ON WHAT MATTERS

DEREK PARFIT

Draft of 28 April 2008

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 5, 6, 11, and 13 to 16. If you are mainly interested in reasons, rationality, and meta-ethics, you might read only Part One and Appendix A.

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 discouraged 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 claims 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. After reading a collection of Sidgwick's 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. [7]

Here are two other 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. [8]
>
> 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? [9]

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. [10] 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. [11]

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. [12]

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 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. [13]

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. [14]

Another passage ends:

> . . . even a man who said 'Evil be though 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. [15]

Sidgwick warned his friends that, because his book attempts to achieve 'precision of thought', it 'cannot fail to be somewhat dry and repellent'. [16] But this precision is often finely expressed. Discussing friendship, for example, Sidgwick refers to

> 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. [17]

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. [18]

> Thus the Utilitarian conclusion, carefully stated, would seem to be this: that the opinion 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. [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]

> . . . 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.[21]

> [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. [22]

Sidgwick's language 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 they do get mildly more adventurous. . . make fairly uncritical use of a notion of purity.' [23] 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. [24]

When people find Sidgwick dull, they are responding, I believe, 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. [25]

Sidgwick is very good at seeing the force of objections to his views. After attending a philosophical debate, 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.

This quality I find very endearing. For example, Sidgwick writes:

> I am trying to finish my review of Martineau's [book]. 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.

It is these virtues that can make Sidgwick hard to read. The problem is that, as C. D. Broad explains, Sidgwick

> incessantly refines, qualifies, raises objections, answers them, and then finds further objections to the answer. Each of

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

Our first reading of the *Methods* is, in a way, the worst, since there is little that is striking or inspiring, and we can easily get lost in the way that Broad describes. 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.

Sidgwick also writes:

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

This is another overstatement. 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'. [26] In the *Methods*, as Broad claims, 'almost all the main problems of ethics are discussed with extreme acuteness'. [27] 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. [28] These facts make Sidgwick's *Methods* the book on ethics that it would most advance the subject for everyone to read, remember, and be able to assume that others have read.

My own debts to Sidgwick are easy to describe. Of my reasons for becoming a graduate student in philosophy in the late 1960s, 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 depressing 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 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 few 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 hope I may persuade some other people to overcome their antipathy to Kant, and to learn from him as I have done.

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. That 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 show other people some atrocious paragraph, and say 'How can it be worth reading anyone who writes like this?' The answer could always be 'What about Kant?'

There are deeper problems. When I become obsessed with Kant, I tried to restate more clearly some of Kant's main claims and arguments, and I found this task very frustrating. I couldn't fit all of Kant's claims together in a coherent view, and many of Kant's arguments seemed to be obviously 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. [29]

It would also have helped me to know that Kant did not have a single, clear, coherent theory. When we ask whether Kant accepts or rejects some claim, the answer is often 'Both'. As Kemp Smith remarks, 'citation of single passages is quite inconclusive'. [30] For example, though Kant writes that 'a human being's duty at each instant is to do all the good in his power', [31] he is not really 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'. [32] 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.' [33] (To avoid provoking Hegelians, we should perhaps say 'one of the least exact'.)

'Consistency', Kant writes, 'is the greatest obligation of a philosopher'. [34] That, I believe, is not true. Clarity is as important. And Kant's greatness chiefly consists in his having many philosophically deep 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 that there is 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, making Kant's view no better than the propaganda of the so-called 'People's Democracies'of the Soviet bloc, in which voting was compulsory, and there was only 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. . . [35]

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. [36]

Some of Kant's weak arguments, moreover, have great charm. When condemning suicide, Kant also 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. [37]

It is the word 'supposedly' that is so endearing here. There is 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. [38]

This conclusion arrives too late.

For another example, consider Kant's claim that, in lying 'even to achieve a 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'. [39] We should ignore such outbursts. On the very next page Kant himself suggests that, if we are asked by an author whether we like his work, we may be permitted to say what 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. . .' [40] Kant seldom uses words like 'most', 'many', 'several', or 'some', preferring to write only 'all' or 'none'. Kant says that he uses 'good' 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* and *forbidden*.

Temperamentally, I am an extremist too, who has to struggle to be more like Sidgwick. But unlike some Kant's other readers, I never hated Kant, and I have now made my peace with him. At Oxford we 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 only be about how much of what Kant wrote is Alpha, and how much is Gamma. And, when we have understood what is Alpha, what is Gamma doesn't matter. [41]

I still believe that, in some ways, 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 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 his arguments often fail, he also gives us some profound truths. Like Sidgwick, I often find him 'quite a revelation.' [42] Kant's books are extraordinarily fruitful and 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.' [43]

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 Part Two, though the book's main arguments are in Parts One and Three. But I discuss only what Kant's formulas imply when these formulas are understood in their most straightforward way. When I ask, for example, whether it is wrong to treat people *merely as a means*, I do not consider what Kant himself meant by this phrase. Except in a few sections, which are mostly in Part Two or some Appendices, 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 became a philosopher, as I have said, 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 revealing 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 irreducibly normative truths.[44] During several visits to Harvard, 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 . . .

ON WHAT MATTERS

SUMMARY

PART ONE

CHAPTER 1 REASONS

1 Normative Reasons

We are the animals that can 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*, 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.

CHAPTER 2 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, aims, or choices. Some of these theories appeal to our actual desires or aims; others appeal to the desires or aims that we would now have, or the choices that we would now make, if we had carefully considered the relevant facts. According to *value-based, 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. Theories of these two kinds often deeply disagree. We ought, I shall argue, to accept some value-based objective theory.

4 Responding to Reasons

Our responses to reasons for acting are voluntary, in the sense that we could have chosen to act differently. But, 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.

5 Object-Given Reasons

The same facts can give us reasons both to have some desire, and to try to fulfil this desire by acting in some way. 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 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 merely 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 they are responses to what is good or bad in these conscious states.

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 defensibly made by those who accept subjective theories about reasons.

6 Subject-Given Reasons

Subjective theories about reasons can take several forms. Some theories appeal to the desires or aims that we would have, or the choices that we would make, after fully informed and *procedurally* rational deliberation.

CHAPTER 3 AGAINST SUBJECTIVISM

7 Why People Accept Subjective Theories

Since so many people believe that *all* reasons are provided by facts about what would fulfil our present desires, aims, or choices how can it be true that 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 may seem to give us reasons.

8 Analytical Subjectivism

Some claims seem important, but are merely *concealed tautologies*, which everyone could accept whatever else they believe. Some people use the words 'reason', 'should', and 'ought' in *subjectivist* senses. When subjectivists use these senses, they do not make substantive claims. Such analytical subjective theories I discuss only in Appendix A.

9 Why We Ought to Reject Subjective Theories

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. 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.

According to 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. Some of our desires can 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. Such desires cannot be defensibly claimed to give us any reasons. So subjectivists cannot defensibly claim that we have reasons to have any desire or aim. Nor can these people even defensibly claim that we have reasons to act in any way.

10 Fully Informed Desires

Some subjective theorists claim that, when we are making important decisions, we should often to try to learn more about the different possible outcomes of our acts, so that we shall come to have better informed desires, and can then try to fulfil these desires. But these people cannot coherently make such claims.

11 Reasons, Motives, and Well-Being

If we accept some subjective theory about reasons, we must deny that events can be good or bad for particular people, or impersonally good or bad, in the reason-implying senses. If these theories were true, nothing could be good or bad in these ways. When some writers claim that some life would be best for someone, they mean that this is the life that, after fully informed and 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. Other accounts fail in other ways.

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

CHAPTER 4 RATIONALITY

12 Practical and Epistemic Rationality

We are rational insofar as we respond well, or correctly, to reasons or apparent reasons. We have some *apparent* reason when we have beliefs about the relevant facts whose truth would give us some reason. Our desires are rational when, if our beliefs were true, what we want would be in some way good, or worth achieving. Our acts are rational when, if our beliefs were true, we would be doing what we had sufficient reasons to do. In most cases, it is irrelevant whether our beliefs are rational.

13 Beliefs about Reasons

Practical irrationality does not, as some people claim, merely involve inconsistency between our desires or acts and our normative beliefs. It can be irrational to want, and to do, what we believe that we have decisive reasons to want and to do.

14 Other Views about Rationality

The rationality of our desires does not depend, as many people claim, on how we came to have these desires, or on whether they are inconsistent, or on whether our having these desires has good effects.

CHAPTER 5 MORALITY

15 Sidgwick's Dualism

We can assess the strength of all our reasons, Sidgwick claims, 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, and revise his conclusion. We ought to assess the strength of all our reasons from our actual, personal point of view. 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. These two kinds of reason *are* comparable, but only very imprecisely. As *wide value-based objective* theories claim, 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. If we know the facts that give us such reasons, either act would be rational.

16 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. Though facts about reasons are, in this way, more fundamental, the rest of this book is about morality. In discussing morality, we shall be discussing some of the reasons that most need discussing, because they raise the most difficult questions. And,jbefore we can decide whether we might have sufficient or decisive reasons to act wrongly, we must know more about which acts are wrong, and what makes them wrong.

CHAPTER 6 MORAL CONCEPTS

17 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. We ought to distinguish, for example, between some act's being wrong in the *fact-relative*, *evidence-relative*, *belief-relative*, and *moral-belief-relative* senses.

18 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 morality, and about practical and epistemic reasons. Rather than proposing a new moral theory, I shall try to develop existing theories of three kinds: Kantian, contractualist, and consequentialist.

PART TWO

CHAPTER 7 POSSIBLE CONSENT

19 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.

20 The Consent Principle

Kant's claim 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 not given 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 claim about consent 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.

21 Reasons to Give Consent

Whether we could achieve Kant's ideal depends on which are the acts to which, if they knew the relevant facts, people could rationally consent, because they would have sufficient reasons to consent. If we ought to accept either some subjective theory about reasons, 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 we ought to accept 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 can argue that, in all such cases, it would be wrong to act in any way to which anyone could not rationally consent.

22 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 some act, but the 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 we must choose between acts that 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. We have another reason to discuss the Consent Principle: it is worth asking whether we could achieve Kant's

ideal.

23 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 refuse consent. That is clearly false. It might be argued 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.

24 Deontic Beliefs

To explain why the Consent Principle does not mistakenly require certain wrong acts, we must appeal to the claim 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.

25 Extreme Demands

The Consent Principle can require us to bear great burdens, when that is our only way to save others from much greater burdens. If this requirement is too demanding, we could revise this principle. But we might still be able to achieve Kant's ideal.

CHAPTER 8 MERELY AS A MEANS

26 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 claim, 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.

Consider some Egoist, whose only aim is to benefit himself. When this man keeps his promises, pays his debts, and saves some drowning child in the hope of getting some reward, he may be treating some other people merely as a means. But

these acts would not be wrong. Kant's claim could be qualified, so that it would not mistakenly imply that such acts are wrong. It is wrong, we might claim, to treat anyone merely as a means, or to come close to doing that, if our act is also likely to harm this person.

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

27 *As* a Means and *Merely* as a Means

It is widely assumed 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.

28 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 lesser harms on people as a means of achieving these 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 9 RESPECT AND VALUE

29 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. But this claim does not help us to decide, in difficult cases, which acts would be wrong.

30 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 these acts or events 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.

31 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.

32 The Right and the Good

The 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. This part of Kant's view may seem to make what Kant calls the 'fundamental error' of the ancient Greeks. But that is not so.

33 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. It is by strictly following certain moral rules, Kant claims, that everyone could do most to produce this world. When Kant, Hume, and others make such claims, they fail to draw some distinctions that we need to draw.

CHAPTER 10 FREE WILL AND DESERT

34 The Freedom that Morality Requires

If our acts were merely events in time, Kant argues, 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 its being true that our acts are merely events in time.

35 Deserving 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. This argument also fails. 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

CHAPTER 11 UNIVERSAL LAWS

36 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 is 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.

37 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, since Kant's formulas would then achieve nothing.

38 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 other people act on various moral maxims. Since our imagined Egoist always acts on a maxim that he could not rationally will to be universal, Kant's formulas imply that all of this man's acts are wrong. This man acts wrongly even when, for self-interested reasons, takes some medicine, pays his debts, and saves some drowning child in the hope of getting some reward. These implications are clearly false. When this man 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, whenever 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. 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 12 WHAT IF EVERYONE DID THAT?

39 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 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 say or think '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 involves the acts with which we are 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 together do, 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*.

40 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 'Have no children, so as to devote my life to philosophy' and 'Consume food without producing any.' 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 can be partly met by pointing out that most people's maxims are implicitly conditional, since they do not apply to cases in which the acts that they describe would have various bad effects.

41 The Ideal World Objection

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, not to use violence even in self-defence, and to act in ways that mistakenly ignore what other people will in fact do. This objection can be answered. Kant's formula does not require such acts.

There is, however, 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* and *contractualist* moral theories. To answer this objection, we should revise Kant's formula in another way. On this revised formula, it is wrong to act on some maxim 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.

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

CHAPTER 13 IMPARTIALITY

42 The Golden Rule

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

43 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.

44 The Non-Reversibility Objection

Many wrong acts benefit the agent but impose much greater burdens on others. The Golden Rule condemns such acts, because we could not rationally choose that 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, some of us could rationally will it to be true both that everyone acts like us, and that everyone believes such acts to be morally permitted. So Kant's formulas mistakenly permit these wrong acts.

This objection applies to many actual cases. One example involves the men who benefit themselves by denying women various opportunities, and giving less weight to their 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 some people who are rich or powerful exploit and oppress some people who are poor or weak.

45 A Kantian Solution

To avoid this and some of our other objections, we should again revise Kant's formulas. The Moral Belief Formula 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 14 CONTRACTUALISM

46 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

> *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.

47 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.

48 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 the relevant facts, and she supposed that she 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.

49 The Deontic Beliefs Restriction

According to Scanlon's partly similar formula, everyone ought to follow 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 claims that his formula gives an account of wrongness itself, or of *what it is* for acts to be wrong. But contractualist formulas are better claimed to describe one of the facts that can *make* acts wrong.

When we apply any contractualist formula, contractualists must claim, we cannot appeal to our intuitive beliefs about which acts are wrong. If we could appeal to such *deontic* beliefs, these formulas would be worthless, and show nothing. To defend this feature of their view, some contractualists claim that we ought to ignore such intuitive deontic beliefs. 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 we ought to accept this formula.

CHAPTER 15 CONSEQUENTIALISM

50 What Would Make Things Go Best

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. According to *Value-based Consequentialists,* the rightness of our acts depends only on facts about how it would be best for things to go. *Direct* Consequentialists apply this test to everything. When these people apply this test to acts, they are *Act Consequentialists*. *Indirect* Consequentialists apply this test 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.

51 Consequentialist Maxims

According to *Maxim Consequentialism*, everyone ought to act on the maxims whose being acted on by everyone would make things go best. Kant's Law of Nature Formula permits some people to be Maxim Consequentialists.

52 to 56 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 what they would have sufficient reasons to choose.
>
> There are some principles whose universal acceptance would make things go best in the impartial-reason-implying sense.
>
> These are the principles whose universal acceptance everyone would have the strongest impartial reasons to choose.
>
> No one's impartial reasons to choose these principles would be outweighed by any set of relevant conflicting reasons.

Therefore

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

Therefore

> It is only these optimific principles whose universal acceptance everyone would have sufficient reasons to choose, and could rationally 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 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 has 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 consequentialist conclusions.

CHAPTER 16 CONCLUSIONS

57 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. 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.

58 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, or gathered. If this theory succeeds, it would describe what these other properties have in common.

For the Triple Theory to succeed, it must be both in itself plausible and have acceptable implications. This theory has many plausible implications. 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 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.

COMMENTARIES

APPENDICES

APPENDIX G KANT'S MOTIVATIONAL ARGUMENT

PART ONE

CHAPTER 1 REASONS

(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 Normative Reasons

We are the animals that can understand and respond to reasons. This ability has given us great knowledge, and power to control the future of life on Earth. We may even be the only rational beings in the Universe. [45]

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 might 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 belief or desire, or our acting in some way. But 'counting in favour of' means roughly 'giving a reason for'. Like some other fundamental concepts, such as those of time, space, possibility, and reality, the concept of a reason is indefinable in the sense that it cannot be helpfully explained merely by using words. [46] 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 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. 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 which gives us this reason. But these claims do not conflict with mine, since these are merely different ways of saying the same things. My doctor might say, 'No, you don't have any reason not to eat any other kind of food, but you will have such a reason after I've told you that eating walnuts would kill you'. It is simpler to say that I already have this reason.

When we must choose between different possible acts, our reasons may conflict, and they can differ in what we can call their force, strength, or weight. If I enjoy walnuts, that gives me a reason to eat them; but, if they would kill me, that 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. Rather than saving one person from ten hours of pain, for example, we might have a stronger set of reasons to act in a way that would both save this person from nine hours of pain, and save each of ten other people from one hour of pain. We would then have *more reason* to act in this second 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. [47] When 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 facts, we are not, in the sense I intend, *responding* to these reasons. This is like the sense in which, if we ignore someone's cry for help, we are not responding to this cry.

There is often nothing that we have decisive reasons to do, or most reason to do, because we have *sufficient* reasons, or *enough* reason, to act in any of two or more ways. Our reasons to do something are sufficient when these reasons are not weaker than, or outweighed by, our reasons to do anything else. We might have sufficient reasons, for example, to eat either a plum or a peach, to give some money either to Oxfam or to some other well-run aid agency, or to choose either law or medicine as a career. When neither of two conflicting reasons would be stronger, that need not be because these reasons are precisely equally strong. Though there are truths about the relative strength of different reasons, these truths are often very imprecise. [48]

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. [49] Even if we never use the phrases 'decisive reason' or 'most reason', most of us often use 'should' and 'ought' in these senses. There is a similar sense of 'must'. These words imply reasons of different strengths. For example, 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 rail. Though the word 'should' is used much more often, I shall mostly use the stronger and less ambiguous 'ought'.

As well as asking what we ought to do in the reason-implying sense, we can ask what we ought *rationally* to do. When we call some act 'rational' we express the kind of praise or approval that we can also express with words like 'sensible', 'reasonable', 'intelligent', and 'smart'. We use 'irrational' to express the kind of criticism that we also express with words like 'senseless', 'stupid', 'idiotic', and 'crazy'. Acts that are open to weaker criticism we can call '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 different 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 implicit assumptions, such as the assumption that we would not harm ourselves or others by eating a walnut, or pushing open some swinging door. If we have certain beliefs about the relevant facts, and what we believe would, if it were true, give us some 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, [50]
>
> 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 decisive reasons *not* to act in this way,

and

> *irrational* if these reasons would be both clear and strongly decisive.

When we know all 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 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 beliefs, 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 our false beliefs can give us reasons, we would have to claim that these reasons have no

normative force, in the sense that they do not count in favour of some act. We would have to ignore such reasons when we are trying to decide what someone has most reason to do, and what this person ought in the reason-implying sense to do. It is better to say that our false beliefs can give us what merely *appear* to be reasons. In the case of the angry snake, since you believe that running away would save your life, you have an *apparent reason* to run away. On the view stated above, we ought rationally to act in some way if we have decisive apparent reasons to act in this way. It is irrelevant whether, because our beliefs are true, these apparent reasons are real.

Similar claims apply to what we actually do. 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 acts are most irrational when these beliefs are conscious. If these reasons would not be clear, or strongly decisive, our act may be only less than fully rational. [51] Similar claims, I shall argue, apply to our desires and aims. *We* are rational if, in our desires, aims, and acts, we respond to reasons or apparent reasons. To be fully rational, we may also need to respond to certain rational requirements, by avoiding certain kinds of inconsistency between our acts, intentions, and some other mental states. In this book I shall say little about such requirements.

Some people claim that, to be rational, we don't need to respond to reasons or apparent reasons, since it is enough to respond to these rational requirements. According to another widely held view, the rationality of our desires and acts depends on the rationality of our beliefs. I shall question these views in Chapter 4.

We can next explain why, though it is facts that give us reasons, what is rational depends on our beliefs. When we are trying to decide what someone ought to do, so that we can give this person advice, it is the facts that matter. 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 such people may not deserve any criticism,

since they may have false beliefs whose truth would have given them sufficient reasons to act as they did. In most cases, that is enough to make their act rational. If you ran away from the angry snake, believing falsely that this act would save your life, your fatal act wouldn't be foolish, stupid, or crazy. You would merely be very unlucky.

On the view stated above, we ought rationally to act in some way if we have beliefs about the relevant facts whose truth would give us decisive reasons to act in this way. In many cases, we do not have such beliefs. When we do not know all of the relevant facts, it may seem that we ought rationally to try to do what we have decisive reasons to do. In some cases, however, that is so. Such acts, for example, may be too risky. It is of great importance what we ought rationally to do when we do not know all of the relevant facts, and how we can best respond to risks and to uncertainty. These questions have been well discussed by many philosophers, economists, and decision theorists. Compared with these questions about what we ought rationally to do, questions about reasons are more fundamental. These are the questions about which different people, and different theories, most deeply disagree. Since I shall mainly be discussing these deeper disagreements, I shall mostly consider cases in which we know all of the relevant facts. In such cases, what we ought rationally to do is the same as what we have decisive reasons, or most reason, to do.

These claims have been about about *normative* or *justifying* reasons. When we have such a reason or apparent reason, and we act *for this reason*, this becomes our *motivating* or *explanatory* 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. [52] 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 which 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. 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 ought to do in the decisive-reason-implying sense, I shall mostly discuss what we have decisive reasons, or most reason, to do.

2 Reason-involving Goodness

We can next consider the concepts *good* and *bad*. When we call something

> *good*, in what we can call the *reason-implying* sense, we mean that there are facts about this thing's nature, or properties, that would in some situations give us or others certain kinds of reason 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', 'best', 'bad', 'worse', and 'worst'. [53]

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

> Ice formed on the butler's upper slopes, [54]

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

> A MAN A PLAN A CANAL: PANAMA,

these senses of 'good', 'bad', and 'best' are not plausibly regarded as reason-implying. 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, Thomas 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 further reason. [55]

This view needs, I think, one small revision. If some medicine is the best, this fact could be truly claimed to give us a reason to take this medicine. But this reason would be *derivative*, since its

normative force would derive entirely from the facts that made this medicine the best. That is why it would be odd to claim that we had *three* reasons to take this 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, they are not worth mentioning in such a claim. [56]

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.

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, and *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 *substantive good* theories, our well-being may also partly consist in some other activities or states, such as loving and being loved, moral goodness, and some 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 is 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 next define another sense in which events can be good or bad for us. When we call some possible event

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

It would be in this sense good for us if we were happy, and bad for us if we were in pain, or if we suffered in other ways. The phrases 'good for us' and 'bad for us' are often used more narrowly, to refer to things that have good or bad effects on our health. Pain and suffering may not be in these ways bad for us. But it is always in itself bad to be in pain. Pain and suffering are bad for us in the sense that these are conscious states that we have self-interested reasons to want not to be in.

We can have strong reasons to care about the well-being of certain other people, such as our close relatives and those we love. Like self-interested reasons, these altruistic reasons are both *personal* and *partial*, since they are reasons to be specially concerned about the well-being of those people who are *related to us* in certain ways. We also have some reasons, I believe, to care about everyone's well-being. Such reasons are *impartial* in the sense that they are reasons to care about anyone's well-being, whatever that person's relation to us.

These reasons are also impartial in the different sense that these are the only reasons that we would have 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 we might have from a financial, aesthetic, or other such point of view. When we think about certain possible events, our *actual* point of view is impartial. That is true when we are considering possible events that would involve or affect 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. Suppose that, after some shipwreck, some rescuers could save either me or many other people who are all strangers to me. I would have strong self-interested reasons to want these rescuers to save me rather than these many strangers. But I would know that, if I were in the position of an impartial observer, I would have more reason to want the rescuers to save many people rather than saving only one.

From an impartial point of view we would all have 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 this point of view. Some people might instead believe that, from an impartial point of view, we would all have reasons to care more about the well-being of certain people, such as those people who are morally best, or have the greatest abilities. Note next that, even 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. [57]

We can now describe another kind of goodness. When we claim that one of two possible events would be

> *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.

It would be in this sense better if some plague or earthquake killed fewer people, or if anyone ceased to be in pain. This kind of goodness I shall call *impersonal*. But this word may be misleading. Such goodness is impersonal only in the sense that,

in calling some event 'good', we don't mean that this event is good *for* some person. Many events are impersonally good because they are good *for* one or more people. And since everyone has reasons to want such events to occur, such impersonal goodness involves *omnipersonal* reasons.

If some 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' and 'good for' in these senses, these are merely briefer ways of claiming that there are such other, reason-giving facts.

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 THEORIES

3 Two Kinds of Theory

The word 'desire' often refers to our sensual desires or appetites, or to our being attracted to something, or 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 to 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 that we do, 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.

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

Our desires are *teleological* or *telic* when we want some event as an *end*, or for its own sake. Our desires are *instrumental* when we want some event as a *means*, because this event would or might cause some other event that we want. We want some acts or other events both as an end and as a means to some other end. Two such acts are taking enjoyable and health-preserving exercise, and eating delicious and nourishing food. When we decide to try to fulfil some 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 some telic desire. For example, I might want medical treatment, 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 chain.

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

As well as wanting such other things, some people do not even want pleasure as an end. Suppose that we know some relentlessly ambitious politician, whom we find basking in the sun, sipping champagne. When we ask this man what he is doing, he replies 'Enjoying myself'. Given our knowledge of this man's character, this reply is baffling. This man never does anything for enjoyment. He then explains that his doctor warned that, unless he allows himself some passive 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 because they would have effects that he wants.

There are two main kinds of view about what we can 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 what might achieve these aims. These reasons are given by facts about the *objects* of these desires or aims, or *what* we want or are trying 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*.

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. Most of these outcomes or acts would be good or bad for particular people, or impersonally good or bad. So we can also call such reasons *value-based*. But, as I have said, such reasons derive their force, not from the goodness or badness of these acts or outcomes, but from the facts that would make them good or bad.

According to another group of theories, our reasons for acting depend on facts about what would fulfil or achieve our present desires or aims, or on facts about what we choose. Some of these theories appeal to our actual present desires, aims, or

choices. Other theories 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 reasons depend on these facts about *us,* and our desires, aims, and choices, we can call them *subject-given*. To save words, I shall sometimes discuss only *desire-based* subject-given reasons, but most of my claims would also apply to *aim-based* or *choice-based* reasons. If we believe that all practical reasons are of this kind, we are *Subjectivists about Reasons*.

These two kinds of theory are very different. According to Objectivists, practical reasons all derive their force from the facts that make certain things relevantly good, by giving us reasons to want these things, and to try to achieve them. According to Subjectivists, we have no such object-given reasons. Some Subjectivists even claim that it is *we* who 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. [58]

These *objective* and *subjective* theories 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 these kinds of theory agree that we have reasons to try to fulfil many of our present desires, they may disagree about how strong these reasons are. On many subjective theories, the strength of these reasons depends on the strength of these desires. On objective theories, the strength of these reasons depends instead on how good, or worth achieving, the fulfilment of these desires would be. Most of us often have stronger desires for what would be less worth achieving. We may prefer, for example, to have some enjoyable experience in the nearer future, though we know that, if we waited, our enjoyment would be greater. And 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. In these and other ways, subjective and objective theories often disagree about what we have *most* reason to do, and about what we ought rationally to do.

There are other, deeper disagreements. 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. Our reasons for acting derive their force, not from the facts that our acts would fulfil some of our desires or aims, but from the facts that give us reasons to have these desires or aims.

