that seems to me better, by making positive and perhaps Platonistic metaphysical claims. This difference is irrelevant here.

[811]Korsgaard (1996B*)* 38.

[812] Korsgaard (1996B) 44.

[813] Korsgaard (1996B) 41 note 68.

[814] For a longer discussion of Korsgaard's view, see Parfit (2006). Even there I say little about some of the most original and central features of Korsgaard's rich and complex view, such as her claims about our practical identity.

[815] Korsgaard (1997) 240.

[816] Korsgaard (1997) 240, my italics. Korsgaard similarly writes: '. . . a realist account of the *normativity* of the instrumental principle is incoherent. For think how the account would have to work. The agent would have to recognize it, as some sort of eternal normative verity, that it is good to take the means to his ends. How is this verity supposed to *motivate* him?' (Korsgaard (2003) 110, my italics).

[817] Korsgaard (1996A) 163-4.

[818] As Nagel writes, 'Only a justification can bring the request for justifications to an end' (Nagel (1997) 1-6.

[819] In my remarks about this question, I am merely summarizing, and oversimplifying, what others have claimed. See, for example, Leslie (1989).

[820] Of several discussions of these questions, I owe most to Leslie (1979) and Nozick (1981); then to Swinburne (Oxford, 1979), Mackie (1982) Unger (1989), and some unpublished work by Stephen Grover.

[821] Credit for such cases may be due to Kavka (1986).

[822] For a similar appeal to the difference between such questions, see Hieronymi, (2005). See also Hieronymi (2006). Hieronymi does not, however, conclude that there are no state-given reasons.

[823] Gauthier (1975). This argument's fullest statement is Gauthier (1986).

[824] In an unpublished paper 'Rational Irrationality', and later in Sections 7-8 of Parfit (1984-7).

[825] This appendix was written in 1994, in response to Gauthier (1997). I have not tried to take into account Gauthier's most recent work.

[826] Since Gauthier means by our *utility* the fulfilment of our *present* considered preferences, what he appeals to is, strictly, the *Deliberative Theory*. But as Gauthier remarks (Gauthier (1986) p. 6), most of his claims apply equally to Rational Egoism. And Gauthier often uses words, like 'benefit' and 'advantage', that refer more naturally to our interests rather than our present preferences. So we can here ignore the differences---though they are often great---between the Deliberative and Self-interest Theories. We can suppose that, in all of the cases we discuss, our present considered preferences would coincide with what would be in our own interests.

[827] What is expectably-best may not be the same as what we can expect to be best. Some acts are expectably-best for us though we can know, for certain, that they will not actually be best for us. Trying to do what is actually best may be, given the risks, irrational.

[828] *Reasons and Persons*, Sections 7-8.

[829] Gauthier gave this reply in Gauthier (1986) (especially, 173-4). In Gauthier (1997), Gauthier later gave up the claim that we could not deceive others. He suggested that, if we remained self-interested, and merely appeared to be trustworthy, that would be worse for us. Thus he writes: 'the overall benefits of being able to promise sincerely. . . may reasonably be expected to outweigh the overall costs of keeping promises when one could have gotten away with insincerity' (p. 26). But if we could get away with insincerity, what are the benefits from being able to promise sincerely? Gauthier might appeal, like Hume, to the benefits of peace of mind, and a good conscience. But that seems insufficient for his purposes. Gauthier also claims that, even if we were generally trustworthy, we would be able to make some insincere promises. But this merely limits the costs of sincerity. It does not suggest that there is any gain. For Gauthier's distinctive argument to get off the ground, he needs, I believe, his earlier assumption that we could not rationally hope to deceive others.

[830] See, for example, Gauthier (1986) Chapter VI.

[831] In *Reasons and Persons*, Sections 7-8.

[832] I also supposed that it might be rational to change our beliefs about rationality. This, too, was intended to help Gauthier's argument. If we did not change our beliefs, we would be doing what we believe to be irrational, and that might seem enough to make our acts irrational. But this element need not concern us here.