4 Responding to 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 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, often 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 avoid some painful ordeal, to make someone feel less lonely, or to see the sublime view from the summit of some mountain. We might also have *extrinsic* telic reasons, which would be provided by facts about some possible event's relation to other events. On one view, for example, the punishment of criminals is good, not only as a means of deterring people from committing other crimes, but also as an end, or for its own sake. But such punishment is believed to be good only when it is deserved. This fact about such punishment is not intrinsic, since people cannot deserve to be punished unless they earlier committed some crime.

We need not discuss extrinsic telic reasons, because such reasons can be explained with claims about intrinsic reasons. Events would be *extrinsically* good by being part of some longer event, or sequence of events, which they make *intrinsically* better. On the view just mentioned, for example, though deserved punishment is only extrinsically good, if someone commits some crime and is later punished, this longer sequence of events is intrinsically less bad, and in this way better, than this person's committing this crime and never getting the punishment that he deserves. [59]

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. We can also respond to our reasons to have some desire by coming to have, and continuing to have, this desire. 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 avoid pain, or want to stay alive.

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.

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 voluntary acts are purely mental. If we find ourselves asking, for example, whether we still have enough time to catch some train, we might voluntarily do a mental calculation to answer this question. But if this calculation leads us to believe that we don't have enough time to catch our train, our coming to have this belief is unlikely to be voluntary. We could not, for example, choose to believe that we can catch our train, because we could run 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 have responded in such ways, and we can often give our reasons. [60]

It is worth asking whether our responses to such reasons might take other forms, by being always or often voluntary. Suppose that, when I am aware of certain facts that give me decisive epistemic reasons to have some belief, I fail to respond in the non-voluntary way, by coming to have this belief. Though I can see smoke and flames rising towards me up the stairs of my

hotel, I fail to believe that my life is in danger. Could I correct my mistake, by choosing to have this belief?

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

Suppose instead that I do believe that the smoke and flames give me decisive reasons to believe that my life is in danger. It is unlikely that I could then choose to believe that my 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 some exceptions. We can imagine that, though I believe that the smoke and flames give me decisive reasons to believe that my life is in danger, I don't yet have this second belief. I might then be able to take a further mental step, by choosing to make myself have this belief for these reasons. But even in this imagined case, my response to these epistemic reasons would still be only partly voluntary. When I saw that smoke and flames were rising up the stairs of my hotel, I did not choose to believe that these facts gave me decisive reasons to believe that my 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 non-voluntary ways, by recognizing or realizing what follows from what, and what must be true.

Similar claims apply to our desires. We cannot choose to have most of our desires, or preferences, because we cannot choose either which are the events that we have reasons to want, or how strong these reasons are. 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, character, and emotions, by making ourselves the kind of person that we have reasons to want to be. But this ability would also be dangerous, like our recently discovered ways of moving our bodies at great speed.

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. [61]

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

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

We might have reasons of all these kinds to have the same desire. If you are in pain, for example, I might have all these reasons to want your pain to end. What I want would be in itself good, and it might have the good effect of allowing you to enjoy life again. My wanting your pain to end may be in itself good, and this desire might 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 be called *object-given*. Many people assume that we can also have *state-given* reasons to have certain beliefs: reasons that are 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 *truth-related*, but *goodness-related*, or *value-based*. Such reasons to have beliefs are often called *pragmatic*.

If we can have such state-given reasons, they would not, I believe, have any importance. When it would be better if we were in some state, we might be able to cause ourselves to be in this state. We would then have practical reasons to act in this way. There would be no point in claiming that, as well as having reasons to *cause* ourselves to be in this state, we would have reasons to *be* in this state. Such extra reasons would be idle cogs, which made no difference.

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 make myself lose weight, to buy a bicycle, to learn how to dance, and

to make some friends. There would be no point in claiming that, as well as giving me reasons to act in these ways, these facts would also give me reasons to weigh less, to own a bicycle, to know how to dance, and to have some friends. My reasons to *be* in these states would add nothing to my reasons to cause myself to be in them.

Suppose next that, though it would be better if we were in a certain state, we have no way of causing ourselves to be in this state. In such cases, we would have reasons to *want* to be in this better state, or to wish that we were in it. There would be no point in claiming that we would also have reasons to *be* in this state. I might have reasons, for example, to wish that I were six inches taller, twenty years younger, and could run faster than a cheetah. There would be no point in adding that I would also have reasons to *be* six inches taller, to *be* twenty years younger, and to *be able* to run faster than a cheetah.

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. There is no point in claiming that we also have *state*-given reasons to *have* this belief or desire. And, as I argue in Appendix B, we have some reasons to reject this claim.

5 Object-Given Reasons

On objective theories of the kind I believe we should accept, when we have reasons to act in some way, these reasons derive their force from the facts that give us reasons to have the desires and aims that our acts are intended to fulfil or achieve. These other reasons are, in this way, more fundamental.

What are most important are intrinsic telic object-given reasons. These are reasons to want some possible event as an end, or for its own sake, which are provided by some of this event's intrinsic features. Different objective theories partly disagree about which events we have such reasons to want. Such theories may appeal, for example, 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 promote only our own well-being. According to *Rational Impartialism*, we all have equal reasons to promote everyone's well-being. These views are both, I believe, too simple. Nor should we assume that object-given reasons are provided only by facts about our own or other people's well-being. Some acts or outcomes have reason-giving features, which make them good or bad, though they are good or bad for no one. Of this great variety of reason-giving facts, it will be enough here to consider certain facts about our own

hedonic well-being.

When we have false beliefs, or are ignorant, it may be rational for us to want what we have no reasons to want, or strong reasons not to want. Since we are here discussing reasons, we can suppose that we know all of the relevant, reason-giving facts.

In such cases, our desires are rational when we want some event that we have object-given reasons to want. Our desires are not rational, or in the old phrase *contrary to reason*, when we want some event that we have such reasons *not* to want, and no reasons, or only weaker reasons, to want. When some desire is strongly contrary to reason, because we want some event that we have clear and strongly decisive reasons *not* to want, this desire is irrational. Desires that are more weakly contrary to reason are merely less than fully rational. There is no sharp borderline here, since irrationality is a matter of degree.

We have some partly desire-like states that are not responses to reasons. One large and important group are the *hedonic likings* or *dislikings* of some actual present sensations that make our having these sensations pleasant, painful, or unpleasant, or in which their pleasantness or painfulness consists. It is sometimes claimed that these sensations are in themselves good or bad, since their intrinsic qualitative features give 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 those sensations that some people love and others hate, such as the sensations that are produced 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 sound of house-flies, the feeling of touching velvet, and the flattening, deadening effect of overhead lights. But the oddness of these dislikes does not make me open to any rational criticism. Whether we like, dislike, or are indifferent to these various sensations, we are not responding or failing to respond to any reasons.

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. The nature or quality of some painful or unpleasant sensations may in part depend on whether we dislike them. Such sensations might be claimed to be in themselves bad when their nature 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. [62]

When we are having some sensation that we intensely dislike, we may also strongly want not to be in this conscious state. Such desires about our conscious states we can call *meta-hedonic*. There are several differences between such dislikes and such desires. What we dislike is some sensation. What we want is

not to be having a sensation that we intensely 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 hot poker 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, 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 other desires.

Here are two such further differences. First, when we dislike some present sensation, it is this dislike that makes our conscious state bad, in the reason-implying sense. Our liking some sensation can, in a similar way, make our conscious state good. Hedonic likings or dislikings can thus be claimed to *create* or *confer* such value or disvalue. That does not show that our meta-hedonic desires, or any of our other desires, can create or confer value, by making their objects good or bad. Our future pains are not made to be bad by our present desires to avoid these pains. And when we are in pain, by having some sensation that we intensely dislike, what makes our conscious state bad is our dislike, not our present desire not to be in this state.

Second, while we have no reasons to like or dislike such sensations, we do have reasons to want to have, or not to have, sensations that we like or dislike. Unlike our hedonic likings or dislikings, our meta-hedonic desires *can* be rational or irrational. These desires provide some of the clearest examples, since we can have strongly decisive reasons for and against having such desires.

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. Unless some other fact gives us some opposing reason, it would be irrational knowingly to prefer the more painful ordeal. Such a preference would be most irrational if we preferred the more painful ordeal simply because it would be more painful. That preference may never have been had. When people prefer to undergo the more painful of two ordeals, that may always be because this ordeal would have some other feature, such as being a punishment that these people believe they deserve, or being a way of showing other people how tough they are.

Most preferences of this kind involve our attitudes to time. We may prefer the worse of two ordeals because of a difference in *when* this ordeal would come. 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 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 an attitude that is partly similar: caring more about our nearer future. Unlike these two imagined attitudes, this *bias towards the near* 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 will make this 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. [63]

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 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.

6 Subject-Given Reasons

Subjective theories appeal to facts about our present desires, aims, and choices. According to the simplest desire-based theories,

> (A) each of us has a reason to do whatever would fulfil any of our own present desires.

For desire-based theories to be plausible, however, they must claim that some of our desires do not give us reasons. Suppose again that, having just angered a poisonous snake, you want to run away, falsely believing 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, though this act would be rational, 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 gives you no reason for acting. Subjectivists might claim that

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

You have no reason to run away, (B) implies, because your desire to run away depends on a false belief. Remember next that our desires are *telic* when we want something as an end, and *instrumental* when we want something as a means to some end. You want to run away merely as a means of saving your life. So Subjectivists might claim that

> (C) reasons are provided only by telic desires.

You have no reason to run away, (C) implies, because this act would not help you to achieve any of your ends.

(B) may seem more plausible than (C). When 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. It may be less obvious why instrumental desires provide no reasons. When such desires do not depend on false beliefs, they may not be in any way mistaken.

Subjectivists can defend (C), 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. The tree's owner tells me that I can have only one of these apples. 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 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, I have no separate 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 the higher apple.[64]

As this example shows, Subjectivists must claim that reasons are

provided only by telic desires or aims. And this claim sufficiently explains why you have no reason to run away from the angry snake. There is no need to claim that your desire to run away depends on a false belief.

We can also have telic desires, however, that depend on false beliefs. Suppose, for example, that I want to hurt you, 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 might claim that, when such telic desires depend on false beliefs, they provide no reasons for acting.

As before, there is another way to describe such cases. Rather than saying that we have no *reason* to fulfil such desires, 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. Whenever we want something to happen, what we really want is for this thing to happen *if* certain facts are as we believe them to be. If these facts are *not* as we believe, our conditional desire could not be fulfilled, and this could be why such desires provide no reasons for acting.

Even if Subjectivists need not claim that reasons are provided only by desires that depend on true beliefs, there is no objection to this claim. Such desires we can call *error-free*. According to what we can call

> *the Instrumental Theory*: We have most reason to do whatever would best fulfil our own present error-free telic desires.

This theory is *instrumental* because it treats our ends as given, and tells us only to take the best means to these ends.

Some Subjectivists appeal, not to our *actual* desires, but to the desires that we would now have if we knew more. That is an obvious way to extend the Instrumental Theory. If we deny reason-giving force to desires that depend on false beliefs, we can plausibly deny that force to desires that depend on ignorance. This distinction is not deep. When I want to hurt you, there are at least 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 acted in this way, I might not know that your motive was to save me from some greater injury. There is little difference between these cases. 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, it seems plausible to take a further step. Subjectivists can claim that, just

as we do *not* have reason to fulfil those of our actual telic desires that we would *not* have if we knew more, we *do* have reason to fulfil the telic desires that, with greater knowledge, we *would* have. As before, this distinction is not deep. If I had known that you had good motives for injuring me, I might not only have ceased to wish you ill, but also come to wish you well. If that is true, Subjectivists might claim, I have a reason to treat you well.

If we appeal to what we would want if we knew more, we can plausibly 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 that we would now have if we knew all of the relevant facts.

Facts count as *relevant*, some writers claim, if our knowledge of these facts would affect our desires. But this criterion is too wide. As Allan Gibbard remarks, if we knew 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 appalled or depressed that we would lose our desire to enjoy ourselves. Such desire-affecting facts, some Subjectivists claim, should not be taken to be relevant. On this view, 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. In certain cases, if we were fully informed, that would change our situation in some way that would give us different reasons. If we knew certain important facts, for example, we might lose our desire to find out these facts, but that would not make it true that in our actual state, in which we don't know these facts, we have no reason to try to fulfil this desire. 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.

Subjective theories can take other forms. Some subjective theories appeal, not to *desire*-based, but to *aim*-based reasons. Other theories appeal to the choices that we would make after some process of deliberation. 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. These two kinds of theory can be stated in deceptively similar ways. Subjectivists and Objectivists might both claim that

> (D) 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. According to Objectivists, when it is true that

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

that would make it true that

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

According to Deliberative Subjectivists, when it is true that

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

that would make it true that

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

These views are very different. On both views, when we are deciding what to do, we should deliberate in certain ways. We should try, for example, to imagine fully the effects of our different possible acts, to avoid wishful thinking, and to assess probabilities correctly. If this is how we decide what to do, we are *procedurally* rational.

Objectivists make further claims about what we have reasons to want, and to do. There are many telic desires and aims, Objecivists claim, that we have decisive object-given reasons to have. We have such reasons, for example, to want to avoid being in agony. To be fully *substantively* rational, we must respond to these reasons, by having these desires and aims, and trying to achieve them.

Deliberative Subjectivists make no such claims. On such views, there is only procedural rationality, since there are no object-given reasons. If we already have certain telic desires or aims, we may be rationally required to do what we know would best achieve these aims. But we are not rationally required to have any particular telic desires or aims.

To illustrate these theories, 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 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 reasons or apparent reasons 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 act in this way. On this view, however, the inference runs the other way. Instead of

claiming that what we ought to choose depends on our reasons, Subjectivists claim that our reasons depend on what, after ideal deliberation, we would choose. If I have decisive reasons to stop smoking, that is because I would choose to act in this way.

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. We can suppose that, in our examples, we know all the relevant facts, and that, after ideal deliberation, we would choose to do whatever would best fulfil our present telic desires or aims. 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 are other disagreements between different subjective theories that we can mention but then ignore. Suppose that, given the facts, all subjective theories imply that I have a decisive reason to stop smoking. On some subjective theories, this reason is given by the fact that

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

On some other theories, this reason is given by the fact that

> (2) this act would preserve my health, giving me many more years of happy 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 preserve my health, but this fact gives me this reason only because I want to achieve this aim. Though (2) is the fact that *gives* me my reason, this reason *depends* on (1), since it is (1) that makes (2) give me this reason. Similar claims apply to the fact that

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

According to the Deliberative Theory, we have decisive reasons to do whatever, after ideal deliberation, we would choose to do. But our reasons to do something cannot be plausibly claimed to be given simply by the fact that we would choose to do this thing.

So Deliberative Theorists should claim that (2) gives me my reason to stop smoking, but that this reason depends on (3), since it is (3) that makes (2) reason-giving.

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 facts give us these reasons. When I say that, on these theories, our desires, aims, or choices can give us reasons, that is short for the claim that certain facts about these mental states can help to make it true that we have reasons to act in certain ways.

There is another kind of theory that I should briefly mention. Subjectivists, I have said, appeal only to claims about procedural rationality. But some writers, though appealing only to such claims, are much closer to Objectivists in their beliefs about what we ought rationally to want, and to do. What we ought to do, these people claim, depends on what we would want, or choose, if our desires were, in strong senses, *coherent* and *systematically justified*. On this view, if we cared about our present agony but not about our future agony, our desires would not be coherent or systematically justified. To be procedurally rational, we must care equally about avoiding agony at any time. These people also claim that, if we cared only about our own well-being, our desires would not be fully coherent or justified. To be fully procedurally rational, we must care about everyone's well-being. According to some Kantians, if we set ourselves any end or aim, we are not fully procedurally rational unless we also value the capacity of all other rational beings to set their ends, and we commit ourselves to treating others only in ways to which they could rationally consent. [65]

These people's theories are, in a way subjective, since they appeal to what we would want or choose if we were fully informed and in these demanding ways procedurally rational. But I shall use 'procedurally rational' in its ordinary thinner sense, [66] and I shall call these people, not Subjectivists, but *Systematic Coherentists*. Like Objectivists, these Coherentists believe that we are rationally required to have certain telic desires or aims, such as a desire to avoid all future agony. That is true, Objectivists believe, because the nature of agony gives us decisive reasons to have this desire. These Coherentists defend such beliefs in a quite different way. On their view, we have no such reasons to want avoid agony. Nor do we have such value-based object-given reasons to care about or to value other things, such as the well-being of others, or rational agency. We are rationally required to have such desires, concerns, and values, not because we have such *reasons* to have them, but because our failure to have them would make our pattern of concern incoherent, or systematically unjustified. Since such views are very different from the views that I call Subjectivist, I shall not discuss them further here. But if we have value-based object-given reasons, as I believe, we should reject

Systematic Coherentism. We should claim that some things matter, in the sense that we have such reasons to care about these things.

CHAPTER 3 AGAINST SUBJECTIVISM

7 Why People Accept Subjective Theories

Many people accept some form of Subjectivism. We ought, I believe, to reject all subjective theories, and accept some objective theory. [67]

Since so many people believe that *all* 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, 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 value-based 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 reasons. [68]

Third, some people accept desire-based theories about well-being. According to such 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 of us fail to distinguish 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 them to achieve their aims, these facts may give us *non*-derivative reasons to act in these ways. When other people have desires or aims that they have no reasons to have, these people may have no reasons to try to fulfil them. But *we* may have such reasons. In helping these people to fulfil their desires or aims, we respect their autonomy, and avoid paternalism. Other people's desires, aims, or choices are, in this respect, like votes, which should be given just as much weight even when the voters have no reason to vote as they do. Many people accept desire-based or choice-based theories because they are democrats, or liberals, who believe that we should not tell other people what they ought to want or choose. [69]

Seventh, when we have some aim, and we know 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, as I have said, that is not so. When our aims depend on false beliefs, or we have no reason to have these aims, we may have no reason to do what we ought rationally to do, by trying to achieve these aims.

Eighth, when people claim that we have reasons to fulfil our present desires, they are often thinking of desires for future activities or experiences that we believe we would enjoy. When these beliefs are true, as they often are, we do have reasons to fulfil these desires. But these reasons are provided, not by the facts that we would be fulfilling these desires, but by the facts that we would enjoy these 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 don't enjoy, their parents sometimes say, 'See! You didn't *really* want that'. Such claims are false, since these children *did* want these things, and the truth is rather that their desires didn't give them reasons. Similar claims apply to our desires to avoid what we believe would be painful, or unpleasant. When people claim that our desires give us reasons, it is often such facts about what we would enjoy, or find painful or unpleasant, that they really have in mind. Such facts give us reasons that are *hedonic* rather than *desire-based*.

Ninth, some people mistakenly believe that hedonic reasons *are* desire-based. When these people think about sensations that are painful or unpleasant, they do not distinguish between our dislike of these present sensations and our meta-hedonic desires not to be having sensations that we dislike. It is our dislike, I have claimed, that makes our conscious state bad, and gives us our reason to try to end our pain. 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 many 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, this 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 many kinds of game, for example, such as games without rewards that involve pure 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 fact would then give us a reason to try to fulfil this desire.

In describing such cases, we need 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*. My reason to stop smoking, for example, is given by the fact that this act would preserve my health, 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 this desire, so this reason is desire-based. The reasons that I have just described are quite different. When some act would give us pleasure, this fact gives us a reason to act in this way. This reason may *causally* depend on our having some desire, since that may be why this act would give us pleasure. But this reason would not *normatively* depend on this desire. If some act would give us pleasure, this fact gives us a reason whether or not this pleasure causally depends on our having some desire. And such hedonic reasons, as I have said, are not desire-based.

There are 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 be worth achieving. But such cases may involve desires that we have no 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, our having this desire 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 aims. When we have decided to try to fulfil some desire, thereby making its fulfilment one of our aims, this decision may give us further reasons to try to achieve this aim. But these reasons would not be provided merely by the fact that we have made this decision and adopted this aim. These reasons would be provided by the facts that, if we do not try to achieve such aims, we shall be on the whole less likely to achieve our aims, 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. [70]

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, it is not surprising that so many people accept subjective theories. Many of these people have various true or plausible beliefs about which facts give us reasons, and they have merely failed to see that these beliefs do not in fact support any subjective theory. These people are not really Subjectivists.

8 Analytical Subjectivism

There is another way in which some people have come to accept subjective theories about reasons. As both Kant and Sidgwick warn, when we think about normative questions, we can be easily misled by claims that seem important but are merely *concealed tautologies*. In Kant's words:

There is no science so filled with tautologies as ethics. [71]

An *open* tautology uses the same words twice, in a way that does not make any *substantive* and informative 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, substantive 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. 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. We might similarly claim that

(4) acts that produce happiness are felicific.

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

(2) acts that produce happiness produce happiness,

and

(5) acts that are felicific are felicific.

As before, these claims are not substantive. 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', this claim *is* substantive, and conflicts with many people's beliefs. Many of

us 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* senses. We can call these people *Analytical Subjectivists*. When some of these people say that

(7) we 'ought' to do something,

they mean that

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

This subjectivist sense of 'ought' we can call the *desire-fulfilment* sense. Some of these people claim that

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

Since these people use 'ought' 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 believed. For (9) to make a substantive claim, these Subjectivists must use 'ought', in some other, non-subjectivist sense, such as what I have called the decisive-reason-implying sense. These people are *Non-Analytical Subjectivists*.

Though Analytical Subjectivism is not a substantive normative theory, this view makes some important claims, which I discuss in Appendix A. I shall here discuss substantive, non-analytical subjective theories.

According to these theories, as I have said, all reasons for acting are desire-based, aim-based, or choice-based. It will be simpler to consider cases in which different subjective theories coincide, because we know all the relevant facts, and the acts that would best fulfil our present telic desires are also what we would choose to do after ideal deliberation. Our deliberation is *ideal* when we are fully informed and procedurally rational. I shall mostly discuss desire-based reasons. These reasons are, in one way, primary, since our aims are often the desires that we have decided to try to fulfil, and our choices are often intended to achieve our aims. Many of my claims about desire-based reasons would also apply to aim-based and choice-based reasons.

In discussing subjective theories, I am using 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* or *include* 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 more particular facts that are involved in romantic or sexual love, or in love for our children. [72] As one example of 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.

9 Why We Ought to Reject Subjective Theories

Subjective theories have implausible implications. Suppose that, in

> *Case One*, I know that some future event would cause me a 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 desires or aims, 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 claim 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 agony. But this claim cannot be made by those who accept desire-based subjective theories of the kind that we are considering. Given their assumptions, these people cannot claim that our *future* desires give us reasons. Most of these people

assume, for example, that all reasons involve actual or possible *motives*, and only our *present* desires could now motivate us.

Some other kinds of theory do claim that we have reasons to want to fulfil, and to try to fulfil, our predictable future desires. 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 rationally 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 now in ways that would cause these future *desires* to be fulfilled. We might be claimed to have similar value-based reasons to do what would fulfil other people's desires, because such acts would promote these other people's well-being.

We can also imagine a temporally neutral desire-based theory about reasons. On this view, we have most reason to do whatever would best fulfil all of our desires throughout our life, whether or not the fulfilment of these desires would be good for us. There could be a similar, personally neutral theory, which claimed that we have most reason to do whatever would best fulfil everyone's desires throughout their lives, whether or not the fulfilment of these desires would be good for anyone. These imagined theories are very different from the widely accepted subjective theories about reasons that we are now discussing. [73]

According to these theories, as I have said, it is only facts about our own present desires or aims that can give us reasons. 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 some approaching fire---I would 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 have no reason to avoid this agony, they would mean only that, after ideal deliberation, I am not motivated to act in this way. Everyone could accept that, in *this* sense, I have no reason to move my hand away from the approaching fire. But many of use the phrase 'a reason' in some other, non-subjectivist sense. We believe that, though I have no desire to avoid this agony, it makes sense to claim that I have a reason to have this desire, and a reason to avoid this

agony, if I can. According to substantive, non-analytical subjective theories, I have no such reasons. That is hard to believe.

If Non-Analytical Subjectivists find this claim hard to believe, they might suggest that my example is no objection to their theory, because this case is purely imaginary. Every actual person, they might say, wants to avoid all future agony.

This reply would not succeed. First, we are asking whether subjective theories imply that we have *reasons* to want to avoid all future agony. To support the claim that we have reasons to have this desire, it is not enough to claim that everyone *has* this desire. These Subjectivists would also have to claim that our having this desire 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 the prospect of future pain, if this pain would be far enough in the future. Of those who have believed that sinners would be punished with eternal agony in Hell, many tried to stop sinning only when they became seriously 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. If we claim that reasons are provided by facts of certain kinds, we cannot defensibly restrict our claim to actual cases. According to subjective theories, what we have reasons to do depends on facts about what would best fulfil our present fully informed desires, aims, or choices. This claim cannot be true in the actual world unless it would also have been true in possible worlds in which evolution produced human beings who were just like actual human beings, except that they did not want to avoid all future agony, or their desires differed from ours in other ways. So we can fairly test subjective theories by considering such cases.

Subjectivists might next reply that, if some normative theory has acceptable implications in all or most actual cases, that may give us sufficient reasons to accept this theory. We might justifiably accept such a theory even if in some unusual or imagined cases this theory seems to go astray.

Some theories 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 applied to cases that involve actual people, subjective theories often have plausible implications. But that is because most actual people often have desires that they have object-given reasons to have, since what they want is in some way worth

achieving. In such cases, subjective theories have the same implications as the best objective theories. To choose between these two kinds of theory, we must consider cases in which these kinds of theory disagree. Such disagreements take their clearest form in some unusual actual cases and some imaginary cases. So we cannot defend subjective theories with the claim that we can ignore these cases, or 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, as 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.

Some Subjectivists might give a different reply. These people appeal to the desires or aims that we would now have, or would now want ourselves to have, if we had gone through some process of fully informed and *rational* deliberation. So these people might claim that

> (F) in such cases, if we were fully rational, we would want to avoid all future agony.

As I have said, however, such claims are ambiguous. Objectivists could appeal to (F), because they make claims about *substantive* rationality. On such theories, we all have decisive reasons to have certain aims, and to be substantively rational we must have these aims. These reasons are object-given, in the sense that they are provided by the intrinsic features of what we would be trying to achieve. One such rationally required aim is avoiding future agony. If we did not want to avoid all future agony, we would not be fully substantively rational, because we would be failing to respond to our object-given reasons to have this desire and aim.

Subjectivists cannot make such claims. Subject-given reasons are provided, not by the intrinsic features of what we want, but by facts about what would best fulfil our present telic desires, aims, or choices. On subjective theories, we have no object-given reasons to want to avoid future agony. Subjectivists appeal to the desires or aims that we would now have after deliberation that was *merely procedurally* rational. On such theories, *if* we have certain telic desires or aims, we may be rationally required to want and to do what would achieve these aims. But we are not rationally required to *have* any particular telic desires, or aims. We can be procedurally rational whatever we care about, or want to achieve.[74] In Rawls' words,

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently.[75]

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

Subjectivists might reply that, even on their theories, we can be rationally required to have certain telic desires or aims. That could be claimed, for example, of any telic desires or aims without which we could not even be rational agents, in the sense that we could not act, or respond to reasons.

This reply does not succeed. There may be some telic desires without which we could not act, or respond to reasons. Perhaps, as Bernard Williams suggests, any rational agent must have 'a desire not to fail through error', and some 'modest amount of prudence'. [76] But these desires cannot be claimed to include a desire to avoid all future agony. We can add that, insofar as some theory requires us to have certain telic desires or aims, this part of this theory is not Subjectivist.

We have seen that, in *Case One*, subjective theories imply that I would have no reason to want to avoid some future period of agony, and no reason to prevent this agony if I can. These claims are hard to believe. Suppose next 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 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 also have no other present desire or aim that would be frustrated by 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 to act on my decision, by causing myself to be in agony. 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. This claim is even harder 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 to be in this state. [77] 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. Subjectivists might also claim that some desires could not possibly be had by any rational agent, since such desires would make it impossible to act, or to respond to reasons. But no such claim applies to a desire to have some future period of agony. Having this desire is not incompatible with being an agent.

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. We assume that the nature of agony gives everyone very strong reasons to want to avoid being in agony. 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 it would be unusual for someone to have this desire. 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. [78]

According to subjective theories, we have most reason to do whatever would best fulfil our present fully informed desires, aims, or choices. We have seen that, in *Case Two*, these theories implausibly imply that I would have most reason to cause myself to be in agony. Subjectivists might now revise their view. They might claim that

> (G) 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 (G), 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 have carefully considered all of the relevant facts. Subjective theories imply that, in such cases,

> (H) 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 have a desire-based reason to want to avoid some future period of pain. Subjective theories imply that

> (I) if some possible event would have effects that we want, this fact gives us a reason to want this event as a means to these effects.

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 imply that I have a reason to want to avoid this headache as a means of helping me to fulfil my desire to win this game. 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, by helping me to fulfil any of my other present desires or aims. On these assumptions, (I) does not imply that I have any reason to want to avoid this agony.

Subjectivists might also claim that

> (J) when it is true either that
>
> > (1) our *having* some desire or aim would have effects that we want,
>
> or that
>
> > (2) 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. [79]

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 and dread about what lies ahead. For these reasons, I don't want to have this desire, and would prefer not to have it. On these assumptions, (J) does not imply that I have any reason either to have my desire to avoid this agony, or to cause myself to keep this desire.

Since I have no *other* present desire or aim that gives me any reason to want to avoid this agony, Subjectivists might now claim that this desire *itself* gives me such a reason. More generally, they might claim that

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

If (K) 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 would give me a reason to want to be in agony. If I wanted to waste my life, this would give me a reason to want to waste my life. *Whatever* we want, we would have reasons to have all of our informed desires or aims. Since these claims are clearly false, [80] Subjectivists must reject (K).

Since Subjectivists cannot appeal to (K), 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.

Consider next a different version of this case. Suppose that, because I *could* avoid this future agony, my having this desire would 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. More generally, Subjectivists might claim that

> (L) if we have some fully informed desire, and our having this desire would lead us to do what would fulfil this desire, these facts would together give us a reason to have this desire.

But if (L) were true, all such fulfillable desires would be rationally self-justifying. If we wanted to be in agony, and our having this desire would lead us to thrust our hand into some fire, these facts would give us a reason to want to be in agony. If we wanted to waste our lives, and our having this desire would lead us to waste our lives, these facts would give us a reason to want to waste our lives. Since these claims are clearly false, Subjectivists must reject (L). These people cannot claim that my desire to avoid my future agony gives me a reason to have this desire. Since I have no other present desire that gives me any such desire-based reason, these people must again admit that, on their view, I have no reason to have this desire. So subjective theories imply that, in both versions of *Case Three*, I have no reason to want to avoid my future agony.

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 we 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 we have no other present desire or aim that would be fulfilled by the achievement of these aims, and no desire or aim that could be claimed to give us a reason have these aims. 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 other good aims.

Return now to the claim that

> (G) 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 (G), 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 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 (G).

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 theories, any such reason would have to be provided by one of our actual present desires, or by some desire that we would now have, or would want ourselves to have, after ideal deliberation. As we have seen, this must be some *other* desire. For this other desire to give us this reason, we must have some reason to have this desire which must in turn be provided by some *other* desire. We cannot have a beginningless chain of such 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 (G), they would have to claim that we never have any reason to do what would fulfil any of our desires. As before, this claim contradicts their view, and is clearly false.

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

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

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. After ideal deliberation, I have decided to cause myself to have this agony, if I can. Since Subjectivists must accept (M), they must now admit that, on their view, I have most reason to cause myself to be in agony. 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.

There are other, similar cases. According to subjective theories, if we had fully 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 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. If these are substantive normative claims, rather than merely the implications of some desire-based or choice-based sense of the phrase 'have a reason', 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. We must admit that, in *Case Two*, I have no reason to fulfil my desire to be in agony. But that does not show that, in *Case Three*, I have no reason to fulfil my desire *not* to be in agony. It might be similarly claimed 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, and 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, and 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. If we make such claims, we have moved to an objective theory. And as we have seen, Subjectivists cannot claim that, for some desire or aim to give us a reason, we must have some reason to have this desire or aim. Subjectivists cannot distinguish in these ways between desires that do or don't give us reasons. So we can now argue:

> (N) If we have desire-based or aim-based reasons, all that would matter is *whether* some act would fulfil or achieve our present fully informed telic desires or aims. It would be irrelevant *what* we want, or would be trying to achieve.

Therefore

> (O) Either *all* such fully informed desires or aims give us reasons, or *none* of them do.
>
> (P) If all such desires or aims gave us reasons, we could 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.
>
> (Q) We could not have such reasons.

Therefore

> (R) None of these desires or aims gives us any reason.

When we want to avoid agony, to be happy, and to achieve other good aims, we do indeed have reasons to try to fulfil or achieve these desires or aims. But these reasons for acting are provided, not by the fact that we have these desires or aims, but by the features of what we want, or are trying to achieve, that make these things relevantly good or worth achieving, and give us reasons to have these desires or aims.

Here is another way to sum up these claims. 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 such events 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 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, to waste our lives, and to try achieve 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 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 either to have any desire or aim, or to act in any way. [81]

10 Fully Informed Desires

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, claims about reasons must apply 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 want to achieve 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 *coincide* with, or imply the same as, the best objective theories.

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

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

According to an implausible imaginary theory, which we can call

> *the belief-based theory*, we ought to believe whatever, after considering the facts, we *do* 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 are known to us 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 wholly ignores our reasons to have our beliefs. And, when this theory has acceptable implications, that is because most actual people do *not* ignore such reasons, but respond to them.

Similar claims apply to theories about practical reasons. According to

> objective theories, we ought to have and to try fulfil the telic desires or aims that we have decisive reasons to have.

According to

> subjective theories, we ought to try to achieve our present fully informed telic desires or aims, whatever these are.

When applied to actual people, subjective theories often have acceptable implications. Since most of us often have the desires or aims that we have decisive reasons to have, subjective theories often imply that we ought to try to fulfil the desires or aims that we have such reasons to have. But that is not what these theories claim. In their claims about what we ought to do, these theories wholly ignore our reasons to have our telic desires or aims. When subjective theories have acceptable implications, that is because most actual people do *not* ignore such reasons, but respond to them. Subjective 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 considering cases in which we know and have carefully considered all of the relevant facts. In many cases, however, we do not know all these facts. Many Subjectivists claim that, in such cases,

> (S) 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, or would now want ourselves to have, if we knew and had rationally considered all of the relevant facts.

These writers also claim that

> (T) when we are making many important decisions, we ought to try to learn more about the different possible outcomes of our acts, so that we can come to have better informed 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 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 carefully consider, what these outcomes would be like. But, if we make this claim, we are assuming some objective theory. We are assuming that

> (U) 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.

Subjectivists cannot appeal to (U). According to these people, no such features of possible outcomes ever give us such object-given reasons. If that were true, we would have no reason to try to discover, and carefully consider, what these outcomes would be like. And we would have no reason to accept (S). 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, or would want ourselves to have, if we had rationally considered all of the facts about the possible outcomes of our acts. If these facts would not give us reasons to *have* these desires or aims, 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. [82]

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. [83]

But most Subjectivists do not see that, given their assumptions, we may have no reason to try to have and to fulfil such better informed desires or aims.

Though Subjectivists cannot coherently claim (S) and (T), they can claim that

> (V) we ought to try to discover the facts about how we can best fulfil our *actual* present desires or aims, whatever these are.

Subjectivists can coherently make this claim because (V) does not assume that the possible outcomes of our acts may have intrinsic reason-giving features. On this view, the relevant facts do not include facts about what the different possible outcomes would be like, except when these are facts about which acts would best fulfil our actual present desires or aims.

Subjectivists can also claim that

> (W) 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 come to 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, (W) implies that most of us ought to try to have such desires or aims. But, as before, these claims would not support subjective theories. Most of us want to have better informed desires or aims because we assume what objective theories claim. We believe that the possible outcomes of our acts may have features that would give us reasons to want to produce or prevent these outcomes, if we can. [84]

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 reasons to try to fulfil our desires, we must have value-based object-given reasons to have these desires. 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 aims, our desires or preferences can break ties, by giving us reasons to adopt one of these aims.

I believe, though not very strongly, that we ought to reject even this hybrid theory. When we have certain desires, that may make it true that we have further reasons to try to fulfil these desires. But these further reasons are provided, I believe, not by the fact that we have 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, I concede, show that we should reject this theory. This question would then remain open. But this question would not have much importance, since this hybrid theory is fundamentally objective and value-based.

11 Reasons, Motives, and Well-Being

We can now return to the ways in which events or outcomes can be either 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.

According to subjective theories about reasons, no events or outcomes 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 have been better fulfilled if, in some remote place or later period of time, some 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 such 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. [85]

We can turn next to the ways in which some events could be better *for* particular people, in the sense of being more in these people's interests, 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 this person to live, for this person's sake.

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. On subjective theories, as we have seen, we have no such object-given reasons.

Most of us want to promote our future well-being. Many Subjectivists assume that, since we have this desire or aim, we have self-interested reasons for acting. But this assumption is a mistake. As I have argued, subjective theories cannot defensibly claim that we have *reasons* to have and to try to fulfil this desire or aim. And that is what is meant by the claim that everyone has self-interested reasons to care about, and try to promote, their future well-being.

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 John 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. [86]

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. [87]

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 in fact choose.

Though it is a normative question which kinds of deliberation are procedurally rational, and in this way ideal, it is a psychological question what, after such deliberation, someone would in fact choose. [88] 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. [89]

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. [90] 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'. [91] 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 carefully 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. But, as I have said, it is irrelevant whether such a case might actually occur. Rawls does not assume that any actual person would choose to spend his life counting the numbers of blades of grass in various lawns. 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, if it is clear enough what these cases would involve. And though it is hard to believe that anyone would choose a life of unrelieved suffering, that is because this life would be so obviously bad for this person in the reason-implying sense. It is hard to believe that anyone could be so irrational. On Rawls's view, however, no life could be bad for someone in this sense.

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. [92] 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 accepts a subjective theory 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 most fulfilled.

We can call these the *hedonistic* and *temporally-neutral desire-based* senses of the phrase 'best for someone'. [93] 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 fulfilled. [94]

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, he would have a smaller sum of happiness minus 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 much less fulfilled. Neither of these claims is normative.

Similar remarks apply to other cases. When people use 'best for someone' in either of these senses, they make themselves unable to express their substantive beliefs about which lives would be best for people, since their attempts to express these beliefs become true by definition. There is no point in claiming that, if some possible life would be the happiest, or would do most to fulfil someone's desires, this life would be the happiest, or would do most to fulfil these desires.

These people *could* make substantive claims 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

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

This means

> (Y) If some possible life would give someone the most happiness, and be the life in which this person's desires would on the whole be most fulfilled, that 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 others would have the strongest reasons to want this person to live, for this person's sake.

This claim *is* normative, 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

> (Z) even if a life of unrelieved suffering would be, in other senses, bad for Blue, this is the life that Blue has most reason 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 cannot have object-given reasons to want ourselves or others to live happy lives, nor can we have such reasons to have any other good aim. And, as I have argued, we cannot be defensibly claimed to 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. [95] 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.

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 give us any such reasons.

These bleak views are seldom defended. Most Subjectivists take it for granted that all reasons are provided by facts about what would fulfil our present desires, or aims. [96]

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 be motivated to do this thing.

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 to reply, 'Because I didn't want to'. In most 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 also argue:

> (3) For some fact to give us a reason, it must be possible that we 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. [97]

But (5) is false. We cannot defensibly claim that, whenever

people are motivated to act for some reason, this reason must be subject-given and desire-based *rather* than object-given and value-based. That claim would have to assume that, for some reason to be object-given and value-based, it must be impossible for anyone to be motivated to act for this reason. And that assumption would be absurd. If some act would achieve some aim that is good or worth achieving, we might be motivated to act for this reason.

There is another line of thought that leads many people to accept subjective theories. These people make some meta-ethical assumptions that I discuss in Appendix A, and shall mention only briefly here. On the best versions of what I call objective theories, the fact that we have some reason is an *irreducibly normative* truth. Of those who accept subjective theories, many are *naturalists*, who believe that there cannot be such truths. According to 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 naturalists give *reductive* accounts of desire-based or aim-based reasons for acting. Some naturalists are Analytical Subjectivists. According to these people, when we claim that someone has a reason to act in some way, we *mean* that this act would or might fulfil one of this person's telic desires or aims, or we mean that after ideal deliberation, this person would be motivated to act in this way, or we mean something else of this kind. According to some *non-analytical* naturalists, though the *concept* of a reason is irreducibly normative, the *fact* that someone has a reason is, or consists in, such a causal or psychological fact.

These reductive desire-based or aim-based theories can seem plausible if, like many writers, we fail to distinguish clearly between reasons and motives, and 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. These may seem the best ways to explain the normativity of these reasons. As three of these people write:

> 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. [98]

Value-based object-given reasons cannot be regarded in such ways, since we have such reasons even if we would *not* be motivated to act upon them. [99]

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, since things matter only in the sense that we

care about these things, and these concerns can move us to act.[100]

Such naturalist accounts of reasons are, I believe, mistaken. I defend this belief in Appendix A, but I shall make one remark here. If naturalism were true, we could not have normative 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.[101]

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

CHAPTER 4 RATIONALITY

12 Practical and Epistemic Rationality

We can now turn from reasons to rationality. 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 reasons in these ways. To *fail* to respond to some reason, we must be aware of the facts that give us this reason. If we do not respond to some reason because we are not aware of the reason-giving facts, we are not failing to respond to this reason.

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 our *apparent* reasons. We have some apparent reason when we have beliefs about the relevant, reason-giving facts, and what we believe would, if it were true, give us some reason. These beliefs should be taken to include implicit assumptions, such as the assumption that some act would not harm ourselves or others. To save words, I call these *beliefs whose truth* would give us some reason. We can now look more closely at how the rationality of our desires and acts depends on our beliefs.

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. I might want to go to sleep, for example, 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 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. [102] I shall now briefly sketch my view about how, when they are justified in such ways, the rationality of our desires and acts *normatively* depends on 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 drug 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.

I believe that:

> (1) 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.
>
> (2) In most cases, it is irrelevant whether these beliefs are true, or rational. Some of the exceptions involve certain normative beliefs.
>
> (3) 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 must choose between several possible acts, but we have no beliefs about the relevant, reason-giving facts, there may be nothing that we ought rationally to do.
>
> (4) Our having some desire is in one way rational when and because this desire itself is rational. But in some cases it would be rational to 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 give some examples in Appendices B and C. [103]

Many people would reject some of these claims. Our desires are irrational, some people suggest, just when they causally depend on false beliefs. [104] 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 believe that smoking will protect my health. 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 our desires 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. We act rationally, in most cases, 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 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 mean, roughly, 'open to strong criticism of the kind that we express with words like "foolish", "stupid", and "senseless"'. It would also be better to make different claims about which desires or acts are open to such criticism.

Of those who claim that the rationality of our desires *normatively* 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 claim, 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.

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 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

(5) smoking protects my health

and that

(6) I am now smoking,

I believe that

(7) I am now protecting my health.

My belief in (7) is in one way irrational, since this belief depends in part on my irrational belief in (5). But my belief in (7) is in another way *rational*. This belief is *rationally derived* from my beliefs in (5) and (6), in the sense that these are beliefs whose truth would give me a decisive reason to believe (7). 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

(8) 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.

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. We can call such desires or acts *rationally supported* by these beliefs. And we might suggest that

(9) 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 (9), 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. 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 (9), there is one way in which Green and Grey are both acting irrationally, since these people's acts both depend on an irrational belief. 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 (9) is fairly plausible, this is not, I believe, the best view. According to (9), our desires and acts can be irrational when and because we are failing to respond to some epistemic reason or apparent reason. My act is claimed to 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 this act *practically* irrational, since what makes this act irrational 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 reasons. [105]

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. To use a different metaphor, when some belief is epistemically irrational, this irrationality can, like a virus, *infect* some of our other beliefs. But, with a few exceptions, this irrationality cannot be transmitted over the gap between our beliefs and our desires and 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, not on whether our beliefs are rational, but 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 the other way round.

13 Beliefs about Reasons

We can have rational beliefs and desires, and act rationally, without having any beliefs about reasons. Young children may respond rationally to certain reasons or apparent reasons, though they do not yet have the concept of a normative 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 fact when he declares that no desires or preferences could be unreasonable.

If we have beliefs about which facts give us reasons, our desires and acts are often rational responses to what we believe. But that is not always true. Most of us want some things that we believe we have no reasons to want and strong reasons not to want. That is true of most of the exhausted parents who want to hit their howling babies, and it is true of me when I want to smash some malfunctioning machine. When we believe that we have no reason to have some desire, and some reasons not to have it, our having this desire is not fully rational. Such desires, we can say, are in one way *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. And I have claimed that, in such cases, it is irrelevant whether our beliefs are true, or rational. These claims do not apply, however, 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, this fact would give us a reason to have this desire, and this desire would be rational. But that does not make it true that we have such a reason, nor does it make our 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 is not enough to give us such a reason, or to make our act rational. Rationality is not so easily achieved.[106]

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

As before, however, that claim applies only to most cases. In

cases that involve beliefs 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 certain beliefs about practical reasons. That is true, for example, when we are motivated by the belief that something would be good, or worth achieving, in the reason-implying sense. [107] Given the very close relation between these desires and beliefs, the rationality of these desires *does* in part depend on the rationality of these beliefs. And, if we have irrational beliefs about practical reasons, and about what we ought rationally to do, our having such beliefs makes us in one way practically irrational.

There is a similar overlap between practical reasons and certain epistemic reasons. For example, we have a practical reason 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 facts give us reasons. It is enough if we are responding rationally to our awareness of the facts that give us reasons, or we are acting on beliefs 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 we also want, intend, and try to do whatever we believe that we have decisive reasons to want, intend, and try to do.

To illustrate these claims, suppose that *Red* has the attitude that I earlier called *Future Tuesday Indifference*. This imagined man cares about his future well-being, including all his future pleasures and pains, *except* when these pleasures or pains would come on any future Tuesday. Given the choice, Red would prefer agony on any future Tuesday to mild pain on any other future day. Red has no beliefs about whether he has reasons to have this preference, and he has no other relevant false beliefs.

When Scanlon discusses this example, 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'. [108]

We are irrational in the ordinary sense when our beliefs, desires, or acts make us open to strong rational criticisms of the kind that people express with words like 'senseless', 'foolish', 'stupid', and 'crazy'. When we are open only to weaker criticisms of this kind, we are merely less than fully rational. If Scanlon is using 'irrational' in this ordinary sense, his claims are not, I believe, justified. When Red prefers agony on a future Tuesday to slight

pain on any other future day, he is not failing to respond to what he believes to be some relevant reason. But, in having this preference, Red is failing to respond to a very clear and strong reason. We all have strong reasons to prefer slight pain to agony. And, if Red's agony would be on a future Tuesday, that does not give him the slightest reason to care about it less. Unless we add some unusual details to this imagined case, these facts are enough to make Red's preference irrational.

Suppose next 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 the nature of agony and slight pain, and about personal identity, time, and the other relevant non-normative facts. And these people all believe that, when other things are equal, everyone has strong reasons to prefer slight future pain to future agony. But these people differ in some of their other beliefs about reasons.

Scarlet differs from Red only in his normative beliefs. On Scarlet's view, we all have reasons to care about our future pleasures or pains, except when these would come on any future Tuesday. Since tomorrow is a Tuesday, Scarlet believes that he has decisive reasons to prefer an hour of agony tomorrow to a minute of slight pain on any other day of the week. Scarlet has this preference, so he chooses the agony.

Crimson's view is closer to the views that many actual people accept. Crimson believes that, though we all 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 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 decisive reasons to prefer five minutes of slight pain later today to six minutes of slight pain tomorrow. Pink, however, prefers and chooses the slightly longer pain tomorrow.

When Scanlon discusses Scarlet, he writes:

> what reason can he have for thinking that it makes a difference whether something happens to him on a Tuesday rather than on some other day of the week? If he has a positive reason for this (even a very implausible one, such as some strange theory of well-being), then I have no hesitation in saying that he is not irrational, just seriously mistaken in his assessment of the reasons that he has. [109]

Suppose that Scarlet accepts some strange theory, since he believes that we have no reason to care about our well-being on future days of the week whose name starts with the letters 'T' and 'U'. On Scanlon's view, since Scarlet has this belief, his preference for the hour of agony is not irrational. [110] As before, if Scanlon uses 'irrational' in its ordinary sense, his claim seems to me unjustified. Scarlet does avoid one kind of irrationality, since Scarlet's preference matches his beliefs about reasons. But in failing to care about this 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 no reason to care about being in agony on future days whose name starts with 'T' and 'U'.

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. [111] 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 the least irrational of these three people.

People are most *clearly* irrational, Scanlon remarks, when they fail to respond to what they themselves acknowledge to be reasons. This remark 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. That is an unacceptable conclusion. 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 our beliefs. Our beliefs are irrational, I would claim, when we are failing to respond to clear and strongly decisive reasons or apparent reasons not to have these beliefs. On Scanlon's view, such beliefs may not be irrational, since we may not be failing to respond to what we ourselves acknowledge to be relevant reasons. Suppose that, though I know that my chance of winning some lottery is only one in a hundred million, I regard this fact as giving me no reason to give up my belief that I shall win. And though I know that everyone else will die, I regard this fact as giving me no reason to give up my belief that I shall live for ever. On Scanlon's 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. On my view, 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 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 committed mass murder, 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.

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 both prefer their future hour of agony to their future minute of slight pain, they do not believe that they have any reasons 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 fully informed, and both now prefer their future agony to their future slight pain, they believe that they have most reason to choose the agony.

On these assumptions, Scarlet and Crimson are still, I believe, irrational. In preferring an hour of agony to a minute of slight pain, these people are failing to respond to a clear and strongly decisive reason. But their beliefs are less 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 some 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 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 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 that they have these reasons. Some Subjectivists may even have these beliefs, and act upon them in their non-academic lives, ignoring or rejecting these beliefs only when they teach or write. (Similar claims apply to those economists who believed, but only when they taught or wrote, that interpersonal comparisons of well-being make no sense.)

I have rejected Scanlon's claim that, when Scarlet and Crimson prefer an hour of agony to a minute of slight pain, these people are not irrational. There may, however, be no disagreement here. I am using 'irrational' in its ordinary sense, to mean 'open to strong criticism of the kind that we also express with words like "foolish", "stupid", and "crazy"'. 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 acknowledge to be reasons, or who are inconsistent in certain other ways. [112]

If Scanlon uses '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 is open to much stronger rational criticism. If this is Scanlon's view, however, it may be misleading for Scanlon to say that only Pink's preference is irrational, since that would suggest that Pink's preference is open to stronger criticism. It is clearer, I believe, and in other ways better to use 'irrational' in its ordinary, wider sense.

14 Other Views about Rationality

We can next briefly consider some other views about the rationality of our desires 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*. Other people mean that this act would be likely to be best for us, or would maximize our expected benefit in an older, temporally neutral sense. We can call these the *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 be likely to be best for us. But these are not substantive claims, which conflict with other claims about what is in other senses rational. These claims are concealed tautologies, whose open forms would tell us only 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 the ordinary senses, which we can also express with words like 'sensible', 'reasonable', 'senseless', and 'foolish'.

According to one substantive view, our desires are rational when our having them has good effects. This view ignores the distinction between some desire itself and our *having* this desire. Whether some desire itself is rational depends, I have claimed, on this desire's intentional object, or the event we want, with the features that we believe this event would have. Whether it is rational for us to *have* some desire may partly depend on some other facts. It may be relevant, for example, how we came to have some desire. If I am in prison, and I know that I shall be painfully killed tomorrow, it might be better for me if I wanted to be painfully killed, since I would then happily look forward to what lies ahead. That might make it rational for me to cause myself to have this desire, if I can. My *having* this desire would then be in one way rational, since I would have rationally caused myself to be in this mental state. But this desire *itself* would still be irrational. This would be a case of rational irrationality. [113]

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 fully rational. Suppose next that, after autonomous deliberation, we want to starve ourselves to death, 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.

On 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 are inconsistent, many writers claim, if they 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, it would not be irrational for me to 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. But, as I have claimed, 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. Our desires might be claimed to be derivatively inconsistent when they depend on such inconsistent 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 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 X to Y, Y to Z, and Z to X. Such preferences are called *intransitive*. If these were mere preferences which did not depend on normative beliefs, it is not clear that they could be claimed to be irrational. This claim is often defended with the remark that, if we had such preferences, we could be exploited. We might be induced to pay some sum of money first to have X rather than Y, then to have Z rather than X, and then to have Y rather than Z. Our money would be wasted, since we would be back with Y, where we started. But this objection appeals, not to the inconsistency of such 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 preferences about three 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. We might believe that X is better than Y, which is better than Z, which is better than X. If such beliefs are inconsistent, as we can often plausibly claim, that might make such preferences derivatively irrational. Though such cases are theoretically very interesting, they do not, I believe, have much practical importance. [114]

The rationality of our desires does not directly 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 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 many 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 of being beliefs about how the spatio-temporal observable universe happens to be. There are some empirical beliefs whose rationality depends only on their content. One example may be Descartes' belief 'I exist.' Perhaps beliefs with this content must be true, in a way that makes these beliefs 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 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 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 denied or ignored by most of the theories that we have been considering.

In rejecting these analogies between beliefs and desires, I am not forgetting that many of our desires depend upon our 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 world, since these beliefs would be true in any possible world.

CHAPTER 5 MORALITY

15 Sidgwick's Dualism

Objective value-based 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 as ends. On some theories, all good ends are in some way good *for* one or more people. Other theories claim some ends to be good in ways that do not depend, or do not depend only, on their contribution to anyone's well-being. Nor is it only outcomes that could be claimed to worth achieving, since some things may be worth doing for their own sake. That might be true, for example, of acts that express respect for people, or some act of loyalty to some dead friend.

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 are doing 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. [115] 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.[116]

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 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 saving 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.

Two reasons are *precisely* comparable when there are precise truths about their relative 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 pick, either of these things. But, on plausible objective theories, most reasons are only imprecisely comparable. That is true even on the simplest forms of hedonism. 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 would not be any precise truth about the relative strength of these reasons. One of these pains could not be, for example, 2.36 times worse than the other. Even in principle, there is no scale on which we could precisely compare the strengths of the reasons that are provided by the intensity or the duration of different pains. And there are only imprecise truths about the relative strength of most other practical reasons. Such truths are least precise when we compare other reasons of different kinds, 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. Such extreme views are hard to defend. Suppose that we are choosing between some architectural plans for some new building. We could rationally choose whatever we had sufficient reasons to choose. If economic and aesthetic reasons

were wholly incomparable, this would imply 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 much very 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.

Perhaps we can 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 would be true if economic and aesthetic reasons were wholly incomparable. 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. [117] 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.' [118]

Sidgwick's Dualism also rests on what Thomas Nagel calls our *duality of standpoints*. [119] We live our lives from our own, personal point of view. But we can think about the world, and our relations to other people, 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. [120] 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. [121]

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. [122] This conclusion seemed to Sidgwick deeply unsatisfactory.