[833] As he wrote (like Queen Victoria), 'We are unmoved' (Gauthier (1986), p. 185).

[834] Gauthier asserted (B)---which he calls his 'second level of commitment'---in Gauthier (1997) 40. I discussed a similar claim, which I called '(G1)' (in Parfit (1984-7) 13). On Gauthier's second level of commitment, it is rational to act on a disposition 'so long as one reasonably expects past and prospective adherence to the disposition to be maximally beneficial'. This claim may seem to mean 'if one both reasonably believes that adherence to this disposition in the past has been beneficial, and reasonably expects that adherence to it in the future will be beneficial'. But this cannot be what Gauthier intends, since it would remove the difference between his second level of commitment and his first level (discussed below). Gauthier must mean: 'if one can reasonably believe that acquiring it was beneficial in one's life as a whole, taking the past and future together.'

Gauthier's move from (A) to (B), or from his third to his second level of commitment, hardly damages his defence of rational morality. On the view defended in Gauthier (1986), for morality's constraints to have rational force for us, accepting these constraints must have been expectably-best for us. On Gauthier's revised view, for these constraints to have rational force, they must also be known not to have been on the whole bad for us. Most of contractual morality's constraints would meet this second requirement.

[835] Perhaps I would have obeyed some order that would have proved fatal.

[836] It may be objected that I acquired too crude a disposition. Perhaps I should have become disposed to ignore threats, except in cases in which I believed that acting in this way would be disastrous. But as Gauthier says, 'I may reasonably have believed that any qualification [to my disposition] would reduce its *ex ante* value, so that unqualified threat-ignoring offered me the best life prospects' (Gauthier (1997) 39). We can add the assumption that only the unqualified disposition would in fact have been as good for me. (There is another reason not to allow this disposition to take this qualified form. If we did, we must allow similar qualifications to the disposition of trustworthiness. As we shall see, that would undermine Gauthier's argument.)

[837] Gauthier endorses the action of a would-be deterrer who, when deterrence fails, disastrously carries out her threat. He writes 'Her reason for sticking to her guns. . . is simply that the expected utility. . . of her failed policy *depended* on her willingness to stick to her guns' (Gauthier (1984) 489.) So what? Her expectation may have depended on that willingness. But why should she remain faithful now?

[838] Note that, in claiming this, I need not appeal to Rational Egoism. I need not assume that this attempt would be rational because it would be likely to be good for me. Since Gauthier rejects Rational Egoism, that would beg the question. But even on Gauthier's theory, it would be rational for me to try to lose this disposition. Suppose that I lose my dispositions whenever they become disastrous. It would be in my interests to have this meta-disposition. So, on Gauthier's theory, it would now be rational for me to act upon it.

[839] Suppose first that, if I tried, I could cease to be a threat-ignorer. As I have just argued, it would then be irrational for me to keep my disposition. If Gauthier accepts this conclusion, could he still assert (B)? Could he claim that, even though it would now be irrational to *keep* my disposition, it must still be rational to act upon it?

There may be certain cases in which, though it would be irrational to keep some disposition, it would still be rational to act upon it. Suppose, for example, that it would be irrational for me to remain prudent. If I did, irrationally, keep this disposition, it might still be rational to act upon it, doing whatever would be best for me. (B), however, is a much stronger claim. According to (B), even if it would now be irrational to keep some

disposition, it *must* still be rational to act upon it, simply because it *once* brought benefits that were greater than its present costs. This claim, I believe, cannot be true. If it is irrational to keep this disposition, why must it be rational, if I do keep it, to act upon it?