Sidgwick's reasoning seems to be this:

> 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.
>
> When we ask this question from our personal point of view, the answer is that self-interested reasons are supreme.
>
> When we ask this question from an impartial point of view, the answer is that impartial reasons are supreme.
>
> To compare the strength of these two kinds of reason, we would need some third, neutral point of view.
>
> 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 this view is too simple, and should be revised. Sidgwick's claims imply that even the weakest self-interested reason could not be weaker than any impartial reason, however strong. 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. These are unacceptable conclusions. If we acted in such a way, the main reactions of others would rightly be horror and indignation. But, as well as being very wrong, our act would not be rational.

Some people would reject that claim. According to these people, if we knew that our act would best fulfil our present desires, or would be best for us, our act, however horrendous, *would* be rational. Of those who hold such views, however, many use 'rational' in either the desire-based or the egoistic sense. If these

people claimed that such an act would be rational, some of them would mean that, in doing what we knew would best fulfil our desires, we would be doing what we knew would best fulfil our desires. Others would mean that, in doing what we knew would be best for ourselves, we would be doing what we knew would be best for ourselves. Everyone could accept these trivial claims.

When I claim that such an act would not be rational, I am not using 'rational' in either of these senses. I mean that, if we acted in this way, we would be seriously at fault for failing to respond to some decisive reasons. Our reason to save ourselves from mild discomfort would be very much weaker than our reasons to save these many other people from death or agony.

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 or believed would make things go best. As an *Act Consequentialist*, Sidgwick believes that such acts are always morally right. Most of us reject this view. It would often be wrong, we believe, to treat people in certain ways---such as injuring, deceiving, or coercing them---even when such acts would make things go best. And the wrongness of such acts, we might claim, would always or 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. We do 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.[123] 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 certain people, we also have *impartial* reasons to care about everyone's well-being. Sidgwick's claims seem to 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? [124]

This egoistic attitude would be, as Nagel remarks, 'very peculiar.' Unless we are sociopaths, or we have been taught to accept some egoistic or desire-based theory, most of us rightly believe that we would have some reason to want any stranger's pain to be relieved. [125] 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.

We can next reject the Two Viewpoints Argument. This argument assumes that, when we are trying to decide what we have most reason to do, we can rationally ask this question either from our actual personal point of view, or from an imagined impartial point of view. We should reject this assumption. It is often 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 can also claim that, to be able to compare partial and impartial reasons, we don't need some third, neutral point of view. We can compare these two kinds of reason from our actual, personal point of view. And some reasons of either kind can be stronger than, or outweigh, some reasons of the other kind.

Sidgwick's view, however, is partly right, since our partial and impartial reasons are only *very imprecisely* comparable. According to what we can call

> *wide value-based objective views*: When one possible act would make things go in the way that would be impartially best, but some other act would make things go best 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 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. [126]

To illustrate such a view, we can suppose that, in

> *Case One*, I could either save myself from some injury, or save some stranger's life,

and in

> *Case Two*, I could save either my own life or the lives of several 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 could 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. These impartial reasons are also *omnipersonal*, in the sense that they are reasons for everyone. 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 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. [127]

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 of *ourselves* and those other people to whom *we* have close ties. 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 simple, and too egoistic. We could often rationally give to some stranger's well-being as much weight, or more weight, than we give to our own well-being. Suppose that, in Nagel's imagined hospital ward, there is only one dose of morphine, which 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 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.

It may seem that, when the stakes are as high as this, we are 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. [128]

There is, however, a strong objection to this view. I have accepted Sidgwick's claim that 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. [129]

This objection can, I think, be answered. Even when the stakes are very high, we are not rationally required, I believe, to give priority to our own well-being. We might appeal instead to a revised version of Sidgwick's view. According to what we can call

> *Pure 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 either rationally 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 *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 *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, Pure 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 Pure Dualists could claim, I could not rationally choose to save this stranger. I ought morally to give priority to my child or friend. I would have various other strong non-moral reasons to act in this way. 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, and I could not 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. I would be rationally permitted to save this stranger only in a version of this case in which I had no such reason-giving and obligation-involving ties to certain other people.

In that other version of this case, I am inclined to believe that I could rationally give up my life to save this stranger. And, if I had no such ties to other people, and merely had to choose between benefiting myself or this stranger, I am inclined to accept Sidgwick's assumption that I could rationally choose to make my decision from an impartial point of view. In such cases, we may be rationally permitted simply to ignore our reasons to be specially concerned about our own well-being. But we need not here decide whether these beliefs are true, or whether, if I gave up my life to save this stranger, my act, though morally admirable, would be less than fully rational.

16 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 whatever 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'. [130]

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 either Moral Rationalism or Rational Egoism, or revise both 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'. [131] 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. [132]

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. [133]

These magnificently sombre claims are, however, overstatements. Sidgwick believed that in most cases duty and self-interest do not conflict, and in such cases Sidgwick's view implies that we have most reason to do our duty. In such a world, the cosmos of duty would not be a chaos. Nor would our system of beliefs about the reasonableness of 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, 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 that 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. [134] 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 deciding how to act, we can ask two questions:

Q1: What ought I morally to do?

Q2: What do I have most reason to do?

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 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 always involves normative reasons or apparent reasons. On the *rule-involving* conception, normativity involves requirements of some kind, or rules that distinguish between what is *correct* and *incorrect*. 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 two conceptions of normativity are very different. On the rule-involving conception, we can create new normative truths merely by proposing and accepting certain rules. 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. And we cannot create such reasons merely by introducing some rule, or requirement.

There is a deeper difference. When there are such rules or requirements, we may have reasons to follow them. But these reasons are often provided, not by the mere existence of these rules, but by certain other facts, some of which depend on other people's acceptance of these rules. If we use words with their correct spelling and meaning, that may make us seem better educated, and help us to be understood. If we drive on the correct side of the road, we shall be less likely to crash. 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.

On some views, it is we who create moral 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.

Moral requirements sometimes 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 believe that, in such cases, 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.

There are also rational requirements. For example, if we believe that we have decisive reasons to act in some way, we may be rationally required either to give up this belief, or to act in this way. Some of these requirements may have little or no importance in the reason-implying sense. Following these requirements may 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. [135]

Return now to our two questions:

Q1: What ought I morally to do?

Q2: What do I have most reason 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. The same applies to morality. If we had no reasons to care about morality, or to avoid acting wrongly, morality would 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 matters, or 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, 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, though we cannot *justify* the reason-involving criterion of importance, we can *use* this criterion. We can truly claim that some things matter in this sense, and that others don't. [136] And it would have great importance if morality did not matter in this sense, 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 most fundamental question is not what we ought morally to do, but what we have most reason to do.

In the rest of this book, I shall 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.

CHAPTER 6 MORAL CONCEPTS

17 Acting in Ignorance or with False Beliefs

Before we ask which acts are wrong, it will help to ask 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 of 'wrong', such as the senses 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, in such cases, everyone always uses 'wrong' in the same sense, which we can call the *ordinary* sense. In many cases, however, people don't know all the relevant facts, and they must act in ignorance, or with false beliefs. In such cases, I believe, different people use 'wrong' in at least partly different senses. To help us to think about such cases, we can use the ordinary sense of 'wrong' to define three other senses. 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. [137]

Acts are in these senses *right* 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 people don't know all of the morally relevant facts, it is enough to ask whether these people's acts are wrong in the ordinary sense. Other writers claim that one of the senses I have just defined *is* the ordinary sense. [138] 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 one 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 writers, when we consider such cases, it is enough to use 'right' and 'wrong' in their belief-relative senses. On this view, it is enough to claim that I act rightly in *Case One*, because I am doing what I believe will save your life, and that I act wrongly in *Case Two*, because I am doing what I believe will kill you.

It is *not* enough to make these claims. We should also claim that, in *Case One*, I act wrongly in the fact-relative and evidence-relative senses, since my act kills you, as on the available evidence it 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, given my false belief about how I could save your life, I ought to do what would almost certainly kill you. We should similarly claim that, in *Case Two*, I act rightly in the fact-relative and evidence-relative senses, since my act saves your life, as it was almost certain to do. This is 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 in your medical condition, 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 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 it would.

According to some writers, 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. It is not enough to say that I acted rightly, in *Case Four*, by doing what was almost certain to save your life. We should also claim that I have murdered you, thereby acting wrongly in the belief-relative and fact-relative senses. Such intentional killings are blameworthy, giving the murderer reasons to feel guilt and remorse, and giving others reasons for indignation.

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 have 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 have conscientiously read these journals, and my fault may be only that I have 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. In this version of this case, I would be much less blameworthy.

According to some other writers, 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 have 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 was doing what I justifiably believed 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 have 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 either of the other senses. And we and others may needlessly disagree. 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 consider the relative importance of these senses. 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, that 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 it would be wrong for me to kill you. So (A) plausibly implies that these acts are 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, that 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 is blameworthy. But this claim 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 is 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 is *not* blameworthy, since this act isn'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 this fact was in no way under my control. Since (B) implies that my act is blameworthy in *Case Five* but not in *Case Seven*, (B) 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 imply (C). Some of these children believe, for example, that well-intentioned acts are blameworthy whenever 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. But when we fully understand blameworthiness 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, that 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. [139] 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. 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. It is fairly plausible to regard murder 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, in the sense of being better supported by the available evidence. The most 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. And that is not a difference in blameworthiness. [140]

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 be morally less blameworthy merely because we are either less successful in hitting our intended target, or are epistemically irrational.

This Kantian view is, I believe, true. Though it can seem plausible to claim that murder is more blameworthy than attempted murder, this claim's plausibility can be 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, when people believe that they are acting wrongly, that is enough to make their act wrong in the ordinary sense; but when people believe that their act is right, that is *not* enough to make this act right. Even if we reject this view, it seems clear that

> (F) in most cases, when someone acts in a way that this person believes to be wrong, that makes this act blameworthy.

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. [141] Some such acts may not be blameworthy.

An act's blameworthiness, I conclude, mostly depends on whether this act is wrong in the belief-relative and moral-belief-relative senses.

We can now ask which are the most important senses of 'ought', 'right', and 'wrong' when we are trying to decide what we shall 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 the cases that we have just been discussing, (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 justifiably 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 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 our act depends on the goodness of its effects or possible effects, we ought to act, or try to act, in the ways whose outcome would be *expectably-best*. [142]

In calling some act's outcome 'expectably-best', we do *not* mean that we expect this act to produce the best outcome. If we closed Gate 3, it would be certain that this act would *not* produce the best outcome, as our act might do if 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 act would save the greatest *expectable* number of lives. These numbers are the actual numbers of lives that each act might save, multiplied by the chances that these acts would save these numbers. 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 we 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 the goodness of these possible effects multiplied by the chance that this act would have these effects. [143]

Expectabilism applies to all cases, including those in which we know which act would in fact make things go best. This act would also be the act whose outcome would be expectably-best.

I have 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 often ought 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 my example, if we could easily find out in which shaft all of the hundred miners are, this discovery would make things go expectably-best, since we would then know how to save all these people, thereby saving ten more people than we would have saved if we had closed Gate 3.

There is another reason why, when we are trying to decide what we ought 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. As Sidgwick points out, similar claims apply to our moral beliefs. Though we know that these beliefs may be mistaken, we cannot try to do what is really right rather than what we now believe to be right. [144]

I claimed earlier that, when we ask whether some act was blameworthy, what is most important is whether this act was wrong in the belief-relative sense. 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. What is important is what we ought to do in the belief-relative and evidence-relative senses. 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 of our acts would be right or wrong in the fact-relative senses.

There is, however, one way in which these fact-relative senses have great importance, and can even be claimed to be fundamental. As well as asking, in some actual case, which acts 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. 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 the ways in which what we ought to do depends on our beliefs, and about how we ought to respond to risks, and to uncertainty. 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. There are also many cases in which these distinctions do not matter. So when making other claims I shall often use 'best' to mean 'best or expectably-best'.

Similar claims apply to theories about reasons. We ought, I have argued, to reject all subjective theories, and to accept some objective theory. Though theories of these two kinds deeply disagree, this disagreement is not about what reasons we would have when we don't know all of the relevant facts.

On both kinds of theory, we can draw some similar distinctions. We can ask both what we *should* or *ought rationally* to do, and what we *should* or *ought* to do in what I have called the *decisive-reason-implying* senses. It might seem that each of these two

oughts could also be used in fact-relative, belief-relative, and evidence-relative senses. That would give us six further senses. Fortunately, things are simpler, since these senses of 'should' and 'ought' already draw one of these distinctions. Some possible act is

> what we *ought* to do in the *decisive-reason-implying* sense just when the relevant facts give us decisive reasons to act in this way.

There is no *belief*-relative version of this *reason*-implying sense of 'ought', since it is only *facts* that give us *reasons*. Some possible act is

> what we *ought rationally* to do just when we have beliefs whose truth would give us decisive reasons to act in this way.

There is no *fact*-relative version of 'ought rationally', since what is *rational* depends on our *beliefs.*

Here is another way to make these points. These two common senses of 'ought' can be regarded as the fact-relative and belief-relative versions of a single underlying sense of 'ought', which we might express with the phrase 'ought practically'. Return to the case in which, while walking through some desert, you have angered a poisonous snake. You believe that, to save your life, you must run away. In fact you must stand still, since such snakes attack only moving targets. We can claim that

> what you *ought practically* to do in the *belief-relative* sense is to run away, because this act is what you would have decisive reasons to do if your beliefs were true.

I express this concept with the phrase 'ought rationally'. We can also claim that

> what you *ought practically* to do in the *fact-relative* sense is to stand still, because this act is what the facts give you decisive reasons to do.

We express this concept when we use the words 'should' and 'ought' in their decisive-reason-implying senses.

18 Other Kinds of Wrongness

We should distinguish, I have just claimed, between several moral senses of 'ought', 'right', and 'wrong'. These senses can all be defined 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 the same or sufficiently similar meanings. This concept refers to the *property* of wrongness. If there are different senses of 'wrong', these 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, I believe, 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, this group of concepts all have a common element which we cannot helpfully explain. To express this indefinable version of the concept *wrong*, I shall use the phrase '*mustn't-be-done*'. [145]

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 suggest thinner definitions. In what we can call the *decisive-reason* senses,

'ought morally' means 'what we have decisive reasons to do',

and

'wrong' means 'what we have decisive reasons *not* to do'.

These senses are misleading. 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 senses, they would claim that

(H) we ought morally to do whatever would be best for ourselves.

But Rational Egoism is not a moral view, but 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. [146]

In what we can call the *decisive-moral-reason* senses,

'ought morally' 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 regarded as 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,

> 'ought morally' means 'what we have morally decisive reasons to do',

and

> 'wrong' means 'what we have such reasons not to do'.

Though they may seem very similar, these senses differ in at least two ways from the *decisive-moral-reason* senses. 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 express 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 *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.

These definitions, we can add, tell us little. If we are trying to decide whether some act would be wrong, it does not help much to ask whether we have morally decisive reasons not to act in this way.

Some people seem to use

> 'ought morally' 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, we have seen, 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 'ought' and 'best' in these impartial-reason-implying senses, it would mean

> (I) 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.

This view we can call *Impartial-Reason Act Consequentialism*. To express this sense of 'ought', we might 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.[147]

When Sidgwick claims that he *ought* not to prefer his own lesser good, he does not seem to mean that such a preference or act would be blameworthy, or unjustifiable to others, or would be an act that would give him reasons for remorse and give others reasons for indignation. Sidgwick seems to mean that, when assessed from an impartial point of view, his reason to give himself some lesser good is weaker than 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 requirement that we promote the well-being of others, it is closer to morality. That makes it, in some ways, a more serious rival, since Impartial-Reason 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'. [148]

(I) 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 the impartial-reason-implying senses. According to subjective theories about reasons, and Rational Egoism, there are no such acts or outcomes, since there are no acts or outcomes that we all have impartial reasons to want or prefer. And, even if (I) were a tautology, Impartial-Reason Consequentialists could make other, substantive claims. If they are hedonistic utilitarians, for example, these consequentialists might claim

> (J) What we ought-impartially to do is whatever would produce the greatest sum of happiness minus suffering. [149]

These people might 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'. That is how this form of consequentialism might be an external rival to morality. [150]

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 moral sense in which everyone uses the word 'wrong'. 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 which 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. The morally-decisive-reason sense of 'wrong' might be claimed to use '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. [151]

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 others hold 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 not, I believe, strong. 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 claim that some act is wrong and you claim that it isn't, we would be disagreeing about whether this act is blameworthy. And when people use 'wrong' in such different senses, that may *increase* their disagreements. In the case just imagined, if we understand each other's use of 'wrong', you may believe that no acts are in my sense wrong, since you believe that no acts are blameworthy because they are forbidden by God. I may 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 that I have just described, if we 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 reactive-attitude or blameworthiness senses, most of us would reject this claim. We would believe that, if I saved myself rather than these two strangers, I would have no reason to feel remorse, and these strangers and others would have no reason to be indignant. But Sidgwick may mean that I would have stronger impartial reasons to save the two strangers. Whatever our moral beliefs, we could all accept that claim.

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

> (K) we ought to do what would make things go best,

and others claim that

> (L) we ought to do what would make things go expectably-best.

If (K) uses 'ought' in the fact-relative sense, and (L) 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 (L) 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 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 apologize, and make amends. When we claim that we ought to do something, we should often be taken to be claiming 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 on what we mean by 'wrong', and on the kind of wrongness to which our use of 'wrong' might refer. 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, of which two examples might be the facts that this act would be dishonest or would cause pointless suffering. On this view, the fact that

> (M) some act is wrong

would be the higher-order fact that

> (N) there are certain other facts that give us decisive moral reasons not to act in this way.

This higher-order fact about our reasons would not give us any independent reason not to act in this way. Though we might claim that an act's wrongness always gave us a reason not to do it, this reason would be wholly derivative, since its normative force would derive entirely from these other reason-giving facts. Since this derivative reason would add nothing, it could be ignored. So, if we used 'wrong' only in this decisive-moral-reason sense, we could conclude that

> (O) when some act would be wrong, this fact would not give us any 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 decide that we had such reasons, we could then ask whether these were *moral* reasons. But this would not be a question about what we ought to do, or had reasons to do. This question would be merely theoretical, like the question of which are the kinds of reason that can best be called economic, or aesthetic. [152]

Many people assume that an act's wrongness gives us strong or even decisive 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 senses, we *can* claim that the wrongness of these acts gives us independent, additional reasons not to act in these ways.

We could make that claim, for example, if 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.[153] We are claiming that this act has the highly distinctive substantive property of being something that *mustn't-be-done*. And, if some acts do have this property, we could plausibly claim that, when some act mustn't-be-done, that gives us a strong reason not to do it. This is one of the senses of 'wrong' with which it seems most plausible to claim that

> (P) when some act would be wrong, this fact always gives us a decisive reason not to do it.

(P) would be just as plausible if we use '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 would give us reasons for remorse and give others reasons for indignation, these facts would all give us 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.

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, and 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 Appendix A, I shall say little about these *meta-ethical* questions. Such questions will be easier to answer when we have made more progress in our moral thinking, and in our understanding of practical and epistemic reasons and requirements. As Rawls and Nagel claim, our moral theories 'are primitive, and have grave defects', and 'ethical theory. . . is in its infancy.'[154]

Rather than proposing a new moral theory, I shall try to learn from some existing theories and moral ideas, and try 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'. [155] I shall ask whether he succeeds.

PART TWO

CHAPTER 7 POSSIBLE CONSENT

19 Coercion and Deception

According to Kant's best-loved moral principle, often called

> *the Formula of Humanity*: We must treat all rational beings, or persons, never merely as a means, but always as ends. [156]

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.[157]

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. [158]

Onora O'Neill similarly writes:

> if we coerce or deceive others, their dissent, and so their genuine consent, is in principle ruled out. [159]

Korsgaard concludes:

> According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrong-doing to others. [160]

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, they could not be deceived. So we cannot possibly deceive people with their consent. This might be why, unlike surgery, deception is always wrong. [161]

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 claims apply to coercion. People could not possibly consent to being coerced, Korsgaard might say, because if people gave consent they would not be being coerced. 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, that these are acts whose

nature makes consent impossible.

20 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. [162]

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'. [163] 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 stronger sense. People can give or refuse consent in this *act-affecting* sense if they have what Korsgaard calls 'power over the proceedings', because we shall treat them 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. [164]

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 mistakenly implies that, in all these cases, whatever we did would be wrong.

We might revise this principle by restricting it to cases in which our acts would have significant effects on only one person, who is someone other than ourselves. Perhaps we ought to let any such person choose how we treat her. But these are not the only cases that Kant has in mind.

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 have *decisive reasons* not to give consent. Kant, I suggest, means

> (E) It is wrong to treat people in any way to which they could not *rationally* consent.

We can call (E) the *Principle of Possible Rational Consent*, or---as I shall say for short---the *Consent Principle*. [165]

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. [166] Consider next Kant's remark 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*. [167]

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', or share, the end or aim of our act. 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 could 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 life be saved, you could not rationally consent to my achieving this aim by stealing from you the medicine that you need to save the lives of yourself and your children. 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 counting the number of blades of grass in your lawn, you could rationally consent to my acting in this pointless but harmless way. 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.

21 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 they could will it to be true that the agent's maxim is a universal law. [168]

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 maxims that we could not will to be universal laws.

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, as your doctor, I

ask you whether you consent to my giving you some medical treatment. For it to be rational for you to consent, you might 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 which is the maxim or policy on which I would be acting, or whether you could will that my maxim 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'. [169] 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 take them to contain no other ideas. [170]

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. [171] 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. [172]

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 make no difference to how they were treated. If that were true, it might be rational for these people not to protest, even if they are treated in ways that are very bad for them, and very wrong. Our question should also be about *informed* consent. When people do not know what effects some act would or 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. So 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 act would not causally affect these people. Two examples would be knowingly 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 making these various claims, 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 be successful, 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 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. Consider, for example,

> *Earthquake*: Two people, *Blue* and *Grey,* are trapped in slowly collapsing wreckage. I am a rescuer, who could prevent this wreckage from either killing Blue or destroying Grey's leg.

If these are the morally relevant facts, it is clear that I ought to save Blue's life. But we can 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 or aims. According to both desire-based 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 Blue's life. [173] Similar claims apply to countless other cases. There are countless right acts to which, according to both desire-based 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 theories, when one possible choice would make things go in the way that would be impartially best, but some other choice would make things go best 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 Blue's life. Grey could rationally regard Blue's well-being as mattering as much as hers, and Blue's loss in dying would be much greater than Grey's loss in losing her leg.

Could *Blue* rationally choose that I save Grey's leg? We could rationally choose to accept some losses, I believe, if we could thereby save others from somewhat smaller losses. But, in this example, there is too great a difference between Blue's loss and Grey's. Blue would not have sufficient reasons to give up her life so that I could save Grey's leg. [174] So the Consent Principle rightly requires me to save Blue's life, since this is the only act to which both Grey and Blue could rationally consent.

Suppose next that, in

> *Lifeboat*, a single person, *White*, is stranded on one rock, and five people are stranded on another. Before the rising tide drowns all these people, I could use a lifeboat to save either White or the five.

These people, we can suppose, are all strangers both to me and to each other, they would all lose as much in dying, and they do not differ in any other morally relevant way. We should make similar assumptions, except when I say otherwise, in all of my later imagined cases. Though some people would believe that I ought to give White some chance of being saved---which might be a chance of one in six or even one in two---most of us would believe, more plausibly, that I ought to save the other five people.

If White could choose how I shall act, she would have sufficient reasons, I believe, to make either choice. White could rationally choose that I save her life, but she could also rationally choose

instead that I save the five. Could the five rationally consent to my saving White 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 my act in the act-affecting sense, thereby choosing how I shall act, this person could rationally choose that I save White rather than the five. The answer is clearly No. If you were one of the five, you would not have sufficient reasons to choose that I save White rather than saving *both* you *and* four other people. You would have both strong personal and strong impartial reasons not to make this choice. On these assumptions, the Consent Principle rightly implies that I ought to save the five, since this is the only act to which both White 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, or deny us any significant benefit. There might be cases, for example, in which we could not rationally consent to being deceived, or coerced, or having decisions made paternalistically on our behalf. We can have strong reasons to want to decide how we live our lives, even when other 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 to them. Blue, I have claimed, could not rationally consent to my saving Grey's leg rather than Blue's life, given the fact that Blue's loss would be so much greater than Grey's. These facts can also be plausibly claimed to make this act wrong. Similar claims apply to the other facts that I have just mentioned. Whenever such facts give some people decisive reasons to refuse consent to certain acts, these facts would also provide moral objections to these acts.

According to the Consent Principle, these moral objections are decisive, since it is wrong to act in any way 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.

If 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 White or the five will be denied the benefit of being saved from death. In this case, I have claimed, (G) is true. White would have sufficient reasons to consent to my failing to save her life so that, before the tide covers both rocks, I could save the five. If White would have such reasons, as I believe, 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 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, I believe, 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.

22 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 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. When we must make some decision on someone else's behalf, it may be enough to claim that we ought to try to decide, and to do, what would be best for this person. It may not be worth adding that it would be wrong for us to act in any way to which this person would not have sufficient reasons to 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*, in the sense that these principles would require 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, I believe, 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 plausibly 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 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.

23 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 was true, that could 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. We can reply that, 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 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. The Consent Principle condemns all such acts. So 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, for example, I ought to save Blue'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 they either do or would refuse consent.

If (J) were true, the Consent Principle would imply the Veto Principle. That would make the Consent Principle clearly false.[175]

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. 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 consent. There are also many acts, however, to which we *could* rationally give such irreversible consent. Before the discovery of anaesthetics, many people rationally gave such consent in advance to painful surgery, permitting the surgeons and their

assistants to use force, if necessary, if the pain later led these people to change their mind.

Such consent was rational, it might be claimed, only because these people knew that the pain of the surgery would be a distorting influence, which might cause them irrationally to change their minds. But we can often rationally give irreversible consent even when we know that we might later change our mind in some way that would not be irrational. 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.

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 Blue's life rather than Grey's leg, it would be dangerous for me to stop. Suppose next that, since Grey knows all of the relevant facts, she 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 Blue'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 Blue'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 assumption, these people would have had sufficient reasons to give their irreversible consent.

Since people could often rationally give such irreversible consent to being later treated in some way without their later actual consent, 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 to such acts. 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. The question whether people *would in fact* consent to some act is, we should note, quite different from the question whether these people *could rationally* 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 relevant facts, or who are too young, or seriously mentally ill, or 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. [176]

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 that 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 other cases, people cannot give valid consent at the time, and have not given or refused consent in advance. We may believe that we should try to treat such 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

> *Painful 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. We ask Green to give irreversible consent to this operation, permitting us to use force, if necessary, if the pain leads her to change her mind.

If Green gave such consent, and the pain did lead her to change her mind, we would be justified in using force to complete this surgery. Since great pain is a seriously distorting factor, Green's withdrawal of consent would not be valid, and could be ignored. Before the discovery of anaesthetics, as I have said, many people consented in advance to having surgery imposed on them, without their consent at the time. The Rights Principle should permit such acts.

Suppose next that, even before this operation starts, 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 may 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 hard 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*, White decides to kill herself. We have strong reasons to believe that, if we forcibly prevented White's act, White's depression would soon lift, and the rest of her life would go well.

Many of us would believe that we could justifiably override White's decision, and use force to prevent her from killing herself. If we accept the Rights Principle, we might claim that severe depression counts as a sufficiently distorting factor, so that White'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 depressed may know the relevant facts, nor are they clearly incapable of making rational decisions. It would be more plausible to claim that, though White's depression does not make her refusal of consent invalid, it makes her less able to make rational decisions, so that White's refusal might be morally outweighed by her decisions at other times. For example, if White has frequent temporary depressions, she may have consented in advance to our preventing her from killing herself while she is depressed. That may be enough to justify our act, though we would here be overruling White's *valid* refusal of consent at the time. And, given the irreversibility of suicide, such acts might be justified even without such earlier consent.

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.

If Brown knew the relevant facts, she would know that blood-transfusions are *not* wrong, and she could then have rationally given irreversible consent 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. Similar claims apply to the Rights Principle, as described above. That principle appeals to the consent that people do give, or would give, if they knew the relevant facts. We may know that, if Brown knew that blood-transfusions were not wrong, she would in fact give her consent. But this may not be enough to justify our saving Brown's life in this way. When people have certain kinds of false belief, such as certain moral or religious beliefs, these people may have rights not to be treated in certain ways, such as having life-saving surgery imposed on them, when it is some such belief that leads them to refuse their consent. The Rights Principle could be revised, so that people with such false beliefs would be regarded as validly refusing consent.

[More to be added here about how the Rights Principle is related to the Consent Principle.]

24 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. As I have said, that may be true of all surgery that is done with anaesthetics. Another, 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 severe 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 give me such consent in advance. The Consent Principle would not then condemn my acts. But we can plausibly claim that, without your *actual* consent either at the time or in advance, these

acts would be wrong.

There may also be some ways of treating people that would be wrong even if these people have actually and rationally given their valid consent. Many people have that view, for example, about *voluntary euthanasia*: or killing someone, as this person asks us to do, for her own good. And some kinds of act 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 the Consent Principle, we must 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 our beliefs about which acts are wrong, our moral reasoning would be circular, or question-begging. Such reasoning could not support our beliefs about the wrongness of these acts.

This objection is, in part, correct. It could not be true both that

> (M) some act is wrong because someone could not rationally consent to it,

and that

> (N) this person could not rationally consent to this act because it is 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, Blue has such a reason to refuse consent to my saving Grey's leg rather than Blue's life. Blue could not rationally consent to this act, not because this act would be wrong, but because Blue'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 moral beliefs.

Return next to the cases in which, when applying the Consent Principle, we ought to appeal to our deontic beliefs. Suppose that, in a variant of *Earthquake*, which we can call

> *Means,* Blue and Grey are trapped in slowly collapsing wreckage. Though Blue's life is threatened, Grey is in no danger. I could save Blue'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 Blue'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 her 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, I believe, be relevantly different from *Earthquake*. In both *Earthquake* and *Means*, either Blue will die or Grey will lose her leg. These cases differ only in how the saving of Blue'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 in advance her irreversible consent to my later saving Blue's life, even though Grey would then lose her leg. And in both cases, since Blue's loss would be so much greater than Grey's, Blue 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 Blue's life by destroying Grey's leg, since that is the only act to which both Blue 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 Blue's life by destroying Grey's leg, this act's wrongness would give Blue 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 conclusion. We are discussing possible consent in the act-affecting sense. For Blue to be able to give or refuse such consent, I must have given Blue the power to choose how I shall act. If Blue 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 Blue 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 Blue to make this choice. Perhaps Blue could rationally choose that I act wrongly, since that choice would save Blue's life. But Blue would also have sufficient reasons to choose instead not to be partly responsible for this wrong act. Since Blue could rationally consent to my failing to save her life by destroying Grey's leg, the Consent Principle would not mistakenly require this act. [177]

This principle may seem to fail in a lesser way, by mistakenly permitting this wrong act. But the Consent Principle does not claim to be the only principle we need. Acts can be wrong for other reasons. So, when this principle does not condemn this way of saving Blue's life, it does not thereby imply that this act is morally permitted.

Similar claims apply to other cases. We are considering cases in which some act of ours would be wrong, not even in part because some people 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 could 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 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.

Some people believe that, even in cases like *Fire*, such voluntary euthanasia is wrong. If it would be wrong for me to give Black this better death, would this act's wrongness give Black a sufficient reason to consent to my failing to act in this way? Some of these people might answer No. These people might agree that, in *Means*, Blue could rationally consent to my failing to save her life by destroying Grey's leg. But Blue's reason to give such consent is provided by the fact that I could save Blue's life only by wrongly injuring someone else. If I killed Black in *Fire* 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.

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, as before, if it would be wrong for me to kill Black without her actual consent, this act's wrongness would have 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 rather than yours. 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 morally outweighs 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 in these and 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. We could explain why this principle should not be applied to cases of this kind. When we apply this principle, we appeal to a thought-experiment, by asking whether other people 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.

If there are cases of this kind, however, we could not achieve Kant's ideal. There would be some morally required acts to which some people could not rationally consent. So we can ask

> Q1: Could we have a duty to choose 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 would be able to act in a way to which you could rationally consent. If it would be wrong for me to save your child rather than mine, this act's wrongness would give you a sufficient reason to consent to my doing my duty, by saving my child.

For another kind of objection, suppose 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 rather than you. 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 you rather than Grey. 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, that would give you both sufficient reasons to consent to this act.

These claims do not show that we could achieve 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. Such objections also show, however, that Kant's ideal makes a significant, substantive claim. Of the plausible or widely accepted views about morality and rationality, it is worth asking which are the views that are compatible with Kant's ideal.

25 Extreme Demands

Suppose next that, in

> *Self*, it is I who am trapped with Blue in slowly collapsing wreckage. I could save either Blue's life or my leg.

On some views, this case is morally just like *Earthquake*. I ought to save Blue's life rather than my leg, since Blue's loss would be much greater than mine. Most of us have a different view. On this view, though it would be wrong for me to save some other stranger's leg rather than Blue'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 Blue had the power to give or refuse consent to my act in the act-affecting sense, thereby choosing how I would act, could Blue rationally choose that I save my leg rather than Blue's life?

The answer may seem to be No. It may seem that, just as Blue could not rationally consent to some stranger's saving my leg rather than Blue's life, Blue could not rationally consent to *my* acting in this way. But this does not follow. 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 Blue from a week of pain, or save some other stranger from only one day of similar pain. There is no other relevant difference between Blue and this other stranger. On that assumption, I would have no reason to give any priority to the well-being of this other stranger. And Blue could not rationally consent to my choosing, *for no reason,* to help the other stranger rather than saving Blue from her much greater burden. That choice would treat Blue as if she were inferior, or didn't even exist.

Suppose next that, rather than saving Blue 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. Though Blue could not rationally consent to my choosing, for no reason, to save some *stranger* from a day of pain rather than saving Blue from a week of pain, Blue may have sufficient reasons to consent to my saving *myself* from this much smaller burden. *This* act would not treat Blue as if she were inferior, or didn't even exist.

In *Self*, however, the stakes are much higher. Blue may not have sufficient reasons to consent to my saving my leg rather than Blue'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 Blue's life? Perhaps not. There may be a difference here between permissibility and wrongness. If I could save Blue's life only by acting wrongly, as we have supposed to be true in *Means*, this act's wrongness, I have claimed, would give Blue a sufficient reason to consent to my failing to save her life. In *Self*, however, I could save Blue's life without acting wrongly. And even if I would be morally permitted to save my leg rather than Blue's life, this act's permissibility may not give Blue a sufficient reason to consent to my failing to save her life.

If this act's permissibility would *not* give Blue such a reason, Blue could not rationally consent to my failing to save her life, so the Consent Principle would require me to save Blue'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 $300 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. [178] 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. [179]

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. [180]

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 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. [181] But we could start by claiming that, in *Self*, I would be permitted to save my leg rather than Blue'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 Blue'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. [182] Some acts are wrong even though everyone could rationally consent to them. Since we need at least one other principle, we can now turn to another part of Kant's Formula of Humanity.

CHAPTER 8 MERELY AS A MEANS

26 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 Kant's

> *Mere Means Principle*: It is wrong to treat anyone merely as a means. [183]

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. Here are two rough definitions. 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,

and

> *merely* as a means if we also 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 view, we must never treat people merely as a means, this requirement would be too weak, and too easily met. 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 slave-owners surely failed to meet Kant's requirement. [184]

This objection shows, I believe, not that we ought to revise this definition, but that we ought to revise Kant's requirement. For a similar example, consider Kant's claim that

> (A) it is wrong for the rich to give nothing to the poor. [185]

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 requirement, 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 treats his slaves, even though he lets 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 as a means. 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. [186] 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 that concern, though relevant, would not be sufficiently important. 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. [187]

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, in most cases, 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. [188]

We could similarly claim that, when this Egoist *keeps* some promise to someone whose help he needs, he wants to make use of another human being 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 view implies that, in keeping his promise and saving this child's life, this man acts wrongly. That is clearly false. [189]

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 he acts wrongly. Our claim should be only that, given this man's self-interested motives, his acts do not have what Kant calls *moral worth*. [190]

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. [191]

In moving to this principle, we would be giving up the view that, if we treat someone merely as a means, or 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. [192]

Return now to

> *Lifeboat*: White is stranded on one rock, and five people are stranded on another. Before the rising tide drowns all these people, I could use a lifeboat to save either White 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, I 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 I know, White is in this other tunnel.
>
> *Bridge*: The train is headed for the five, but there is no other track and tunnel. White is on a bridge above the track. My only way to save the five would be to open, by remote control, the trap-door on which White is standing, so that she would fall in front of the train, thereby triggering its automatic brake.

In all three cases, if I save the five, White would die. But White's death would be differently causally related to my saving of the five. In *Lifeboat*, I would let White die because, in the time available, I could not save both White and the five. In *Tunnel*, I would save the five by redirecting the train with the foreseen side-effect of thereby killing White. In *Bridge*, I would kill White as a means of saving the five. These six people, we should suppose, are all of about the same age, none of them is responsible for the threats to their lives, nor are there any other morally relevant differences between them.

It might be claimed that, in *Bridge*, I would not really be *killing* White as a means of saving the five. I would be merely using White's body as a means of stopping the train, and I would be delighted if White 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 kills or injures this person, as we knew that our act was certain or likely to do, 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*, I either may or ought to save the five. Some people would believe that, in both *Tunnel* and *Bridge*, it would be wrong for me 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. Some 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 I *would* be acting wrongly if, in *Bridge*, I killed White 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, White could rationally consent to my saving the five rather than her. [193] If the choice were White's, she would have sufficient reasons to save her own life, but she would also have sufficient reasons to save the five rather than herself. Since White could rationally consent to my saving the five, the Consent Principle does not condemn this act.

Similar claims apply to *Tunnel*. As before, if the choice were White's, she would have sufficient reasons to save either herself or the five. It would make no relevant difference that she would here be saving the five by redirecting the train so that it would kill her instead. This way of dying, we can suppose, would be no worse for White. Since White could rationally save the five and kill herself by redirecting the train, she could also rationally consent to *my* redirecting the train. So the Consent Principle does not condemn this act.

Similar claims apply to *Bridge*, in which I could save the five only by killing White. If the choice were White's, she would have sufficient reasons to jump in front of the train, so that it would kill her rather than the five. And compared with being killed as a side-effect of my saving the five, in *Tunnel*, it would be no worse for White to be killed as my means of saving the five, in *Bridge*. White might even have more reason to prefer to be killed as a means, since her death would then at least do some good. Since White could rationally kill herself as a means of saving the five, she could also rationally consent to my treating her in this way.

It might be objected that, since it would be wrong for me to kill White as a means, White could not rationally consent to this act. But, if White consented to my killing her as a means, this act would not be wrong. So, even if this act would be wrong without White's consent, that would not give White any reason to refuse consent.

Suppose next that I accept the Consent Principle, and I always act upon it, so that this principle governs my acts. I might then think:

> According to this principle, it is wrong to treat people in any way to which they could not rationally consent.
>
> White could rationally consent to my killing her as a means of saving the five.

Therefore

> Even if White 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 my treatment of White would be governed by the Consent Principle, I would neither be treating White 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 me to kill White, without her consent, as a means of saving the five. But that is not implied by these Kantian principles.

27 *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*. 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 hurting me as a means of getting this information, but she isn't *using* me, or treating *me* as a

means. This distinction can be easily missed, since it is sometimes drawn only by emphasizing different words. If my doctor later gave me some painful medical treatment, she would be *treating* me merely as a means, but she wouldn't be treating *me* merely as a means.

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 my act, in *Bridge*, if I killed White as a means of saving the five. I would not be limiting the harm that I imposed on White. And I would have acted in the very same way even if I had regarded White as a mere means. That may seem enough to justify the charge that, in acting in this way, I would be treating White merely as a means. On this suggestion,

> (C) we treat someone merely as a means if
>
> > (1) we harm this person, without 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 I use remote control to cause White to fall onto the track, so that White's body would stop the runaway train. My aim is to ensure that the five will be saved. I also try, however, to save White's life by running to the track, so that I can jump in front of the train before it reaches White. If my attempt succeeds, I would not be treating White merely as a means, since I would be killing myself for White's sake. It would make no relevant difference, I believe, if I failed to reach the track in time. Nor would it make such a difference if, though I would have sacrificed my life to avoid killing White, this was never possible. In both versions of *Bridge*, my act may be wrong. And, if it is, what makes it wrong may be the fact that I would be *killing* White *as a means* of saving the five. But I would not be treating *White herself* 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. . . [194]

Korsgaard similarly writes:

> Coercion and deception are the two ways of using others as mere means. [195]

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.

On Kant's view, Korsgaard elsewhere 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. . . [196]

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'. [197] But, when poor people feed themselves with the scraps that others throw away, they do not treat these other people as a mere means.

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. [198]

Deception and coercion are not, however, appropriate ways of

treating things, since neither is even possible.

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 risky, 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. [199] 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. [200] That would make it trivial to claim that these chosen principles are true. [201]

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, [202] 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 best be done, I have said, 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.

28 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. [203]

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 I could save either White or the five,
>
> *Tunnel*, in which I could redirect a runaway train so that it kills White rather than the five,

and

> *Bridge*, in which I could save the five only by killing White.

According to one view, in all three cases, I ought to save the five. It makes no difference whether, in saving the five, I would be killing White. When people's lives are threatened, we ought to do whatever would save the most lives.

According to a second view, I 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 me to save the five in both *Tunnel* and *Bridge*, since these ways of saving the five would both kill White. As before, it makes no difference whether I would be killing White as a means.

According to a third view, I ought to save the five in *Lifeboat*, and I would be at least permitted to save the five in *Tunnel*, but it would be wrong for me 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 I would be killing White 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. [204] 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 Vew, 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 I could justifiably redirect the runaway train so that it would kill White rather than the five, it would be wrong for me to save the five *by* killing White. I have *not* been arguing against this view.

The answer to Q2, I believe, is always or nearly always No. If I killed White in *Bridge* without her consent, I might not be treating White merely as a means, or be close to doing that. My treatment of White 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 I had been closer to the train, I would have saved the five by killing myself rather than White. But these facts would not, I believe, affect whether my act would be wrong. If it would be wrong for me to kill White as a means of saving the five, this act would be wrong *whether or not* I would also be treating White merely as a means. Even if I was *not* treating White merely as a means, and was not even close to doing that, these facts would not justify my 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'. [205] 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 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. [206]

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 9 RESPECT AND VALUE

29 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. [207]

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. [208]

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. [209] 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. [210] 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. [211] Though (C) cannot directly solve all moral problems, this principle provides, Wood claims, 'the correct basis for deciding moral questions'. [212] To support this claim, Wood points out that in his longest book about morality, Kant often makes remarks that seem to appeal to (C). [213]

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'. [214] 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'. [215] 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. [216] And, when he condemns lying even 'to achieve a 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'. [217] 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'. [218]

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.

30 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'. [219] 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.' [220] 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'. [221]

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 properties or features that would, in some situations, give us or others reasons to respond to these things in certain ways. [222]

Some of these 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, processes, 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 can 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 can 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, or to happen, and to make them actual if we can. Possible acts and other events would be good as a means when our making them happen, or be actual, would be an effective way of achieving some end. [223]

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 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. [224]

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.'[225] 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 can 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. [226]

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. [227]

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 lives, and to help these people in other ways to achieve these good ends.

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'. [228] 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. [229]

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.

31 Kantian Dignity

We can now turn to 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. 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. [230] 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. [231] 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.'[232]

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. [233]

A third such end is what Kant calls the *Highest* or *Greatest Good*. [234] 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. [235] Kant claims that 'we ought to try to promote' this end, and that 'reason. . . commands us to contribute everything possible to its production.'[236]

There may be a fourth such end. Kant calls rational beings 'something whose existence in itself has absolute worth'. [237] And he writes that, if there were no rational beings, the Universe would be 'a mere waste, in vain, without a final purpose'. [238] These remarks suggest that, on Kant's view, the continued existence of rational beings is another end-to-be-produced with supreme value. [239]

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',[240] and a person whose will is good 'is of no greater value' than someone with an ordinary or a bad will. [241] 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 positive 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'. [242] Wood even calls Kant's claim about rationality's value 'the most fundamental proposition in Kant's entire ethical theory'.[243]

Like having a good will, rationality is in part an end-to-be-produced, or promoted, since 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*. [244] 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. [245] 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 can suppose that we could save these people from pain during needed 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. [246] 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 be almost as 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. [247]

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', [248] and his uses of 'humanity' are shifting and vague. Kant does condemn some vices, such as gluttony and drunkenness, on the ground that they interfere with our rational activities or abilities. [249] 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'. [250] On Kant's view, 'past moral philosophy . . . mistakes the nature of the good'. [251]

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. [252] 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 F, 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. [253] 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 they ought to be treated only in certain ways. This value is a kind of *status*, or

what Herman calls 'moral standing.' [254] 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 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.

32 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. [255] Kant also writes:

> Everyone ought to strive to promote the Greatest Good. [256]

> the moral law commands me to make the greatest possible good in a world the final object of all my conduct. [257]

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.[258]

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. [259] 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 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 assumes, 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. [260] 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

> *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. [261]

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 trivial claim, whatever their moral beliefs. Moore's *Principia* does not put forward a substantive moral view. [262]

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. [263]

Surprisingly, Kant also writes:

> All imperatives are expressed by an 'ought'. . . and say that. . . some act would be good. [264]

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.' [265] 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, [266]

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*. [267] As we have seen, however, Kant himself describes an ideal world which he calls the 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. ' [268]

This is not, I believe, the best way to interpret K3. Kant, I suggest, uses 'the Greatest Good' in the 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* value-based.

33 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. [269]

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). [270]

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. We should not assume that earlier thinkers drew all the distinctions that we now draw. It

was widely assumed, when Kant wrote, that it is by following the rules of common sense morality that everyone could best promote everyone's happiness.

This assumption is also fairly plausible. As Sidgwick later argued, if everyone always tried directly to maximize happiness, there would probably be less happiness than there would be if everyone tried instead to follow the rules of common sense morality. 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 we 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 now draw some other distinctions that many earlier thinkers did not draw. If we suppose that everyone will accept some set of rules, some possible rules would be

> *optimific* in the sense that these are the rules whose acceptance by everyone would be most likely to make things go best. [271]

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 best, or expectably-best.

Of these who accept either of these views, most now believe that we have to choose between them, given the fact that

> (A) breaking some optimific rule is sometimes likely 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 (A). These

people believe that

> (B) it is by always trying to follow the optimific rules that everyone would be most likely to make things go best, or expectably-best.

Moore came close to accepting (B). In trying to do the most good, Moore claims, we ought always to try to follow certain optimific common sense rules.[272] If (B) were true, these two forms of consequentialism would 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 (B) is true, 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.

In some kinds of case, this view can seem implausible. One example are the 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.

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, 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'. [273] 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 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

> (B) 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. [274] 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. [275]

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 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. . .[276]

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. [277]

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 (B) 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 (B). 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. If consequentialists accept this view, they have to choose between Act and Rule Consequentialism.

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 which 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 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. [278] 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. [279] These claims about desert cannot be plausibly derived from, or claimed to be supported by, Kant's other formulas. [280] 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. [281]

Kant's claims about desert are, I believe, false. And, as I shall now argue, Kant came close to seeing that.

CHAPTER 10 FREE WILL AND DESERT

34 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. [282] 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, they might have undetermined origins in the timeless noumenal world. That 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. [283]

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? [284] 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 right---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 (D) and (E) claim, 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 is enough that

> 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 were causally inevitable, we would have acted differently only if we could have 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 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 believe that we ought to have acted differently, this belief may cause it to be true that in similar cases, without any miracle, we *do* in *future* act differently.[285] 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 simple mechanical device that turns by itself when it has been wound up. 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 possible different 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 obvious that many of our acts *do* depend on our decisions, fatalism can seem 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 supernatural power would have had 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 can seldom know in advance what we shall decide to do. And if we make better decisions, and act upon them, things will be likely to go better. That is 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 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. [286]

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 best escape from the advancing smoke and flames. 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. [287]

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 assume correctly that our acts will depend on our decisions, and 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, except our motives and other mental states. We could often act in any of several ways, *if* we chose 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 choice would have been causally

impossible. This, I believe, is the true view. This is the sense of 'could' and 'can' with which we can justifiably claim that 'ought' implies 'can'. And in this sense we have the freedom that morality requires.

Though Kant here suggests that, for practical purposes, the kind of freedom that we need is compatible with determinism, his dominant view is clearly incompatibilist. Kant even claims that noumenal timeless freedom is the keystone of his entire philosophy. He would not have made that claim if he had believed that, even if all of our acts were causally determined, we could have the freedom that morality requires.

According to the argument given above, more briefly stated:

> (H) 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 argument that is summed up in premise (H). 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 compatible with determinism. Even if our acts are causally determined, we could have the kind of freedom that morality requires.

35 Deserving 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 a way that could make them deserve to suffer. That is true, for example, of young children, or 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, however, a better way to explain why these people cannot act wrongly. Young children and these insane people cannot recognize or respond to moral reasons. But ordinary sane adults can recognize and respond to such reasons. That is enough to justify our belief that, as sane adults, we can be 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 we ought to act in certain ways, and that no one could ever deserve to suffer.

According to premise (J), if our acts were merely events in time, we could not be responsible for our acts in some way that could make us deserve to suffer. 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 is ever responsible, I believe, in this suffering-deserving way.

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 can be partly uncaused, or random. But such claims may not apply to our decisions to act, and to other mental events. Most neuroscientists believe that these mental events consist in, or causally depend upon, physical events in our brains which *are* fully causally determined, because they occur on too large a scale to be affected by random events at the level of sub-atomic particles. Some people, however, 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 consist in, or fully causally depend upon, physical events in our brains.

To justify the belief that we can deserve to suffer, however, it is not enough to justify 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. [288]

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 writers 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 writers try to describe some third possibility. Some writers appeal to our rationality. When we claim that someone acted *for some reason*, 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 the occurrence of 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 the kind of freedom that could make it true that we can deserve to suffer because of what we did. But because these writers believe that, as agents, we are fully part of the spatio-temporal world, they cannot intelligibly claim that the causing of an act by an agent is *not* an event.

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. . . [289]

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. [290]

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 were merely events in time, we could not be responsible for our acts in any way that could make us deserve to suffer. [291]

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 has one advantage. We should not expect, Kant claims, to understand this noumenal timeless world. All we can understand is the spatio-temporal phenomenal world. 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 how so many free decisions 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. There are other problems. There is no intelligible way in which our many moral decisions within our lives could depend on what we timelessly decide in the noumenal 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 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 causally inevitable, we could deserve to suffer in Hell. On such a view, to deserve to suffer, we don't have to be free, or to be in any way responsible 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 expect not to understand how we can have timeless freedom, it is less plausible to claim that we should expect not to understand how we could deserve to suffer. We have no similar 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 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 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 he 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 our own character. I can merely hope that Kant would have continued to believe (V), and would have therefore concluded 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. 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 die. 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. [292]

I have been discussing some of Kant's claims about the ideal world that we ought always to strive to promote. But these are not the most valuable parts of Kant's moral theory. Many other writers have claimed that the two greatest goods are virtue and happiness. And Kant says little to defend his assumption that it is by following his other formulas that we can 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 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 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 in ethics'. [293]

PART THREE

CHAPTER 11 UNIVERSAL LAWS

36 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', [294] 'Let no insult pass unavenged', [295] 'Make lying promises when that would benefit me', 'Give no help to those who are in need', [296] and 'the maxim of self-love, or one's own happiness'. [297]

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. [298]

This formula needs to be explained. In one passage, Kant refers to a maxim's being 'a universal permissive law'. [299] 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. [300]

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) claimed it to be wrong to act on any maxim that it would be logically impossible or inconceivable for everyone to accept, this formula would fail to condemn most wrong acts. We can easily conceive possible worlds in which everyone accepts bad maxims, such as the maxims 'Kill, deceive, and coerce other people when 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 also 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 to whom some maxim applies. The maxim 'Care for my children', for example, applies only to parents.

This formula would 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, for example, 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. (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 *successfully* to act upon it, in the sense that their act would achieve their aim.[301]

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 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, (A) to (D) 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. [302]

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. [303] 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. [304]

Korsgaard similarly writes:

> lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive. . . [305]

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 anyone to benefit themselves by telling some lie. 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. [306]

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'. [307] 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'. [308] 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. [309]

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.[310] 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 practice of morally motivated, trust-involving promises would have ceased to exist. Kant's formula therefore condemns such lying promises.[311]

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 claim, because if everyone believed such acts to be permissible, that would undermine a valuable social practice.

(G) is 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 acted on the lying-promiser's maxim, the practice of making promises would be a 'vain pretense', or sham. Suppose that some German soldier acted on the maxim 'Swear this oath but disobey all immoral commands'. We could similarly claim that if all these soldiers had acted on this maxim, the practice of swearing oaths of unconditional obedience would have been a vain pretense or sham. But, as Kant claims, everyone ought to disobey immoral commands.

For a clearer test of (G), we can suppose that, during the Second World War, some non-Jewish German 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 help Jews 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 act.

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. [312] 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 someone's life 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 promise. Kant's lying 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 promise is wrong. Similar claims apply to my example. My imagined person saves someone's life 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 implies that lying to save this person's life would be wrong. The most important difference between these acts is in what they are intended to achieve; and this difference is ignored by (G).

As these cases show, (G) is unacceptable. [313] This formula 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.

According to Hegel and some other writers, this formula 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. It is true that, if all rich people acted on this maxim, that might abolish poverty, thereby making it impossible for anyone later to act on this maxim. But (H) would not condemn these people's acts, because by giving to the poor these people would *achieve* their aim. [314]

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, that 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. [315]

Consider next those men who accepted codes of honour, like the code that led the Russian poet Pushkin to his fatal duel in the snow. Suppose that Pushkin had accepted the maxim 'Fight duels to show my courage, but always shoot to miss'. 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 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, she suggests, is

> (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. [316]

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.' [317] 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. [318] But this claim is also false. Two people can simultaneously deceive each other. And there can be mutual coercion. 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 might reply that, to show that (I) condemns deception and coercion, it is enough to appeal to the weaker 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) condemns playing competitive games with the aim of winning. [319] Perhaps we could accept that conclusion. But there is nothing wrong with acting on the maxim 'Become a doctor', even when, by applying and being admitted to some medical school, we prevent 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'. [320] 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, [321] pay the debts on our credit cards before interest is charged,[322] buy only second-hand books, or give surprise parties. [323]

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 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.

37 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. [324]

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 claims, 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. [325]

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 I believe that, for our choices to be rational, we must also respond 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. [326] 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. [327]

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 everyone's *being* permitted to act on this maxim, but by everyone's *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. [328]

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. [329] 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'. [330] 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 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. So we should appeal to our own beliefs about rationality and reasons, since we are then appealing to what we believe to be the truest or best view.

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 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, these wrong acts are not rational, since everyone is 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 his belief that this man is rationally required to give such help. So, 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. [331] 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.