If I have irrationally remained prudent, there is a different explanation of why it can be rational to act upon this disposition. Doing so will be better for me. The rationality of this act need not be defended by an appeal to the rationality of the disposition, or of my having kept the disposition, upon which I act. Things are quite different with ignoring your threat, in a way that I know will be disastrous for me. If this act is to be claimed to be rational, that can only be by an appeal to the rationality of the disposition on which I am acting. And if it is now irrational for me to keep this disposition, there seems no reason to conclude that, if I keep it, it must be rational for me to act upon it.

Suppose, next, that I could *not* lose my disposition, even if I tried. Gauthier might say that if, that is true, it is not irrational for me to keep this disposition. This is not something that I *do*. But it *would* be irrational for me to keep it, if I *could* lose it. This seems enough to undermine the claim that it must still be rational to act upon it.

[840] (C) is one interpretation of what Gauthier calls the 'weakest' version of his view, or what he calls his first level of commitment. On this view, he writes, one should act upon some disposition, even though one's actions are 'costly. . . only so long as one reasonably expects adherence to the disposition to be prospectively maximally beneficial' (Gauthier (1997) 39).

When Gauthier talks of 'adherence' to this disposition being beneficial, he must mean continuing to *have* this disposition. *Acting* on this disposition may be, as he agrees, costly. I shall also take 'adherence' to mean 'present adherence'. Though Gauthier might mean 'adherence now *and in the future*', that would make his claim less plausible. It would not cover cases where it would be advantageous first to acquire and then to lose some disposition. (Suppose that, while it was indeed better to acquire some permanent disposition than not to acquire it at all, it would have been expectably-best to acquire it simply for a time. Acquiring this permanent disposition was not then, as Gauthier requires, 'maximally beneficial'.)

[841] My drug-induced insanity, Gauthier claims, is 'the rational disposition in such situations, and the actions to which it gives rise are rational actions' (Gauthier (1997) 38). Gauthier means only that it is in my interests to have this disposition *now*. He is not here concerned with a choice between two permanent dispositions. If I had to choose my disposition, not just until the police arrive, but for the rest of my life, it would be better to remain sane and give the man my gold.

[842] Gauthier (1986) (*passim).*

[843] Gauthier might extend his claim about translucency. He might say that we could not have reason to believe that, if we broke our promises, we could keep this fact secret. But this reply would jettison what is novel in Gauthier's view, since it would revert to the ancient claim that honesty is always the best policy.

[844] There is one reading on which this claim must be true. It may be said that, if we are able to suspend our disposition, we were not *truly* trustworthy. But this reading is irrelevant since, for Gauthier's purposes, all that matters is whether we *appeared* trustworthy. It would be quite implausible to claim that, if we break some agreement, we cannot have earlier appeared to be trustworthy, even if, at the time, we sincerely intended to keep this agreement.

If this claim is to help Gauthier's case, he must make other revisions in his view. He writes: 'a disposition is rational if, among those humanly possible, having it will lead to one's life going as well as having any other' (Gauthier (1997) 31). This appeal to *human* possibility seems at odds with other parts of Gauthier's view. He claims elsewhere that we should not ask which dispositions are in general rational, since the answer may depend on a particular person's circumstances. Thus he writes, 'there need be no one disposition that, independently of an agent's circumstances, is sufficient to ensure that his life will go as well as possible, and thus I do not need to suppose that there need be a single supremely rational disposition' (Gauthier (1997) 31-2). A person's circumstances can surely include what is possible for this person.

This appeal to human possibility also raises a problem for Gauthier's argument. Trustworthiness is *not* the disposition that, among those *humanly* possible, is most advantageous. It would be more advantageous to appear to be trustworthy but to be really prudent; and that is surely possible for some human beings. If Gauthier appeals to what is humanly possible, he would have to judge trustworthiness to be an irrational disposition, even when it is had by people for whom, since they could not deceive others, it is the most advantageous possible disposition.