38 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 first that I wrongly steal some wallet containing $63 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 this maxim to be 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. So Kant's formulas mistakenly permit my act.[332] Similar claims apply to other highly specific maxims. When wrong-doers act on such maxims, they could often rationally will that their maxims be universal, because they would know that other such acts would at most be very rare, and would therefore make very little difference. Kant's formulas 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 she is doing.[333] In a realistic version of my example, I would have stolen from my victim even if her wallet had contained only $62, or if she were dressed in red, or eating blueberries, or 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 condemn my act, in the sense of implying that this act is wrong.

These remarks do not fully answer the Rarity Objection. Even if actual wrong-doers never act on such highly specific maxims, we can imagine such people. Kant's formulas ought to be able to condemn these imagined people's acts.[334] 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 always acts on

the Egoistic maxim: Do whatever would be best for me.

This Egoist 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. Such a world would be much worse for him. Egoists have strong self-interested reasons to want everyone else to accept and follow, not the Egoistic maxim, but various moral maxims. Since this Egoist always acts on a maxim that he could not rationally will to be universal, Kant's formulas imply that all of his acts are wrong. This man acts wrongly, not only when he steals, breaks promises, and harms other people, but also when, for self-interested reasons, he pays his debts, keeps his promises, and helps other people. These are unacceptable conclusions. When this Egoist saves some child from drowning because he hopes to get some reward, his egoistic motive makes his act have no moral worth. But he is not acting wrongly.

It might be claimed that, when this man saves this drowning child, *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. [335] Kant similarly claims that to fulfil a duty of gratitude, we must feel grateful. [336]

These distinctions cannot answer this objection to Kant's formulas. My Egoist could not fulfil any duties of virtue, since he never has the right motive. But, as Kant claims, 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 man's 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. Our problem is that, 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 thinks it worth taking these risks, since he could then hope to get a greater reward. On the suggestion we are now considering, if this man saves this child 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'. [337] 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 some people act on this maxim, or policy, it may be simply false to claim that these people accept and act on any other, less general policies.

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 claims suggest and we can here suppose, Kant accepted the maxim 'Never lie'. This maxim is condemned by Kant's Law of Nature Formula. 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. [338] So Kant's formula implies that, whenever Kant acted on this maxim by telling anyone the truth, he acted wrongly. 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 an act would be the only way to save some innocent person's life. So Kant's formula implies 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 might be the maxims 'Torture others for my own amusement' and 'Prevent pointless suffering'. When applied to such maxims, Kant's formula succeeds. 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 'Do whatever would be best for me' and Kant's maxim 'Never lie'. In proposing his Law of Nature Formula, Kant overlooks such mixed maxims. Kant's formula assumes that acting on some maxim is either always wrong, or never wrong. When applied to mixed maxims, Kant's formula fails, since this formula condemns some acts that are 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 formula mistakenly implies, but doing their duty. 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'. [339] Herman claims that, despite a 'sad history of attempts. . . no one has been able to make it work'.[340] O'Neill suggests that, in some cases, Kant's formula may 'give either unacceptable guidance or none at all'. [341] Hill doubts whether, when used on its own, Kant's formula can provide 'even a loose and partial action guide'. [342]

Because they believe that Kant's formula cannot provide a criterion of wrongness, some of these people 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. [343] O'Neill suggests that Kant may intend his formula to provide a test, not of which acts are wrong, but only of which acts have moral worth. [344]

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* or are *contrary to duty*. And 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? [345]

> 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.[346]

Kant also claims that his formula 'determines quite precisely what is to be done. . . with respect to all duty in general'.[347]

These 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 Kant's formula, we can first redescribe 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.

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 the Egoist or Kant would or might act wrongly. But, because Kant's formula appeals to the agent's maxim, this formula mistakenly implies

that the Egoist acts wrongly when he saves this child, and that Kant acts wrongly whenever he tells someone the correct time.

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.[348]

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 countless 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 could 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 could 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, or 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 she is directly trying to achieve. We should also describe the effects which this person believes that her acts might have. What people *intentionally do* is not the same as what they *intend*. To give Sidgwick's example, if some Russian nihilist 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. [349]

When we describe people's acts, we are usually describing what these people are intentionally doing. [350] It is sometimes unclear how we ought to describe some act. It may be unclear, for example, how much we ought to include in some act's foreseeable effects, or what we should 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 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. [351] But this narrower sense of 'maxim' would add nothing to the morally relevant descriptions of people's acts.

It might now 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 improve them, 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'. [352] But, as Kant points out, wrong-doers do not in fact will that their maxims be universal, so that 'there is really no contradiction' in these people's wills. [353]

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. [354] But as we have seen and O'Neill elsewhere claims, [355] 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 when we apply Kant's Moral Belief Formula we ask whether we could rationally will it to be true that some subjective principle is a universal moral law. My 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 principle that permits such acts.

This revision keeps Kant's concern with principles and moral laws. [356]

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 is irrelevant that Kant is acting on a maxim that he could not rationally will to be universal. [357] 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, such as 'Never break the law', which 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 12 WHAT IF EVERYONE DID THAT?

39 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 will 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 these ways. 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*. [358]

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. 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 herself 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.

If each of us benefits herself, each of us is doing what is certain to be better for herself, 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 herself, *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 herself. [359]

These claims are *not* about what are misleadingly called *repeated prisoner's dilemmas*, which I discuss only in a note. [360]

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. [361]

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. [362] 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 she herself did not contribute. She would avoid the costs to herself, and she would be no less likely to receive the greater benefits from others. In many of these cases, the public good is avoiding 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 catches 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. 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, they 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 giving benefits to ourselves or 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 or fines that are 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. These solutions are also seldom achieved, since there are few people who are both conscientious Kantians and have applied Kant's formula to these cases.

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 herself 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 say or think '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 the other way round. 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 benefit ourselves, or 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 herself, or for her 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. [363]

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. [364]

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. [365]

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.

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, when most other people are *not* doing what we should all ideally do, we are excusably permitted, as a defensive second-best, to give the lesser benefits to ourselves, our children, or our other M-related people. [366]

We can now turn to some cases in which Kant's formulas do less well.

40 The Threshold Objection

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.

Whether some act is wrong, however, 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. [367] Consider next the maxims: 'Consume food without producing any', 'Become a dentist', and 'Live in Iceland, to absorb the spirit of the Sagas'. [368] 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 would be 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 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*. In such cases, if someone says 'What if everyone did that?', it is often enough to reply 'Most people won't.'

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. [369] 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 others act 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 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. [370] 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.

If we acted on these unconditional maxims, we would be having no children, or producing no food, when 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. So LN3 does not mistakenly imply that these acts would be wrong.

41 The Ideal World Objection

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 are 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 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. [371]

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 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. [372]

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. [373]

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. [374] 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 people in my position 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. [375] 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 see this problem in its clearest form, consider

> 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 just 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. [376]

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 would 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. [377] 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. [378] 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. [379]

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 however many people 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'. But, 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 what other people have done, are doing, or will do. 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 below which acts of some kind 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. [380]

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. [381]

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, as I argued earlier, it is enough to reply 'Most people won't'. [382]

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. 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. [383]

We can now turn to some simpler and more fundamental questions.

CHAPTER 13 IMPARTIALITY

42 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. [384]

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. [385] Though Kant calls his formula 'the supreme principle of morality', he dismisses the Golden Rule as 'trivial' and unfit to be a universal law. [386] 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, if they could thereby be excused from benefiting others.

This objection backfires. These people ought to 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 could 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 applies to Kant's own formula. According to this formula, these people ought to help others if they could not will that the maxim of not helping 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.' [387] 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 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 others 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, if we

> had the choice and were rational, we would choose that others treat us.

To save words, I shall talk of the ways in which we *would rationally choose* that we be treated.

When we apply the Golden Rule, it is sometimes enough to ask whether we would rationally choose that, in the actual world, we be treated in some way. Torturers, for example, would not if they were rational choose to be tortured. But, when considering many kinds of act, we must ask how we would rationally choose that we 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 rationally choose that other people give us no food. If we have just eaten well, and have a well-stocked kitchen, our answer to that question might be Yes. We ought to ask whether we would rationally choose that, even if we ourselves were starving, other people give us 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 choose that we ourselves be treated. I admit to my hotel anyone who is not black. I would happily choose that I 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 choose that he himself 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 himself being treated as he treats women, he may have to imagine that he is a woman.

In a fuller statement, then, the Golden Rule could be

> G3: We ought to treat others only in ways in which, if we were rational, we would choose that we ourselves be treated, if we were going to be in these other people's positions, and would be relevantly like them.

The phrase 'would choose' can be misleading. In applying G3, we should not ask how, if we were in these other people's positions, we would *then* choose that we be treated. We should

ask how we would *now* choose that we 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 want or choose that we be treated, if we were going to be in this person's position, and 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 rationally choose that I 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 beggars.

By appealing to G4, each beggar could argue that I ought to give her my whole loaf.

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 rationally choose that I 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, and that I would be acting wrongly if I shared my loaf equally between these people.

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 choose that we 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 being 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 would later divide and become two people, one in Green's position and the other in Grey's. [388] On Richard Hare's proposal, I should imagine that I would later live lives that would be just like those of Green and Grey, not simultaneously, but one after the other. [389] On John Harsanyi's proposal, I should imagine that I 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. [390]

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 being in the positions of Green and Grey, I ought rationally to choose that I be saved from death in one of these positions rather than being saved from blindness 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. [391] These judges should ask how they would rationally choose that they 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. [392] 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 rationally choose 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 choose that we 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 rationally choose to do if I were going to be, not only in Grey's position, but also in mine. As in *Case Three,* I ought rationally to choose that I be saved from death in one of these positions rather than being saved from blindness 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 choose 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. [393]

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 could we rationally choose that we be treated in these ways if we were going to be in the positions of all 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'. [394]

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 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 it would be rational for us to choose 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 possible events that 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 are going to be in other people's positions, and would 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 are going to 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 often cannot be compensated by benefits to other people. In ignoring this 'separateness of persons', we are ignoring facts that may give us reasons to accept principles of distributive justice. [395]

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 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.

43 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 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 are clearly 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.

44 The Non-Reversibility Objection

There is another, similar, but practically more important objection to Kant's formulas. The Golden Rule makes us more impartial by requiring us to treat everyone as we would rationally choose that we ourselves 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. [396] 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. Kant's formula works best when it is applied to those wrong acts with which we benefit ourselves in ways that impose much greater burdens on others. The Golden Rule condemns such acts, since we could not rationally choose that 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 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 that everyone acts in this way. Everyone *does* act in this way. Every hotel owner excludes blacks. And I could rationally will 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. [397] 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 that most people do not accept, and on which, he assumes, no one would want everyone to act. Two examples are the maxims of self-interested deception and theft. In acting on such maxims, Kant's wrong-doers make unfair exceptions for themselves. To condemn such acts, we can claim that these wrong-doers could not rationally will it to be true that everyone acts like them.

Our white racist is in a different position. When this man acts on the maxim 'Exclude blacks from my hotel', he is doing what, in his social world, all hotel owners do. So it does not help to ask, 'What if everyone did that?' Nor would it help to ask

whether this man could rationally will it to be true that everyone believes his acts to be morally permitted. This man would be happy if no one believed his acts to be wrong. Kant did not consider cases of this kind. When Kant imagines some wrong-doer asking 'Could I will that my maxim be a universal law?', he assumes that this person's maxim *isn't* such a law. [398] But in some cases, like that of this white racist, some wrong-doer's maxim is already a universal law, since this maxim is already acted on by all of the people to whom it applies.

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 this 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 some 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 treat women as inferior, denying women various rights and privileges, and giving less weight to their 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 could men or slave-owners rationally choose that they 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 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 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 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. [399]

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.

45 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. [400] 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. [401] 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*. [402]

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. [403] 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. [404]

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. [405] 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. [406] 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. [407]

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 several 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. [408]

Though Kant never appeals to what everyone could rationally

will, that may be only because he assumes 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. [409]

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 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, whenever they can.

As I explain in a note, however, it is enough to appeal to MB5. [410]

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?"' [411]

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. [412] 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 14 CONTRACTUALISM

46 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 acceptance by everyone 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 other principles whose universal acceptance would be best only for certain people. What would be best for men, for example, might not be best for women. It may seem that, in such cases, there would be no principle whose choice would be rational for everyone in self-interested terms. But the Rational Agreement Formula applies only to those principles to whose acceptance it would be rational for *everyone* to agree. There would be no point in our choosing principles whose acceptance would be best for ourselves, if these are principles that some other people could *not* rationally choose.

What we could rationally choose would also depend on the effects of our failing to reach agreement. Some contractualists tell us to suppose that, if we failed to agree, no one would accept any moral principles, so no one would believe that any acts were wrong. That would be likely to be bad for everyone. In this amoral *no-agreement world*, Hobbes writes, 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. [413] 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, thereby trying to induce others to choose these principles. 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. 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, however, we would have to defend the claim that everyone *would* have rationally reached agreement on these principles. And this claim would be hard to defend. When David Gauthier appeals to his 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.' [414] 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 choose the principles that everyone else would choose. But how could we predict what others would choose? If we were merely trying to reach agreement on how some fixed set of resources would be shared between us, we might be able to solve this coordination problem. It might be 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 there is no such obvious solution. In trying to predict which principles other people would choose, each of us would be groping in the dark. So in the decisive final vote, there would be no single set of principles that it would be rational for everyone to choose. [415]

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 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 them. 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. [416]

Gauthier argues that, if we are prepared to accept his minimal version of morality, his contractualist theory shows 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. Nor, I believe, could other Hobbesian theories succeed. Hobbesian contractualists give unsound arguments for unacceptable conclusions.

Kantian contractualists, like Rawls, reject these conclusions. As Rawls writes, 'to each according to his threat advantage is not a conception of justice'. [417] So we can now turn to Rawls's view.

47 Rawlsian Contractualism

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*. [418]

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. [419]

In considering Rawlsian Contractualism, we can start with Rawls's assumptions about rationality and reasons. Rawls accepts a desire-based subjective theory, which claims it to be rational 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 it to be rational to

try to do what 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. [420]

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. [421]

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. As Rawls writes,

> knowing that people are rational, we do not know the ends they will pursue, only that they will pursue them intelligently. [422]

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. . . [423]

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. [424] 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. [425]

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 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'. [426] There might be no principles on whose acceptance it would be rational for everyone to agree. If, in contrast, 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. [427]

One of Rawls's main aims, he writes, is to produce a systematic

theory which provides an alternative to all forms of utilitarianism. [428] It is surprising that, in trying to achieve this aim, Rawls proposes his version of contractualism. If our moral reasoning appeals to a combination of self-interested rationality and impartiality, we should expect this reasoning to support some view that was, or was close to being, utilitarian. [429] As Rawls himself points out, utilitarianism is, roughly, self-interested rationality plus impartiality. [430]

Rawls is aware of this problem. On 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. [431] 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. [432]

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, 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'. [433] 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. [434]

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 argument for 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. [435] We would then be closer to achieving Rawls's aim of 'the thickest possible veil of ignorance'. But this version of contractualism cannot 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. [436] 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'. [437] 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. [438] 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. [439]

As Rawls points out, 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 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 valid, 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 extra year of life, rather than giving fifty extra years to each of a thousand other people---people who, without these 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. [440]

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 this argument cannot 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 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. [441]

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. [442]

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. [443]

He also writes

> for the contract view, which is the traditional alternative to utilitarianism, such a conclusion would be a disaster. [444]

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 support this conclusion. [445]

48 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 carry out a single thought-experiment, in which 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 *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 *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 the 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 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 facts.

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. [446]

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 greater power, not to those who *most* need morality's protection, but to those who *least* need such protection, because they have more abilities, or control more resources.

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 us to impose great burdens. And since Rawls appeals to the principles whose choice would be rational in self-interested terms, and he has no relevant objection to the Equal Chance Formula, Rawls cannot plausibly deny that we could rationally choose utilitarian principles, or other similar principles, running the small risks of bearing some great burden for the sake of much more likely benefits. [447]

Since the Kantian Formula requires unanimity without appealing either to a veil of ignorance or to a need to reach agreement, 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, this claim could be plausibly denied.

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 we ought to accept either Rational Egoism, or some desire-based subjective theory, the Kantian Formula would not succeed. If each person supposed that she herself had the power to choose which principles everyone would accept, 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 acceptance would best fulfil everyone's informed desires.

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 wide value-based objective theories of the kind that I believe to be 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. [448]

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. If everyone could rationally choose two or more seriously conflicting principles, this formula might permit too many acts. 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. [449] 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

> (1) some quantity of unowned goods can be shared between different people,
>
> (2) 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

> (3) if these goods were equally distributed, that would produce the greatest sum of benefits.

It seems clear that, in such cases, everyone should be given equal shares.

The Kantian Formula appeals to the principles that everyone could rationally choose, if each person supposed that everyone would accept whatever principles she herself chose. We 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 present 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 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.

49 The Deontic Beliefs Restriction

When we apply Kantian or contractualist formulas, as I have often said, we cannot appeal to our beliefs about the wrongness of any of the acts that we are considering. We can next look more closely at this *Deontic Beliefs Restriction*.

We can also introduce another version of contractualism. According to

> *Scanlon's Formula*: Everyone ought to follow the principles that no one could reasonably reject. [450]

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. [451]

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 Scanlon's appeal to the Deontic Beliefs Restriction. [452] Suppose again that, in

> *Means*, Grey and Blue are trapped in 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.

We may 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 saving someone else's life.

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. [453]

Blue might 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 much less bad than losing many years of life.

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, that would be like pulling on our bootlaces in an attempt to hold ourselves in mid air. To vary the metaphor, 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. [454]

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 an arm, dying is a much greater burden. This is one of the kinds of fact which, on Scanlon's view, can provide reasonable grounds for rejecting some moral principle. When Grey defends the Harmful Means Principle and rejects the Greater Burden 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 deontic beliefs.

Like Rawls, Scanlon proposes his contractualism partly as a way of avoiding Act Utilitarianism, or *AU*. [455] In one way, however, 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. [456]

But when we apply some contractualist formula, and follow the Deontic Beliefs Restriction, we cannot appeal to such convictions.

Even without appealing to such moral convictions, however, Scanlonian Contractualists could reject Act Utilitarianism. Consider

> *Transplant*: White is in hospital, to have some minor operation. I am White's doctor. I know that, if I secretly killed White, her transplanted organs would be used to save the lives of five other people. [457]

According to

> AU: We ought always to do, or try to do, whatever would benefit people most.

This principle requires me to save these five people by killing

White, since that is how I would benefit people most. Most of us would believe this act to be wrong.

We can plausibly defend this belief by appealing to 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. [458] 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*. [459]

This argument illustrates another way in which, if we appeal to a contractualist formula, that makes a difference to our moral reasoning. If we consider *Transplant* on its own, we could ignore this argument. Since I could save the five by secretly killing White, my 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*, my act would be secret, and would therefore produce no anxiety or mistrust. [460]

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, I could redirect a driverless, runaway train, so that it kills White rather than five other people,

and that in

> *Bridge*, I could save the five only by using remote control to make White fall in front of the train, thereby killing White, 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 others.

What is morally important, on this view, is how my saving of the five would be causally related to the act with which I kill White. It would be wrong for me to save the five in *Transplant* and *Bridge by* killing White, but it would not be wrong for me to kill White in *Tunnel*, since I would here be killing White, not as a means of saving the five, but only as the foreseen side-effect of redirecting the train. When we apply Scanlon's Formula, can we plausibly defend this distinction?

The answer, I suggest, is No. When we consider the possibility that we shall be involved in 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. I might have sufficient reasons to help you, for example, if you ask me for such help, though I would not have such reasons if you tried to get me to help by deception or coercion. But, when someone else's act would kill us but would 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 my acts in *Tunnel* and *Bridge*. If White could *not* reasonably reject some principle that would permit me to kill her in *Tunnel*, it seems doubtful that she could reasonably reject every principle that would permit me to kill her 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 threats to people's lives, such as some fire, flood, avalanche, or driverless run-away train. [461] The Emergency Principle condemns my saving the five by killing White in *Transplant*, since I am here White's doctor. But this principle permits me to save the five in a way that kills White in both *Tunnel* and *Bridge*, because these are non-medical emergencies, and in these cases White would be a stranger to me.

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 my saving of the five and my killing of White, but the *personal* relation between me and White in *Transplant*, and the other differences between medical and non-medical emergencies. 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.

Suppose that, after thinking hard about these imagined cases, we believe that I would be morally permitted to kill White in *Tunnel* by redirecting the train away from the five. We also believe, however, that it would be wrong for me to kill White in *Bridge* as a means of saving the five. 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.

According to a second reply, when we are trying to decide whether some act is wrong, nothing is achieved by asking whether we believe that such acts are wrong, since that merely restates our question. But this reply misunderstands the way in which good moral reasoning consists in part in appeals to moral intuitions. In trying to decide whether some kind of act is wrong, we should compare what seem to us to be the most plausible relevant principles. We ought to choose between these principles in part by asking whether, in the other cases to which they apply, these principles have implications that conflict with our moral intuitions, by requiring acts that we believe to be wrong, or condemning acts that we believe to be right. Such thinking may lead us to revise either some of these principles, or some of these intuitive beliefs, or both. In thinking about morality in this way, we are trying to achieve what Rawls calls 'reflective equilibrium'.

A third 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.[462] 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. [463] 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. [464] If we are constructivist contractualists, and we decide 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 mere 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. [465]

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 being a lying promise, or causing pointless suffering---these facts may non-causally make these acts wrong. Being a lying promise isn't the same as being wrong. But, if some act is a lying promise, 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. [466]

Scanlon once claimed that his contractualism gives an account, not of what *makes* acts wrong, but of wrongness itself, or *what it is* for acts to be wrong. This claim was, I believe, a mistake. To see why, we can first consider another statement of the Kantian Contractualist Formula:

> 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 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 some act is disallowed by such a principle, that is *what it is* for this act 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. If we could defend such claims about all kinds of wrong act, the Kantian Formula would give a further, unifying explanation of these wrongness of these acts.

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 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. According to this version of Scanlon's view, 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 these other senses wrong.

If contractualists make such claims, they can defend the Deontic Beliefs Restriction without rejecting our moral intuitions as worthless, or illusions. On these versions of contractualism, it is only *while* we are asking what these contractualist formulas imply that we should not appeal to our beliefs about the wrongness of any 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. [467]

CHAPTER 15 CONSEQUENTIALISM

50 What Would Make Things Go Best

Before we ask what is implied by Kantian Contractualism, it may help to say some more about the goodness of outcomes.

Pain is bad, some of us 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. That may be why he claims that it cannot be contrary to reason to prefer our own acknowledged lesser good to our greater good. Hume often uses 'good' and 'evil' merely to mean 'pleasure' and 'pain'. [468]

While Hume would have thought it trivial to claim that pain is evil, Kant sometimes rejects this claim. For example, he 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. [469]

When Kant claims that pain cannot be evil, he means that pain cannot be morally bad. Like Hume, Kant seems here to be unaware of, or to forget, the reason-implying sense in which it is bad to be in pain. [470]

So does Ross. If some outcome would be bad, Ross assumes, we have a strong moral reason to prevent this outcome, if we can. Because we have no such reason to prevent our own 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. [471] 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.' [472] Some writers claim that, though outcomes can be good or bad for particular people, there is no sense in which outcomes could be impersonally good or bad. [473] 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.

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 impersonally bad in the sense that we all have reasons to regret anyone's being in pain, whatever that person's relation to us. Such badness involves *omnipersonal* reasons. In the same way, we all have impartial 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.

Though many outcomes are impersonally bad because they are bad *for* one or more people, outcomes can be bad in other ways. If we continue to overheat the Earth's atmosphere, those who live in future centuries would be much worse off than the different future people who would later live if instead we ceased to behave in this selfish way. It would be worse if the future quality of life would be in this way much lower, even though, because it would be different people who would later live, this outcome would be worse for no one.

In what follows, I shall use 'best' in the impartial-reason-implying sense. There are often two or more possible outcomes that could be called 'equal-best'. But this phrase misleadingly suggests that there are precise truths about the relative goodness of different outcomes. In most cases there are no such truths. When we describe such cases, it would be clearer to say that there are two or more possible outcomes that would not be worse than any of the others. [474] To save words, however, I shall use 'best' to refer to all such outcomes.

Value-based consequentialists believe that

> (1) whether our acts are right or wrong depends only on facts about how it would be best for things to go.

These people may have conflicting beliefs about what is good or bad. Some consequentialists are *utilitarians*, who believe that

> (2) 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 take the goodness of outcomes to depend in part on other facts. They may, for example, believe that

> (3) 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 the people who were worse off.

The word 'consequentialist' is in one way misleading, as is talk of the goodness of 'outcomes', since these words suggest that on these moral theories all that matters is the future, and the effects of people's acts. The goodness of some outcomes, consequentialists can claim, depends in part on facts about the past. It might be better, for example, if benefits went to people who had earlier been worse off. And some acts may be in themselves good or bad events. Kind acts may be good even when they fail, and the badness of cruelty may not be only in its effects. So when consequentialists claim that some act would *make* things go best, they may not mean that this act would *cause* things to go best. The doing of this act may be part of how it would be best for things to go.

All consequentialists appeal to claims about what would make things go best. *Direct* Consequentialists apply this test 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 could make things go better or worse. When these people apply this test to acts, they are *Act* Consequentialists. Some of these people claim that

> (4) everyone ought always to do whatever would in fact make things go best.

Others claim that

> (5) 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*. [475]

As I have said, however, we should use 'ought' in several senses. If (4) used 'ought' in the fact-relative sense and (5) used 'ought' in the evidence-relative sense, these claims would not conflict, and Act Consequentialists could accept them both. Nor would either of these claims conflict with a version of (5) which used 'ought' in the belief-relative sense. Since we often don't know which acts would in fact make things go best, (5) is in practice more

important than (4). But 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 test directly to some things but only *indirectly* to others. *Rule* Consequentialists apply this test directly to rules or principles, but only indirectly to acts. Some of these people believe that

> (6) 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 claimed to be wrong 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 views overlap with those systematic forms of *virtue ethics* which appeal to the character-traits and other dispositions that best promote human flourishing. There could be many other forms of Indirect Consequentialism. [476]

51 Consequentialist Maxims

Some consequentialists might apply their test 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:

> (A) 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.
>
> (B) 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.

(A) is Kant's Law of Nature Formula. If (B) is true, this 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, (B) 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 *fortunate* people. These 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 that assumption, (B) 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. [477] 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 some people's true needs as agents would be best fulfilled.

(B) 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. [478] Some of these people would have desires that would be best fulfilled if everyone acted on the optimific maxims.

Rational Egoists might reject (B). 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 would be glad if things went in the ways that would be impartially best. If things went in these ways, that would be best for at least a few of these people. These people could rationally will it to be true that everyone acts on the optimific maxims.

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. So Kant's Law of Nature Formula 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 could call this the *Relativism Objection*. This objection is met when we 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.

As I have argued, we have other strong reasons to revise Kant's formulas in this way, and reasons to make these formulas apply, not to our maxims, but to the morally relevant descriptions of our acts. These two revisions lead us back to the Kantian Contractualist Formula. So we can now ask what this formula implies.

52 The Kantian Argument

According to the *universal acceptance* version of Rule Consequentialism, or

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

Such principles we can call *UA-optimific*.

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 principles whose universal acceptance would make things go best.
>
> (D) These are the principles whose universal acceptance everyone would have the strongest impartial reasons to choose.
>
> (E) No one's impartial reasons would be decisively outweighed by any set of relevant conflicting reasons.

Therefore

> (F) Everyone would have sufficient reasons to choose that

everyone accepts these UA-optimific principles.

(G) There are no other significantly non-optimific principles whose universal acceptance everyone would have sufficient reasons to choose.

Therefore

(H) It is only these optimific principles whose universal acceptance everyone would have sufficient reasons to choose, and could rationally choose.

Therefore

These are the principles that everyone ought to follow.

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 Rule Consequentialist principles.

When we apply the Kantian Formula, we ask which principles each person could rationally choose, if this person supposed that 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, anyone could rationally choose what they would have sufficient reasons to choose.

We can next suppose that, as (C) claims, there is some set of principles that are UA-optimific. Of all of the principles that everyone might accept, these are the principles whose universal acceptance would make things go best in the impartial-reason-implying sense. When we consider some kinds of case, there might be two or more such optimific principles that were significantly different. That would raise some questions that would be best considered later. We should first try to get the main outlines right.

If everyone accepted these optimific 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. [479]

According to premise (E), no one's impartial reasons to choose these principles would be decisively outweighed by any set of 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 might be. It will be enough here to consider those other principles that are *significantly* non-optimific, in the sense that their 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.

53 Self-interested Reasons

In asking whether premise (E) is true, we should consider the strongest reasons that anyone might have not to choose 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.

You might be such a person. Suppose that, in

> *Lifeboat*, you are stranded on one rock, and five people are stranded on another. Before the rising tide covers both rocks, I could use a lifeboat to save either you or the five. You and the five are all strangers to me, and are in other ways relevantly similar.

Any optimific principle would require me 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 your rock is nearer to me. According to what we can call

> *the Nearness Principle*: In such cases, we ought to save the group that is nearer to us. [480]

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 you strong impartial reasons to choose that everyone accepts the Numbers Principle. But you would know that, if you made this choice, I would act on this principle by saving the five, and you would die. [481] We can suppose that, in dying, you would lose many happy years of life. That would give you strong self-interested reasons *not* to choose the Numbers Principle, since if you chose instead that everyone accepts the Nearness Principle, I would save your life. According to premise (E), these self-interested reasons would not decisively outweigh, or be stronger than, your impartial reasons to choose the Numbers Principle. Is that true?

According to subjective theories about reasons, the answer depends on your desires or aims. If you cared enough about the well-being of other people, you could rationally choose that everyone accepts the Numbers Principle. But we cannot assume that, in this and other similar cases, you and everyone else in your position would have sufficient desire-based or aim-based reasons to choose the optimific principles. So, if we ought to accept some subjective theory, premise (E) would be false, and there would be no principle about these cases that everyone could rationally choose. As I have argued, however, we ought to reject subjective theories, and accept some value-based objective theory.

According to one such theory,

> *Rational Egoism*: Each of us is rationally required to give supreme weight to our own well-being.

On this view, premise (E) is false. You could not rationally choose that everyone accepts the Numbers Principle, since that choice would be worse for you. But we ought, I believe, to reject this view.

According to a view at the opposite extreme,

> *Rational Impartialism*: Each of us is rationally required to give equal weight to everyone's well-being.

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). You would be rationally required to choose that everyone accepts some optimific principle, such as the Numbers Principle. [482] But we ought also, I believe, to reject this view.

According to

> *wide value-based objective theories*: When one possible act would make things go in the way that would be

> impartially best, but some other act would make things go best either for ourselves or for those 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 some such views, 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*, you could rationally choose that I save you; but you could also rationally choose instead that I save the five. So you could rationally choose that everyone accepts the Numbers Principle. These claims, I believe, are true.

According to some more egoistic objective views, we are rationally required to give strong priority to our own well-being. You would not have sufficient reasons to give up your life unless you could thereby save as many as a hundred or a thousand other people. But in the thought-experiment to which the Kantian Formula appeals, you would have the power to choose which principles everyone would accept, both now and in all future centuries. The principles you chose would be accepted by many billions of people. If you chose that everyone accepts the Numbers Principle rather than the Nearness Principle, your choice would affect how people would later act in very many other cases of this kind. Though you would die, your choice would indirectly save at least a million other people. So, even on these more egoistic views, you would have sufficient reasons to give up your life to save these very many other people.

This case is only one example. But if, as I believe, you could rationally choose this optimific principle even at the cost of your 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. [483]

Suppose next that I am mistaken. We ought, I have claimed, to reject Rational Egoism. But there is another, more plausible view that is relevant here. On this view, though 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 when, as in *Lifeboat*, this burden would be as great as dying young, and thereby losing many years of happy life. You could not rationally choose the Numbers Principle, because you could not rationally choose to die, however many other people's lives your 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 argument that we are now considering, but also to the Kantian Contractualist Formula. Just as you could not rationally choose any principle that required or permitted me to save the five, the five could not rationally choose any principle that required or permitted me to save you. In this and other such cases, there would be no principle that everyone could rationally choose, so there would be no principle that, according to the Kantian Formula, everyone ought 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 clearly the wrong 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 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 between principles that apply to such cases, they would be rationally required to give absolute priority to the saving of their own lives, so there would be no principle that all these people could rationally choose. The Kantian Formula would fail, when applied to such cases, because these people's choices would not be even weakly impartial, but would be wholly self-interested. To avoid this objection, we could revise this formula. When we apply the Kantian Formula to cases of this kind, we might appeal to the principles that the people in these groups could rationally choose from an impartial point of view. Or we might partly follow Rawls, by adding a *local* veil of ignorance. We would then ask which principles these people could rationally choose if they did not know whether they were in the smaller or the larger group. On both these versions of the Kantian Formula, these people could 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

telling us to suppose that *all* principles would 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 various ways, some of which I mentioned when discussing Rawlsian Contractualism. 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 very 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, everyone 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 these ways.

54 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.

It might be claimed 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 you to act in this way. This claim may 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 one way 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 some 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.

The objection that I have just rejected could, however, be transferred 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 stronger reasons to choose that I accept some principle that would require me to save your child.

You would have such stronger 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. Nor can we appeal to the *deontic* reasons that might be provided by the wrongness of any of the acts that we are considering. It is worth repeating why we cannot appeal to such beliefs and such reasons. If we claim that

> (1) 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

> (2) we could not all rationally choose any such principle because such acts are wrong.

It would be similarly pointless to claim both that

> (3) everyone ought to follow certain principles because these are the only principles that everyone could rationally choose,

and that

> (4) 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, we would be going round in circles, getting nowhere. So, when we apply the Kantian Formula, we must ignore our beliefs about which acts are wrong. We can appeal to these beliefs only at a later stage, when we have worked out what this formula implies, and we are asking whether 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. It may be hard not to be influenced by our beliefs about your duties to your child. So it will help to change our example. Suppose that, in

> *Fourth Lifeboat*, I could save either the person whom you love most or five other people. These six people are all strangers to me.

Any optimific principle would require me to save the five other people rather than your most-loved friend. 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 instead that I accept some principle that would require me to save your most-loved friend.

Though this claim has some plausibility, it is not, I believe, true.

It may seem absurd to deny that you would have decisive reasons to make the choice that would save the person whom you love most. 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. . . [484]

I am not, however, appealing to the kind of impartial system that Williams here movingly rejects. First, on the optimific principles that we are considering, we are often morally permitted or required to give strong priority to the well-being of those people to whom we have close ties. These principles would not require us to save several strangers rather than one of our children, or someone else whom we love.

Second, in arguing that we could all rationally choose that everyone accepts the optimific principles, I am not assuming that

we are rationally required to give equal weight to everyone's well-being. My argument allows that we are often rationally permitted to give strong priority to our own well-being and the well-being of those to whom we have close ties. This argument assumes only that we would also be rationally permitted to give significant weight to the well-being of strangers.

As I have argued when discussing *Lifeboat*, your most-loved friend could rationally choose that everyone accepts some optimific principle. Though your friend would then die, her choice would indirectly save very many other people's lives. That would give your friend 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, that 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. Perhaps we could not have sufficient reasons choose to save five, or fifty, or even five hundred other people rather than the person whom we love most. But if, in our imagined case, you chose that everyone accepts some optimific principle, your choice would indirectly save a much greater number of other people's lives. However much you love your friend, you would also have sufficient reasons, I believe, to make the choice that would save these very many other people. If that is true, these cases provide no 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 your friend. That is clearly an unacceptable conclusion. This objection is like the one that appeals to High Stakes Egoism. Love can be a kind of egoism on someone else's behalf. When applied to such cases, the Kantian Formula would fail because some people's choices would not be even weakly impartial. To avoid this objection, we could revise this formula so that these people would make their choices from an impartial point of view, or from behind a local veil of ignorance. But, when the stakes are significantly lower, we could still appeal to the unrevised formula. 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 point of view, everyone could rationally choose. And, as before, I believe that we do not need to make this revision.

55 Other Non-Deontic Reasons

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.' [485] 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 would 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, they would 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. We may be forced to choose between 'perfection of the life, or of the work'. [486] 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 typescript of my great unfinished novel or the only typescripts of five similarly great unfinished 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 books rather than mine. But these reasons would not, I believe, outweigh my impartial reasons to choose this principle. I could rationally give up my book to save five other similarly great books. If my belief were mistaken, we could again revise the Kantian Formula so that these choices would be impartial. That would make little difference, since such cases would be rare.

There is another, more important possibility. Suppose that some optimific principle would require us to do something that we believe to be wrong. When we apply the Kantian Formula, as I have said, we cannot appeal to our belief that certain acts are wrong, nor can we appeal to the *deontic* reasons that the *wrongness* of these acts might provide. But we can appeal to the facts that, in our opinion, *make* these acts wrong. And we might claim that

> (K) these wrong-making facts would give us decisive *non-deontic* reasons not to choose any optimific principle that would require us to act in these ways.

To illustrate this claim, suppose that, if we acted in certain ways, we would be injuring, deceiving, and betraying certain other people for our own convenience. As well as making these acts wrong, these facts about these acts might always or often give us decisive non-deontic reasons not to act in such ways. Some of these reasons would be provided by the ways in which it would be bad for these other people to be injured, deceived, and betrayed, and be bad for us to be someone who acts in such ways. These facts might also give us decisive non-deontic reasons not to choose any principle that would require such acts.

The optimific principles would not require us to injure, deceive, and betray others for our own convenience. But these principles would require some acts that many people believe to be wrong. These principles would, for example, require some people to use artificial contraceptives. If we believe that such acts are wrong, could we also claim that we had decisive *non*-deontic reasons not to choose any principle that would require such acts? The answer is clearly No. If it would not be wrong to use artificial contraceptives, we would have no strong reasons not to act in this way, and no strong reasons not to choose any principle that would require such acts.

Consider next hastening our death to avoid suffering, and lying to some would-be murderer to protect his intended victim. The optimific principles would permit many such suicides, and would require all such lies. We might have decisive reasons not to act in these ways if, as Kant believed, such acts are wrong. But if such acts are not wrong, we would have no strong reasons not to act in these ways, and no strong reasons not choose any principle that would permit or require such acts.

For another example, remember that, in

> *Bridge*, a runaway train is headed for the five. If I caused White to fall in front of the train so that White's body would trigger the train's automatic brake, I would kill White but would thereby save the five. There is no other way in which anyone could save the five.

Suppose that, as I have claimed, the optimific principles would require me to act in this way. Suppose next that we believe both that such acts are wrong, and that what makes them wrong is the fact that they involve

> (1) killing some innocent person as a means of saving several other people.

We might then claim that

> (2) if some act is of the kind described by (1), this fact gives us a decisive non-deontic reason not to choose any principle that would require us to act in this way.

To assess this claim, we might suppose that it would *not* be wrong for me, in *Bridge*, to kill White as a means of saving the five. We might then ask whether, if this act would not be wrong, I would have a decisive non-deontic reason not to act in this way. But this question may be hard to answer, since we may find it hard to suppose that such acts are not wrong.

It may help to turn to cases in which we have changed our mind about whether some act is wrong. Suppose that, in

> *Bomb*, I could save the five by throwing a small bomb which would explode in front of the train. This bomb would also kill White.

Many people would believe this act to be wrong. When they considered such acts, some people have appealed to the claim that the negative duty not to kill has priority over the positive duty to save people's lives. These people believed that

> (3) it would be wrong to act in a way that saves several people's lives, if this act would also kill some innocent person.

These people might have claimed that

> (4) if some act is of the kind described by (3), this fact gives us a decisive non-deontic reason not to act in this way.

But this claim, I believe, is false. Remember that, in

> *Tunnel*, I could redirect the runaway train onto another track so that it would kill White rather than the five.

This imagined case has been much discussed, though it has little practical importance, because it seems a counter-example to the widely accepted view that the duty not to kill has priority over the duty to save people's lives. When they considered cases like *Tunnel*, many of those who held this view changed their mind, since they ceased to believe (3). On the view to which these people turned, we are morally permitted to redirect some kinds of threat to people's lives, such as a runaway train, or an avalanche, or flood, if we would thereby make this threat kill fewer people. [487] I would be morally permitted to redirect the train in *Tunnel*, even though my act would save the five in a way that would also kill White.

When these people changed their view, by ceasing to believe (3), they would have rightly rejected (4). The fact that my act would kill White in *Tunnel* gives me a strong non-deontic reason not to act in this way, since it is awful to kill an innocent person. This may be why many people believe that I would be only morally permitted, and not morally required, to act in this way. But this reason would not be decisive. If I would be morally permitted to redirect the train, the fact that I would be saving

several people's lives would give me sufficient reasons to act in this way.

Similar claims apply to *Bridge*, in which I could save the five only by killing White. The fact that I would be killing White as a means may give me a decisive reason not to act in this way. But this fact would give me such a reason, I believe, only if and because this fact would make this act wrong. This fact would not give me a decisive *non-deontic* reason. We may find it hard to assess this last claim, because we may find it hard to imagine that killing someone as a means is not wrong. But there is a very close analogy between claims (2) and (4). If we ought to reject (4) in cases like *Tunnel*, as I and many others believe, we ought also to reject (2).

Return now to our main argument. When we apply the Kantian Contractualist Formula, we ask which are the principles whose universal acceptance everyone could rationally choose. We would all have strong impartial reasons to choose the optimific principles. According to premise (E), these reasons would not be outweighed by any conflicting reasons. We are now supposing that some optimific principle requires some act that we believe to be wrong. When we apply the Kantian Formula, we cannot appeal to our belief that such acts are wrong. But we might appeal to the facts that, in our opinion, *make* such acts wrong. According to the objection we are now considering,

> (K) these wrong-making facts would give us decisive *non-deontic* reasons not to choose any optimific principle that would require us to act in these ways.

There are, I have said, many wrong-making facts that give us decisive non-deontic reasons not to act in certain ways. But I suggest that

> (5) if the optimific principles would require some kind of act that we believe to be wrong, the facts that, in our opinion, make such acts wrong would not *directly* give us decisive reasons not to act in this way. These facts would give us such reasons only if, and because, they would make such acts wrong.

We should expect (5) to be true. If the optimific principles require some kind of act, we must all have strong impartial reasons to want everyone to act in this way. If we did not have such reasons, the optimific principles would not require such acts. If these acts would be wrong, that might give us decisive reasons not to act in this way. But, if these acts would *not* be wrong, and we would all have strong impartial reasons to want everyone to act in this way, we should not expect to have decisive *non*-deontic reasons not to act in this way.

Of the facts that could be most plausibly claimed to give us such

reasons, one example is the fact that some act would kill some innocent person. If even this fact would *not* give us such a reason, as I have just argued, we should expect the same to be true of the other facts that can make acts wrong, such as the facts that can make it wrong to tell certain lies, steal certain things, or break certain promises. [488] When everyone would have impartial reasons to want us to act in these ways, these other facts would give us decisive reasons only if, and because, they would make these acts wrong. If that is true, as I believe, we can reject (K). I defend this belief further in Appendix H.

There is, I believe, no other strong objection to premise (E). 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, and could therefore rationally choose, that everyone accepts the optimific principles.

56 What Everyone Could Rationally Will

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, that would make things go in ways that would be impartially much worse than the ways in which things would have gone if everyone had accepted the optimific principles. In nearly all such cases, things would also go much worse for some unfortunate people. [489] These people could not rationally choose that everyone accepts this non-optimific principle, since they would have both strong impartial reasons and strong personal reasons not to make this choice. In *Earthquake*, for example, Black could not rationally choose any principle that required me to save Grey's leg rather than Black's life. And, in *Lifeboat*, none of the five could rationally choose any principle that required me to save you rather than saving all of the five. So, as (G) claims, there are no significantly non-optimific principles whose universal acceptance everyone would have sufficient reasons to choose.

(B), (F), and (G) together imply

> (H) It is only the optimific principles whose universal acceptance everyone could 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 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. We can call this the *Kantian Argument* for Rule Consequentialism, or the *First Convergence Argument*.

This argument, we may suspect, must have at least one consequentialist premise. If that were true, this argument would be uninteresting, and unimportant. 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 the goodness of outcomes. This argument also assumes that there are some principles whose universal acceptance would make things go best. But this assumption is not consequentialist. We could believe that there are such optimific principles, but also believe that some of these principles are false, since 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 a 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 they were choosing in some way that would make these choices sufficiently impartial. Some contractualists claim that, to achieve such impartiality, it is enough to appeal to what 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 she had the power to choose which principles we would all 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 any consequentialist premise, this argument has a 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 consequentialist conclusions.

CHAPTER 16 CONCLUSIONS

57 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? If the answer is Yes, the Kantian Contractualist Formula requires us to be Act Consequentialists.

As Sidgwick argued, AC is not in this sense optimific. [490] If we were all Act Consequentialists who always tried to do whatever would make things go best, these attempts would often fail. When predicting the effects of different 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 also, and most importantly, be bad in other ways. For it to be true that everyone accepted and always tried to act on AC, most of us would have to lose too many of the strong loves, loyalties, personal aims, and other motives that make our lives worth living. For these and other similar reasons, we can claim that

> (L) 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 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 or require us to give some kinds of strong priority to the well-being of our close relatives and friends, and of those people to whom we are related in various other ways, such as our pupils, patients, clients, colleagues, customers, 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.

The Kantian Argument that I have defended could be revised to show that

> (M) 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

> (N) 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. [491]

(N) 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 this world would differ from 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. [492] For some of the reasons Sidgwick gave, we can claim that

> (O) 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. [493] 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. If (N) is not true, this version of the Kantian Formula does not require us to be Act Consequentialists.

As I have claimed, however, this formula does require us to follow the principles that are UF-optimific. And compared with the UA-optimific principles, these UF-optimific principles are more similar to AC. [494] 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 Kantian Contractualism and Rule Consequentialism 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. 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 about people's motives, desires, and dispositions. And we can ask which motives we ought to have, and what we ought to be disposed to do. This 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 versions, and use these different senses of 'ought' and 'wrong', in such different parts of our moral theory. [495]

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. These possibilities may raise some problems. Though I think it likely that such problems could be solved, I shall not discuss them 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 other 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.
>
> (P) 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.
>
> (Q) 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 what I earlier called its most unified and harmonious form.[496] 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 followed would best achieve this aim.

Of this argument's premises, KC is Kantian Contractualism. My defence of (H) above could be turned, with some revisions, into a defence of (P). (Q) is Kant's description of what he calls the Greatest Good.

We ought, I have argued, to revise (Q). It would be bad, Kant claims, if people had more happiness than they deserve. And some of Kant's claims imply that some people deserve to suffer, and that it would be bad if such people suffered less than they

deserve. But Kant also claims

> (R) 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

> (S) All of our decisions *are* merely such events.

Therefore

> (T) No one could deserve to suffer.

If we subtract Kant's claims about desert, Kant's ideal 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 more happiness, but also if more of this happiness came to people who were less happy. We could also add that our well-being does not consist merely in happiness and avoiding suffering, and that the goodness of different states of the world would in part depend on some other facts.

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. [497] 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, that is bad for both the murderer and his victim, and is 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 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. [498]

As I have said, 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. [499]

This claim is a hedonistic version of Rule Consequentialism.

I shall now sum up several of my claims. 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.
>
> (U) There are certain principles whose being universal laws would make things go best.
>
> (V) 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 (U) is true. I have, I believe, successfully defended (V). 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.

58 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, there must be stronger 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 might call this *the Second 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.

Like Kantian Contractualism, Scanlonian Contractualism can take different forms, since there are different views about what are reasonable or admissible grounds for rejecting some proposed moral principle. On some of these views, we could reject at least one of this argument's premises. But this argument shows, I believe, that at least one version of Kantian Contractualism could be combined with at least one version of Scanlonian Contractualism. It is a further question whether these would be the best versions of these theories. I discuss that question further in my response to Scanlon's Commentary below.

This combined theory, as I have argued, could 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. [500]

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 property of being a lying promise or causing pointless suffering. 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 reasons not to act in these ways.

If we accept this 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 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 steal some object that its owner never uses, or fail to help some people who are in great need, or add our bit to pollution, 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, one of the only principles whose being universal laws everyone could rationally will, and 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'. [501]

For some moral theory to succeed, it must have plausible implications. The Triple Theory has many such implications. But we might find that, after we have carefully considered all the

relevant facts and arguments, this theory still conflicts with our intuitive beliefs about the wrongness of certain acts. If these beliefs are very strong, or such conflicts are quite common, we could then justifiably reject this theory. But if these conflicts are significantly less deep, or less common, we could justifiably revise these intuitive beliefs. [502]

We have 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 some principle is unreasonable, it could not be justified. If Scanlon's Formula seems undeniable, however, that is because it 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 false 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 false, because such acts are not wrong. And, in making such a claim, we could appeal to our intuitive beliefs about which acts are wrong. It is only *while* we are applying some contractualist formula that we cannot appeal to these beliefs. Though Scanlon's Formula is in itself very plausible, we could justifiably reject this formula if its implications conflicted too deeply with some of our other moral beliefs.

Though Kantian and Scanlonian Contractualism could be combined, that might not be true, I have said, of the best versions of these theories. If these best versions could not be combined, we would have to choose between them.

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.[503]

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, 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 with these 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.[504]

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 are the *principles* that 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 first 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 to try to achieve. That is what most of us believe, unless we have been taught otherwise by some philosopher, economist, or other social scientist.

I later argued that, with some revisions and additions, Kant's most important claims are these:

> Everyone ought to treat everyone only in ways to which they could rationally consent.
>
> 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.
>
> Everyone ought to try to promote a world of universal virtue and happiness.
>
> 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.
>
> 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 have strong reasons, I believe, to accept these claims.

I shall not try to summarize my other claims. But I shall end by mentioning one way in which some of these claims have seemed to me worth defending.

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 belief that all questions must have answers. But some disagreements are not of these kinds. 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.

COMMENTARIES

RESPONSE TO SUSAN WOLF

1 Actual and Possible Consent

Susan Wolf makes several claims that seem to me both true and important. And we disagree, I believe, less deeply 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 involves treating people in ways to which they *actually* consent. But I do not abandon this idea. I claim that many acts 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 can plausibly appeal to

> *the Rights Principle*: Everyone has rights not to be treated in certain ways without their actual consent. [505]

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. I argue that Kant means

> (E) It is always wrong to treat people in ways to which they could not rationally consent, if we gave them the power to choose how we treat them.

This claim is plausible and might be true. I call (E) the *Principle of Possible Rational Consent*, or for short (but perhaps misleadingly) 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 easily 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. And this principle, I argue, never conflicts with the Rights Principle, which often requires explicit consent. 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.

Kant's claims about consent give us what I call *Kant's ideal*. We cannot always treat people only in ways to which they do in fact, or would in fact, consent. But we might be able to treat everyone only in ways to which they could rationally consent. I argue that, on some plausible assumptions, we could achieve this ideal, and that, at least in most cases, this is how we ought to act. Since Wolf does not discuss these claims, she may not share my high opinion of this part of Kant's view.

2 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. [506]

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 this act would be wrong, since only the saving of a child could justify imposing such an injury on someone else. Acting on these beliefs, you save your child by causing Black to lose only 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', such claims seem to me clearly 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 second 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 an injury as great as destroying one of Black's legs. Your act is wrong, we can add, even though you are *not* treating Black merely as a means.

Return next to

> *Bridge*, in which I could save five people's lives by using remote control to cause *White* to fall in front of a runaway

> train.

Wolf claims that this act would 'very definitely' treat White merely as a means. In some versions of this case, I argued, I would *not* be treating White merely as a means. But this fact, I also claimed, would not justify my 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 is *harder* to defend the belief that such acts are wrong. On Wolf's view, it is not enough to appeal to the claim that such acts harm certain people as a means, since we must also defend the further claim that these acts treat these people merely as a means. On my view, to condemn harming people as a means, we need not defend that much more doubtful claim.

3 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. Wolf claims that such acts may be wrong 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 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 some autonomy-infringing principles would be optimific, since their acceptance would save more people from death or other burdens. Wolf also claims that

> (F) everyone could rationally choose that everyone accepts some *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 optimific Rule Consequentialist principles.

To assess this objection, we can again suppose that in

> *Tunnel*, I could redirect some runaway train so that it kills White rather than five other people.

Wolf's autonomy-protecting principles would condemn my saving the five in this way, since this act would impose a great burden on White. 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. 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 some autonomy-protecting principle, I 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 me 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 would be *optimific*, since this would 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 me not to save their lives. But Wolf might be right to claim that the five could rationally choose this autonomy-protecting principle. This principle might be optimific.

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 other than mine. She 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 reducing 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 would 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*, I would be morally permitted to redirect the train so that it kills White rather than the five. But Wolf does discuss *Bridge*, in which I could save the five only by killing White.

Most of us would believe that, in *Bridge*, it would be wrong for me to save the five by killing White. According to Wolf's autonomy-protecting principles, it is wrong to impose great burdens on people without their consent. These principles do not distinguish between *Tunnel* and *Bridge*. In both cases, if I save the five, my act would impose a great burden on White, by killing her without her 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, White and the five would all have strong reasons to prefer to be the ones calling the shots, deciding what would happen to their bodies, and whether they would live or die.

If we believe that my saving the five would be wrong in *Bridge* but not 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 I save the five, my act would also kill White, but only in *Bridge* would I be killing White as a *means* of saving the five.

I claimed that, in *Bridge*, the optimific principles would require me to save the five by killing White. Wolf questions this claim.

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'. This suggestion might be correct. As Wolf claims, it can be hard to predict whether some principle would be optimific. When I discussed *Transplant*, I made similar claims. 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 my saving the five by killing White, 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, because this view would not conflict, in such cases, with our moral intuitions.

4 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 normative 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. ' [507]

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.[508]

Even if we hope that Kantian Contractualism does not imply Rule Consequentialism, my argument for this conclusion may be sound, or valid.

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, which claims that we have rights not to be treated in certain ways without our actual consent. For some of the reasons 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 me to save the five by killing White. As we have seen, Wolf questions this claim, since she suggests that the optimific principles might condemn such acts. But, 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 require me, in *Tunnel*, to redirect the runaway train so that it kills White rather than the five. Like most other people, I can accept that conclusion. But this formula would also require me, in *Bridge*, to save the five by killing White. And, like Wolf, I find this claim implausible. Intuitively, this act seems to me wrong.

This intuition is not, however, very 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 White to be killed as a means. And the Kantian Formula provides an argument against this intuition. If we were choosing the principles that, in such cases, everyone would follow, we would have more reason, I believe, to choose principles that required me to save the five. That is most clearly true if we did not know in whose position we would be, since this choice would then be more likely to save our lives. Though I am still inclined to believe that it would be wrong for me to kill White as a means, this intuition is not strong enough to lead me to reject the Kantian Formula. /[509]

We have strong reasons, I believe, to accept this formula, and to act on the optimific principles of Kantian Rule Consequentialism. But there might be other cases in which this moral theory

conflicts more strongly with my and other people's moral intuitions. If that were true, our reasons to accept this theory might be outweighed by stronger conflicting reasons.