[845] At one point, Gauthier may make this move. While honesty is the best policy, Hume writes, there may be some exceptions. According to Hume's 'sensible knave', he is wisest 'who observes the general rule, and takes advantage of all the exceptions.' Gauthier replies that, to be rational, we must be disposed to keep our promises, since this disposition will be best for us. He then writes, 'such a person is not able, given her disposition, to take advantage of the "exceptions'"; she rightly judges such conduct irrational.' (Gauthier (1986) 182.)

[846] See pages 000.

[847] In the doctrine that 'ought' implies 'can', the sense of 'can' is compatible with determinism. If that were denied, and we assumed determinism, we would have to claim that *every* act is rational.

[848] It would of course be better if I merely appeared to be insane. But we can suppose that this is not possible, since if I had not taken the drug, the robber would know this. (Perhaps one of the drug's effects is a characteristic look in the eyes; or perhaps I can convince the robber only if he sees me drink this drug.) Being actually in this state is then the disposition that is best for me.

[849] Gauthier (1997) 37.

[850] Provided, of course, that these bad effects do not outweigh the good effects of my disposition. Gauthier need not claim that, if I killed myself or my children, that would be rational.

[851] It may be said that, in one respect, Gauthier's view is less extreme than Hume's. Even if my act has bad effects, these must be outweighed by the good effects of having my disposition. But we can remember here that, on Gauthier's main view, I maximize my utility if I fulfil my present considered preferences, and these need not coincide with my interests. As on Hume's view, these preferences could be as crazy as we can imagine. The difference between these views is that, on Hume's view, for my act to be rational, I must at least be trying to fulfil my aims, while on Gauthier's view, my acts need only be the side-effects of a state the having of which will achieve these aims.

[852] 'Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices' (Gauthier (1986) 187).

[853] It may seem that, if that is true, breaking our promises cannot be better for us. But this may not be so. The bad effects come, not from our breaking of these promises, but from the fact that we are both translucent and disposed to break our promises whenever this will be better for us.

[854] It is worth explaining why. In our assessment of the good or bad effects of our dispositions, we include the acts to which these dispositions would or might lead. If it is best for us to have some disposition, even though this will lead to acts which are bad for us, those effects must be outweighed. Since the assessment of our dispositions includes the assessment of our acts, but goes beyond it, this is the assessment that tells us what on balance will be best for us.

[855] Gauthier (1986) 170.

[856] It may be questioned whether G tells us, if we can, to acquire these dispositions. That does not follow from the fact that, if we do, that will be better for us. If G does not tell us to act in this way, that would be an objection to G, and would again undermine Gauthier's argument. But Gauthier might claim that, in trying to acquire these dispositions, we would be acting on an advantageous, or maximizing, meta-disposition.

[857] He would admit that, in practice, few of us are always rational. But he might claim that, in assessing the plausibility of these theories, we should consider what would happen if we always did what they told us to do. He might then claim that, if we fully followed S, we would always maximize at the level of our acts.

[858] It may be objected that, if we cannot always do what E claims to be rational, E cannot claim that we ought to do so. 'Ought' implies 'can'. But this confuses two questions. When I say that we cannot always do what E claims to be rational, I mean that this is not causally possible. This is the kind of possibility that is relevant when we are comparing the effects of our having different dispositions. The sense of 'can' that is implied by 'ought' does not, as Gauthier agrees, require such causal possibility, since this other sense of 'can' is compatible with determinism.

[859] It may seem that, if we cannot always do what E tells us to do, there is no way of predicting when we shall follow S. That is not so. Suppose that we are now always disposed to do what we believe to be rational. If we know that we can acquire maximizing dispositions, we shall then do so, even though we know that this will cause us later to act irrationally. Acquiring these dispositions is, according to E, the rational thing to do. It is only *after* acquiring these dispositions that we shall start acting in ways that E claims to be irrational.

[860] Gauthier (1984) and (1985).

[861] Gauthier (1985) 159-61.

[862] McLennen (1988).

[863] Such a claim is fairly plausible in the case of trustworthiness, the disposition that is Gauthier's chief concern. If we could not conceal our intentions, as he assumes, it might be better for us if we intended to keep our promises, even when this way of acting would be worse for us. Unless we have this intention, others might exclude us from advantageous agreements. And for us to be able to form this intention, we might have to believe that it is rational to keep such promises.

[864] In a letter to me.

[865] See Gauthier (1986) (p. 182) and Gauthier (1997) 31). (But see also Gauthier (1986), pp. 170 and 158.)

[866] Gauthier (1997) 36.

[867] At one point, Gauthier comes close to accepting (D). He cites my book's version of (D)---there called '(G2)'---and writes, 'to this extent I accept. . . (G2)' Gauthier (1997) 40.

[868] It may seem that, in making these remarks, I have presupposed a naively realistic view. Gauthier might say that a normative theory could not be *true*. But this would not rescue Gauthier's argument. Even on a noncognitivist view, we must give some content to the notion of a normative belief. We must be able to claim that an act *is* rational, and be able to assert or deny different theories. My remarks could be restated in these terms.

[869] Lewis (1985).

[870] Gauthier (1985) 159-61.

[871] Gauthier (1997) 30.

[872] Gauthier (1997) 36.

[873] Gauthier (1997) 38.

[874] Gauthier (1986), 17.

[875] [Acknowledgments to Otsuka, Persson, and Cullity.]

[876] This may not be the best description of what makes these acts permissible. For another account, see Kamm (2007).

[877] Though Kant assumes, in the *Groundwork*, that there are no such objective ends-to-be-produced, that does not explain his claims in passage (A) quoted above. Kant here writes that all imperatives either *represent* some act as a necessary means to some subjective end, or represent some act as necessary in itself. This claim is about the content of possible imperatives. (A) cannot be read as claiming that, though some imperatives represent some act as a necessary means to some objective end-to-be-produced, no such imperatives are valid, because there are no such ends. So it seems that, in this passage and in his later arguments, Kant overlooks this kind of imperative. Given Kant's love of taxonomies which are exhaustive in the sense of covering every possibility, Kant's overlooking of these imperatives is a mystery. I suggest one possible explanation in note 000 below.

[878] The phrase 'for its own sake' can be used, we should note, in a slightly different sense. Our acts have moral worth, Kant claims, only when we act '*from* duty', or for the sake of duty.

When we act on some deontological principle, such as a requirement not to lie, we may both be acting in some way for its own sake rather than as a means of producing some effect, and be doing our duty for its own sake. But we might also act from duty on some purely teleological principle, such as one that requires us to do what would benefit others. Though we would then *do our duty* for its own sake, our duty would be to *act* in this way, *not* for its own sake, but as a means of benefiting others. (Reference to Korsgaard.)

[879] We can now suggest one way in which Kant may have overlooked the possibility of categorical teleological imperatives. Kant may have had in mind three of the distinctions that I have just drawn. When considering imperatives that require us to act in some way, Kant may have seen that any such imperative must either

motivate us only with the help of some desire,	or	motivate us all by itself,

and must either

apply to us only if we have some desire,	or	apply to us whatever our desires,

and must either

tell us to act in some way as a means of achieving some end,	or	tell us to act in some way for its own sake only.

If Kant did not distinguish clearly between these distinctions---as is suggested by the fact that he uses 'formal' and 'material' to express all three distinctions---this may explain why he misdescribes the third distinction, claiming that all imperatives tell us to act in some way either for its own sake only, or as a means of achieving some *desired* end. The other two exhaustive distinctions both refer, in their left-hand side, to our desires. By adding this reference to desires, Kant may have drawn the third distinction in a way that is *not* exhaustive, since it overlooks those imperatives that tell us to act in some way as a means of achieving some categorically required end.

[880] Reference to Allison and others.

[881] Guyer (1992) 325-6.

[882] Potter (1998) 40.

[883] Schneewind (1998) 318.

[884] Gregor (1963) 78-9.

[885] Thus, after writing that only 'lawgiving form. . . can constitute a determining ground of the will', and commenting on that claim, Kant concludes that 'the fundamental law' is 'So act that the maxim of your will could always hold at the same time as a principle in a giving of universal law'CPR, 29-30.

[886] Kant's 'refutation' contains another argument. Kant writes:

> Because the impulse that the representation of an object possible through our powers is to exert on the will of the subject in accordance with his natural constitution belongs to the nature of the subject---whether to his sensibility (inclination and taste) or to his understanding and reason, which by the special constitution of their nature employ themselves with delight upon an object----it would, strictly speaking, be nature that gives the law; and this, as a law of nature, must not only be cognized and proved by experience---and is therefore in itself contingent and hence unfit for an apodictic practical rule, such as moral rules must be. . (G 444)

Kant again concedes here that, when some principle gives us some 'object', or end, we might be moved to act upon this principle, not by our inclinations, but by our reason. When applied to such principles, Kant's argument is this:

> (1) If we believed that there was some end which we were required to try to achieve, and we were moved to act on this belief by our reason, this motivation would depend on our natural constitution. It would be a natural feature of us that we were, in this way, rational, being able to be moved by our belief in this requirement.
>
> (2) Since our being moved by this belief would depend upon our nature, it would really be nature, not reason, which gave us this requirement.
>
> (3) Since natural laws are contingent, but moral requirements must be necessary, this requirement could not be a moral law.

Though this argument raises deep and difficult questions, it cannot be sound. We might similarly claim that, since our ability to reason logically depends on our nature, logical laws must be natural and contingent. Kant would rightly reject that claim. And to protect his Formal Principle from this argument, Kant must claim that our ability to act on his principle does *not* depend on our natural constitution. Kant might say that we act on his principle not as natural but as noumenal beings. But even on that assumption, this argument could not show that there are no true substantive principles. As before, if there are such principles, we could moved to act upon them in whatever way in which we could act upon Kant's Principle.

[887] For an excellent discussion of both these arguments, see Kerstein (2002) Chapter 7. There is much else in Kerstein's book which goes beyond, and may partly correct, my brief claims in this Appendix.

[888] Irwin (1996) 80.

[889] See, for example, 'What is Orientation in Thinking', VIII, 145, 303-4, and G 448.

[890] Hill (1992) 88.

[891] Korsgaard (1996A), 22.

[892] Korsgaard (1997).

[893] Kant writes, 'from the problematic and pragmatic', which are his names for the two forms of hypothetical imperative.

[894] He writes: 'even if there have never been actions arising from such pure sources, what is at issue here is not whether this or that happened; that, instead, reason by itself and independently of all appearances commands what ought to happen; that, accordingly, actions of which the world has perhaps so far given no example, and whose very practicability might be very much doubted by one who bases everything on experience, are still inflexibly commanded by pure reason.' G 407-8.

[895] Of Kant's grounds for making this assumption, another may be his view that, for our acts to have moral worth, 'it is essential. . . that the moral law determine the will directly'(CPR 71). If no principle could directly motivate us, none of our acts, on this view, could have any moral worth. Suppose, however, that our acceptance of the moral law motivates us, not directly and all by itself, but only with the help of a standing desire to do our duty. It would be implausible to claim that, when we act on this desire, doing our duty because it is our duty, our acts have no moral worth.

[896] Both this and the previous quotation apply specifically to the Formula of Universal Law. This remark refers to 'ask yourself whether, if the action you propose were to take place by a law of nature of which you were yourself a part, you could indeed regard it as possible through your will. . . . if you belonged to such an order of things, would you be in it with the assent of your will?

[897] Reference to Darwall.

[898] Rawls (1971) 30-3.

[899] Rawls (1999) 524-5.

[900] *Religion within the Limits of Reason Alone*, translated by T. Greene and H. Hudson, Harper 1960, 25.