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.

I agree that we do not need a single supreme principle. But we do, I believe, need a *single true morality*. My main aim is not to find a supreme principle, but to find out whether we can resolve some deep disagreements. As Wolf claims, it would not be a tragedy 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 our belief that there are such principles might be an illusion.

RESPONSE TO ALLEN WOOD

1

I have learnt a great deal from Allen Wood's books and articles, and I am delighted 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. [510] 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. [511]

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. [512]

Since I have tried to show that Kant's Formula of Universal Law may provide a 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 that is not so. I accept Wood's claim 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, so that it applies not to maxims but to principles. According to one version of 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. [513]

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 presents as 'the idea of the will of every rational being as a will giving universal law. ' [514] 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. [515]

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.' [516] But, as Wood also claims, FA does not give us a reliable criterion of wrongness. [517] If we ask which are the principles that we *would in fact* choose, we cannot assume that everyone would choose the same principles. Nor could we predict which principles other people would choose.

We ought, however, to revise FA, so that this formula refers to the principles that it would be *rational* for 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: Everyone should try to follow the principles that it would be rational for everyone to choose as the principles that would govern everyone's conduct.

This claim states another version of my Kantian Contractualist Formula. [518] 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 would not be surprising if, as Kant claims, these different 'ways of representing the principle of morality are, fundamentally, only so many formulas of precisely the same law'. [519]

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 such criterion, whether some act would be wrong. Wood does not claim either (1) or (2), nor I believe is either claim true. So Wood, I believe, should agree that it is important whether Kantian Contractualism provides a good criterion of wrongness. And that, I have argued, may be true.

2

I turn now to Wood's discussion of Kant's Formula of Humanity. When Kant applies this formula, he claims it to be wrong to treat people in ways to which they could not rationally consent. This claim, I argued, is both defensible and worth making. I am glad that, in his commentary, Wood seems to agree.

Though Kant's Formula of Humanity rightly implies that it is wrong to *regard* anyone merely as a means, whether our *acts* are wrong, I argued, 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. [520]

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 claimed, is too vague to provide a criterion of wrongness. Wood agrees. [521]

Unlike me, however, Wood believes that FH is the most important of Kant's formulas. 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. . . [522]

Kant's claims about this value, Wood suggests, describe 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*: 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 which gives morality its 'end or point', thereby showing how morality's demands on us 'make sense'. [523]

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. [524] 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'. [525] Wood calls rational nature 'the underivative objective good'. [526] Kerstein similarly writes that humanity is 'absolutely and incomparably good', [527] and Korsgaard writes that, on Kant's view, humanity must be treated 'as unconditionally good'. [528] [529]

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*. [530]

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, and would not have been Kant's view, 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 always to be treated only in these ways.

This version of Wood's Thesis does not 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 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. [531]

These kinds of rationality, Wood writes, have 'the absolute worth that grounds morality'. [532]

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. [533]

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. [534]

Wood even suggested that, on Kant's view, the 'rational choice of ends is the act through which objective goodness enters the world'. [535]

This cannot, however, be 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 of reasons and of the goodness of our ends, and he calls these views 'good Kantianism'. [536]

Our non-moral rationality may have some kinds of value, to which I shall return. But such rationality cannot be claimed to have supreme goodness of a kind that could ground 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. [537]

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. [538]

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. [539]

Wood's Foundational Thesis might appeal instead to the *exercised* capacity to give and obey moral laws, which is roughly what Kant calls a 'good will'. Kant claims that such wills are supremely good. So Wood might 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. [540] 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. ' [541]

This argument is not, I believe, sound. For something's goodness to give us a reason for acting, this thing need not exist.

Many of our acts are intended to achieve some merely possible good end. So Kant might claim that

> (4) the supreme goodness of good wills gives us our reason to try to have such a will, and to act rightly. [542]

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. [543]

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'.

3

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 wrong acts 'all involve treating that value. . . with disrespect'. [544] If the value of rational nature was the goodness of good wills, these claims would not be plausible. 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 supreme goodness of good wills. From Kant's claim about the goodness of such wills 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. [545] Kant's ethics is grounded, *not* on the goodness of good wills, but on what Wood calls the 'absolute worth of rational nature'. [546]

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 could agree that, on Kant's view, morality is grounded on the goodness of good wills and deserved happiness. 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.

This value 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 people have this status does not help us to decide how people 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'. [547] On Kant's view, Herman similarly writes, 'the domain of "the Good" is rational activity and agency: that is willing'. [548]

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, is this Kant's view. On Kant's view, as Herman notes, what is good is only *good* willing. [549]

Even if rational agency is not supremely good, such agency might be claimed to have what Wood calls 'the basic value to be respected'. [550] 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. [551]

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 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. [552]

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. . . [553]

Rational agency seems here to be regarded as having the kind of value that some people claim for chastity, and self-defence to be regarded as like the protection of our chastity. I doubt that this is really either Kant's or Herman's view. Aggressive violence *is* wrong, not because it devalues rational agency, 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 reasons to want not to be in.

Consider next these 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. [554]

> the fundamentally valuable thing. . . is a rational being, a person – or, more precisely, rational nature in a person.

Both these more precise claims are, I believe, mistaken. We should 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*. [555]

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. [556] 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*.

RESPONSE TO BARBARA HERMAN

In her some of her brilliant discussions of Kant's Formula of Universal Law, Barbara Herman claimed that this formula cannot provide a criterion of wrongness. [557] Despite 'a sad history of attempts', she wrote, '. . . no one has been able to make it work'. [558] Herman, I have argued, was right. In its present form, Kant's Formula cannot succeed. But, if we revise this formula, we may be able to make it succeed.

In her commentary, to my surprise, Herman seems to argue that Kant's Formula does not need to be revised. [559]

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. [560] And my imagined ruthless gangster acts wrongly, she suggests, when he buys a cup of coffee.

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 is 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. 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 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 also 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'.

These claims could be easily misunderstood. Duties of justice *are*, on Kant's view, moral duties. As Kant writes

> all duties, just because they are duties, belong to ethics. [561]

When Kant claims that our failure to fulfil duties of justice makes our acts 'illegal', he does not mean that such acts are against the criminal, state-based law. He means that such acts are against the *moral* law. Kant 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 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.

So Herman might claim that, even when some act is morally permissible and in conformity with duty, this act might be in these other senses wrong.

If we distinguish these senses of 'wrong', my argument could become:

> (D) When my Egoist pays his debts, keeps his promises, 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 next 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. [562] 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? [563]
>
> 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. [564]

As these and many other passages together show, Herman cannot defensibly reject premise (E). [565] Since this argument is sound, Herman should accept my conclusion. 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 Herman's claims may 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 Kant accepted the maxims 'Never steal' and 'Never break my promises'. Kant could not have rationally willed it to be true that no one ever steals or breaks some promise, even when these are the only ways to save some innocent person's life. So Kant's Formula would imply that, whenever Kant acted on these maxims, by returning someone's property or keeping some promise, he would be acting wrongly. These claims are also clearly false. As before, to avoid this objection, Kant's Formula must be revised.

2

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 morally relevant description of the acts that we are considering.

Herman does not discuss my proposed revisions. But some of Herman's claims 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 more truly Kantian version of Kant's Formula, which we can call

the Second 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 senses, wrong.

As we have seen, Herman suggests 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 Second 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, or keeps some promise, 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 these acts moral worth. It would be irrelevant whether Kant was acting on some maxim, such as 'Never lie' or 'Never break my promises', which he could not rationally will to be universal. So the Second 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 does Kant act inappropriately when he acts on the maxim 'Never lie' by telling someone the correct time of day. So the Second 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 always 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 admits, we can fulfil these duties whatever our motive.

Given his self-interested motive, my Egoist could not fulfil some duties of virtue. And, if he acted wrongly, he would also be acting for the wrong motive. But the Second Formula should not 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 their conscience, or obey some sacred text, or do what their parents told them to do, they are acting on incorrect principles, or using unsound moral reasoning. 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 acts on the maxim 'Never Lie' by telling someone the correct time of day, Kant's Formula falsely implies that this act is wrong. Herman might reply that Kant's act *is* in one sense wrong, since this act is 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.

But, as Herman here claims, 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. [566]

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 good sense in which it was wrong for this gangster to pay for his coffee.

In the passages that I have just been discussing, Herman well describes some of the ways in which it matters 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.

3

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 her 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.

As before, 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 Kant's

> *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 one statement of 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. As I argued, for example, 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 stating 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.[567]

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:

> (H) According to Kant's Formula, we act wrongly just when we act on some maxim that fails a certain test.

Therefore

> (I) Kant's Formula implies that, if some maxim fails this test, it is always wrong to act upon it, and that, if this maxim passes this test, it is always permissible to act upon it.
>
> (J) 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 what is involved in some maxim's being unable to be willed to be universal. 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 my promises'. 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. [568]

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 of whether this formula implies that lying is always wrong. She 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 people's lives. But Herman objects to the way in which I would reach this conclusion. She sums up my way of 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 outcome is non-morally best, in what I call the impartial-reason-implying sense, just when this is the possible outcome that, from an impartial point of view, everyone would have most reason to want. 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

> (K) 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

> (L) 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 (K). For example, she writes

> states of affairs are not possible bearers of value in Kantian ethics. [569]

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. [570]

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. [571]

We could all accept *these* claims. As Kant rightly remarks when discussing the Stoic, it is not morally bad to be in pain. But pain can be bad in a different, non-moral sense, since pain is a state that we all have non-moral reasons to want not to be in. And, as I have claimed, outcomes can be non-morally good or bad, or good or bad for people, in such reason-implying senses.

Herman seems to accept my claims 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. [572]

And she writes that we could not rationally

> will a world where one's life can have no value in this reason-giving sense. [573]

If we accept some desire-based or aim-based subjective theory about reasons, I have argued, we cannot claim that we all have reasons to care about our own and other people's well-being. But Herman seems to rejects such theories, and to assume that we have such reasons. [574]

Herman also seems to believe, however, that when we apply Kant's Formula we should not appeal to such reasons. When she describes my way of applying Kant's Formula, she 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.

Our next question is: Why not? When we apply Kant's Formula, we ask whether we could rationally will it to be true 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?

In considering this question, we can turn to Kant's discussion of his imagined rich and self-reliant man, whose maxim is not to help others when they 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. [575]

Kant seems to be 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 when discussing Kant's claim:

> It is surely no crude mistake. . . to interpret this passage as making some kind of prudential appeal.

But Herman then argues that, in 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 be better for him, since 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. She claims that, if that is how we apply Kant's Formula, 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 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 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. [576]

This objection does not, I believe, 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. [577]

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 which 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). [578]

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'.[579]

If this argument succeeded, that would show only that, according to Kant's Formula, it is wrong never to help others who are in need. That 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 argued, many actual cases in which Kant's Formula clearly fails.

The most important cases raise what I called the Non-Reversibility Objection. This objection can best 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 could not rationally want other people to 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 take 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 rationally will it to be true both that they and other rich people continue to give nothing to the poor, and that everyone, including the poor, believes that their failure to give 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 slave-owners, or to those men who treat women as inferior, denying them certain rights and privileges, and giving less weight to their well-being. These men could rationally will 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. As I have claimed, there are many cases in which those who act wrongly could rationally will it to be true that their maxim is universal. To avoid these objections, Kant's Formula should appeal instead to what everyone could rationally will.

Return now to our main question. 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 good or bad for ourselves or others, and to our reasons to care abour our own and other people's well-being. My appeal to such reasons, Herman claims, makes my proposed Kantian Contractualism a 'hybrid reconstruction', which departs too far from the best elements in Kant's view.

That, I believe, is not true. When Kant applies his formula, he appeals to our reasons to care about our own well-being. And, as we have seen, so does Herman in her proposed Kantian argument for beneficence. Though Herman suggests that, when we apply Kant's Formula, we should not appeal to such reasons, claiming that 'such considerations cannot be legitimately introduced', she says nothing to defend the claim that appeals to such reasons are not legitimate. And this claim clearly needs some defence. When we ask whether we could rationally will it to be true that some maxim is universal, why should we ignore our reasons to care about our own and other people's well-being, and our other non-moral reasons to care about what happens? [580]

Return now to Herman's claim that Kant's Formula of Universal Law cannot provide a criterion of wrongness. Despite 'a sad history of attempts', Herman wrote, '. . . no one has been able to make it work.' I have argued that, if we revise Kant's formula, we *can* make it work. Herman, I hoped, would agree that this history of sad attempts has a happy ending.

My hope was dashed. Herman rejects my attempt to make Kant's formula work, writing 'the mismatch of methods is too profound'. Herman cannot be objecting to my proposed revision, which appeals to the principles that everyone could rationally will to be universal laws. This revision is not un-Kantian, since it makes Kant's Formula of Universal Law closer to Kant's Formulas of Autonomy and of the Realm of Ends. Herman's objection seems to be that, if we apply this revised formula in a way that appeals to our non-moral reasons, we shall reach un-Kantian conclusions. Kant, we all know, was a deontologist, not a Rule Consequentialist.

This objection is not, I believe, decisive. Like many earlier writers, Kant seems to have assumed that it is by following various deontological principles that we can make things go best. On that assumption, as I have said, deontologists can even be *Act* Consequentialists. As I argue in a passage that Herman does not discuss, Kant's actual Law of Nature Formula, when combined with Kantian assumptions about rationality, permits some people to be Maxim Consequentialists. So we

should not be surprised to find that, when combined with other assumptions about rationality and reasons, my revised Kantian Formula requires everyone to be Rule Consequentialists. And, if we *are* surprised, that is no objection to my argument for this conclusion.

RESPONSE TO T. M. SCANLON

1 Giving Organs

Scanlon's Commentary starts with a very illuminating discussion of Kant's assumptions about what we could rationally will. Since I accept all of Scanlon's claims, I shall add only two brief remarks. According to what Scanlon calls 'Kantian constructivism', claims about reasons must be grounded in 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. I also believe that, for Kantian moral theories to succeed, they must appeal to 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. Such claims are 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.[581]

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 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.'[582]

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.[583]

Since it is GBC which raises this problem 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.

2 The Individualist Restriction

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. [584]

These reasons give each person what we can call *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 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.[585] 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'. [586] 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. [587]

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 moral difference, Utilitarians believe, when such 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.

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'. [588] 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. . . [589]

On the simplest form of Scanlon's Individualist Restriction, benefits to different people cannot ever be morally summed. In applying Scanlon's Formula to 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*. [590]

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. [591] Scanlon also suggests that, when one burden is not much smaller than another, the numbers count. Rather than saving one person's life, for example, it might be right to save many other people from being paralyzed. These suggestions raise some difficult questions, which are irrelevant here. To avoid such questions, 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 these people. 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 them 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 that we could give to Blue by 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 a mere 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 we treated Blue.

Though Scanlon's Formula may give the right 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 case in which there is no such difference, so that the *baseline* is *equal*. 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 treat Blue	0	100
We treat the others	100	90

If we do nothing, Blue and the others would all be equally badly off, since they would all have as much pain. According to Scanlon's Individualist Restriction, benefits to different people cannot be morally summed, so we ought again to treat Blue, 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 even to each of 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. In such cases, I believe, 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 even if only ten other people were saved from 10 of their 100 days of pain. The same total sum of benefits would then be shared much more equally between Blue and these other 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 ten other people were saved from 10 of their 100 days of pain. But this outcome would be better, not because there would be less 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. [592]

It may help to vary the example. Suppose that Blue and several other people are all aged 20, 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 extra years of life, whatever the number of the other people to whom we could instead give 5 extra 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 extra years of life rather than giving 5 extra years to each of ten, or twenty, or even more of these other people.

This claim is plausible, however, only because 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. Suppose that, in *Case Six*, our 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 treat Blue	70	30
We treat the others	30	35

On Scanlon's view, we ought to give Blue her extra 40 years of life rather than giving 5 extra years even to each of a *million* of these other people. As before, that is clearly false. And what makes it false is not merely that, compared with 40 extra years, 5 million extra years of life would be a vastly greater total sum of benefits. These benefits would also be much better distributed between different people. It would undeniably be better if, rather than Blue's living to the age of 70 rather than 30, a million other people 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. [593] In cases like *Four* and *Six*, rather than giving to a single person some great benefit, we ought to give a sum of smaller but significant benefits to many other people who are just as badly off.

These cases show, I believe, that Scanlon ought to drop his Individualist Restriction. /[594] 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 the cause of the problem. 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 view 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. These people must be able to appeal to the combined strength of their grounds to reject any principle that did not require us to give these benefits to them.

As these cases show, it is not only Utilitarianism that gives weight to the numbers of people who receive benefits and 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. [595]

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 way 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.

In his earliest statement of his theory, Scanlon appealed to this view, writing:

> when we consider a principle 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.[596]

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. [597] 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 Blue from some great burden, or give much smaller benefits to many other people. Scanlon claims that, in such cases, the numbers don't count, so that we

ought to save Blue 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, as we have seen, Scanlon's view also implies that, in *Case Four*,

> (B) we ought to save Blue from her 100 days of pain rather than saving each of 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 return to the view 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 the million other people are as badly off as Blue, facing the same great ordeal, these people's claims to be saved from their pain are as strong as Blue's. So they could reasonably reject any principle that did not require us to save them from a much greater total burden of ten million 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 responds 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.

There are other ways in which, if Scanlon dropped his Individualist Restriction, he would strengthen his contractualist theory. 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 only Grey 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 of us would believe.

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 would conflict even more deeply with most people's moral beliefs.

Scanlon's Formula would support these moral beliefs if Grey could reasonably reject every such 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. [598]

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 others to take Grey's organ without Grey's consent. 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 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* would 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 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 how 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 would rise by only a few days or hours. This benefit might be heavily 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 some other known particular people, these acts would also produce much anxiety, conflict, and mistrust. We may 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 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. Since 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, 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 extra years of life. [599]

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. 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. [600]

3 Scanlon's Claims about Wrongness and the Impersonalist Restriction

There are, I believe, some other ways in which Scanlon should revise

and thereby strengthen his contractualist theory.

In his book Scanlon claimed that, rather than describing the facts that can *make* acts wrong, his contractualism 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. [601]

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 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 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 a 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 that, when some act is wrong in his contractualist sense, that makes this act wrong in the justifiabilist, blameworthiness, and reactive-attitude senses. These four senses of 'wrong' are all definable abbreviations. So this version of Scanlon's Formula could be more fully stated as

> SF3: 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 this form. [602]

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.[603]

Of those who reject appeals to the goodness or badness of outcomes, some claim that there is no sense in which some outcomes can be worse than others. 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. [604]

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. [605]

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 his contractualist sense of 'wrong', 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, as I have said, Scanlon no longer claims that his formula gives an account of what it is for acts to be wrong. Scanlon now makes the much more defensible and important claim that, when acts are wrong in his contractualist sense, that makes these acts wrong in one or more other, non-contractualist senses. And Scanlon could not say, that, when we ask which acts are wrong in these other senses, claims about the goodness of outcomes are *by definition* irrelevant.

If Scanlon keeps his Impersonalist Restriction, he might claim that his formula describes one of the facts that can make acts wrong in these other senses. There are some facts that always make acts wrong. It is always wrong, for example, to torture others for our own amusement, and it may always be wrong to treat people in ways to which they could not rationally consent. Scanlon might similarly claim that, when acts are wrong in his contractualist sense, this fact always make these acts wrong in one or more non-contractualist

senses.

For Scanlon's view to take this form, however, he would have to claim that, when some act is in his sense wrong, this fact always has moral priority over facts about impersonal values, or the goodness or badness of outcomes. And Scanlon could not, I believe, defend this claim. What we owe to others could not, for example, have absolute moral priority over facts about the pain of non-human animals. We would be morally permitted or required to fail to keep some fairly unimportant promise, if we could thereby save several animals from great pain. [606] Nor could what we owe to others have such absolute priority over some of the other facts that can make outcomes better or worse. Some examples are the ways in which some economic or technological policies would make things go much worse in the further future. Suppose that, if we follow one of two energy policies, that would greatly lower the quality of life that would be had by people more than a century later, but because this policy would also affect who it is who later lives, this great lowering of the future quality of life would be worse for no one. Such policies may be clearly wrong. But in some such cases, this wrongness could not be fully explained in Scanlon's contractualist terms, if we can appeal only to *personal* grounds for rejecting principles. There may be no one who could reasonably reject every principle that would allow such policies, since there may be no person or group of people on whom the choice of such policies would impose any burden, or great enough burdens, or whose moral claims would be unmet in other relevant ways.

Consider next the *Retributive Principle* which requires us to give criminals the punishment that they deserve, even when such punishment would have no good effects. On Scanlon's present view, 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 this way. Scanlon rejects the Retributive Principle, I believe rightly. But Retributivists might reply that it would be in itself good if people get what they deserve. [607] In rejecting this reply, Scanlon might claim that what we owe to others has absolute priority over moral claims that appeal to the goodness of outcomes. 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 Scanlon could not defensibly claim that what we owe to others has absolute priority over facts about the goodness of outcomes, his view could take two forms. If he keeps his Impersonalist Restriction, Scanlon might retreat to the claim 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 we had other, strong enough moral reasons to act in these ways. This version of Scanlon's view might seem disappointingly weak. But that may not be true. Scanlon might 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. Scanlon's 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 stronger claim that acts are wrong in other senses *just when* they are wrong in Scanlon's contractualist sense. If that were true, Scanlon's contractualist theory would unify, and help to explain, all of the more particular ways in which some acts are wrong. This version of Scanlon's view, though making a stronger claim, would be less distinctive. I shall not try to decide here whether this would be the best version of Scanlon's view.

4 The Convergence Arguments

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 Second 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 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 the Individualist Restriction, and appealed instead to a stronger form of the Contractualist Priority View. We can here suppose that Scanlon also drops the Impersonalist Restriction, and that he makes claims, not about wrongness itself, but about what makes acts wrong. I shall ask whether my Convergence Argument succeeds when applied to this revised version of Scanlon's view, which I shall call the *New 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.

This argument shows, I believe, that the Kantian and New 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.

On these assumptions, the Kantian Contractualist Formula requires 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.

If (3) were true, Scanlon's Formula would require us to save Blue from this burden. Kantian and Scanlonian Contractualism would here conflict.

To assess these claims, we must first know 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. 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 Seven*, 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, in such cases, there must always 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, 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.' 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. And it is an open question, I believe, whether we could 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 to make moral mistakes. Given the technological developments of the last two centuries, such cases now have great importance. But, for two reasons, we can ignore such cases here. First, these cases raise some difficult problems which are not relevant to the question of whether Scanlonian Contractualism might conflict with Kantian Contractualism and Rule Consequentialism. If Scanlon's Formula would require us to ignore some such very small benefits or burdens, the same might be true of the Kantian Contractualist Formula and the optimific Rule Consequentialist principles.

Second, we shall be asking whether Scanlon's Formula might conflict with the optimific principles. So we need to consider cases in which we believe that 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 the Telic Priority View. That assumption makes this form of Consequentialism closer to Scanlonian Contractualism.

Return to *Case Three*, in which the 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 in all these cases, we should suppose that each day of pain is an equal burden. On the 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. 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 Five*, in which the 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, because their universal acceptance would make things go much worse. *Slightly* non-optimific principles raise some complications that are 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 everyone could rationally choose some optimific principle even if that choice would impose some great burden on them, I was discussing cases in which, by choosing such a principle, these people would indirectly save many other people from *similarly great* burdens. In *Lifeboat*, for example, if you chose the Numbers Principle rather than the Nearness Principle, you would die, but your choice would indirectly save many other people's lives.

In Rawls-Scanlon Cases, no such claim is true. If Blue chose 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 other people from *much smaller* burdens. So these are the cases in which it would be most plausible to claim that some people could not rationally choose the optimific principles.

We ought, I believe, to reject even this plausible claim. Return to *Case Three*, 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 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 a so 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 counter-example to my argument for Kantian Rule Consequentialism. For these and some of the other reasons that I gave in Chapter 15, 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 Second Divergence Argument, according to which Kantian Rule Consequentialism can be combined with Scanlonian Contractualism. When applied to these cases, my argument would in part be:

> (D) Since the Many could *not* rationally choose any Blue-protecting principle, there must be facts that give these people strong grounds 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, Scanlon's Formula would require us in such cases to follow these Blue-protecting principles. Scanlonian Contractualism would here conflict with Kantian Rule Consequentialism.

Is (G) true? We are supposing that, in *Case Three*, for the optimific principles to require us to benefit the Many rather than Blue, it would have to be true that 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 Five*, 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 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. [608]

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. [609]

These suggestions are, I believe, mistaken. Though the Priority View can take a contractualist form, it can also plausibly take a *telic* form, which makes claims about the goodness of outcomes. [610] 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 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, I believe, plausibly take a fairly strong form.

5 The Independence of Scanlon's Theory

Scanlon might, however, claim that, compared with this Telic View, the *Contractualist* Priority View can plausibly take an even *stronger* form. That might be enough to make (G) true.

Return for example to *Case Seven*, 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, 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 each of some vast 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 return here 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 Eight*, 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 gives 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. Scanlon's 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. /[611] /[612]

Suppose next that, in

> *Case Nine*, 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. Scanlon's 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.

Eight and *Nine* 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, Scanlon's Formula supports these principles in a stronger and decisive way. But suppose next that, in *Case Ten*, if people were given equal shares, the total sum of benefits would be smaller. In such cases, there might be some people who could not rationally choose the Principle of Equal Shares, since some unequal distribution would be much better for these people and be only slightly worse for others. 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 Cases Seven to Ten show, when we apply Scanlon's 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. Scanlon's 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 the outcome might be 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 example of a different kind, suppose that in

> *Case Eleven*, 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 principles 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 principles 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.

There are, I believe, several other ways in which Kantian and Scanlonian Contractualism may have different implications, some of which conflict. [613] There are also, I believe, some lesser ways in which Kantian Contractualism and Rule Consequentialism may have different implications. 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.

According to my Convergence Arguments, we could accept a Triple

Theory, which combines Rule Consequentialism with both Kantian and Scanlonian Contractualism. If may seem that, if these theories sometimes have different implications, my Convergence Arguments fail, and we must reject the Triple Theory.

That is not, I believe, true. . . [unfinished]

A: NORMATIVITY, NATURALISM, AND NONCOGNITIVISM

By asking some questions, we can distinguish several views:

Are some normative claims intended or believed to state truths?

Yes — Cognitivism

No — Non-Cognitivism

Are there any normative truths?

Yes

Are these truths irreducibly normative?

Yes

Are these truths about what exists in some non-spatio-temporal part of reality?

Yes — Platonism

No — Non-Platonic Non-Naturalist Cognitivism

No — Nihilism

No

Are the concepts with which we state such truths irreducibly normative?

Yes — Non-Analytical Naturalism

No — Analytical Naturalism

These distinctions are rough, and do not cover all the possibilities. We ought, I believe, to accept some form of Non-Platonic Non-Naturalist Cognitivism. I argue here that we ought to reject both Naturalism and Non-Cognitivism. Platonism and Nihilism I hope to discuss elsewhere.

PART ONE

1 Normative and Natural Concepts, Claims, and Facts

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 we shall be considering, most refer to *properties*, such as the properties of being new, glittering, a poet, a convincing argument, the brightest star, or the first man to walk on the Moon. When we claim that some concept *refers* to some property, we are not claiming that anything *has* this property. No one is the first man to walk on the Sun; and, on some views, no acts are wrong.

The same word can have different senses, or meanings, thereby expressing different concepts. 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 'refute' to mean 'deny', and I deliberately use the word 'event' in a wide sense that covers acts, processes, 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, duty, good, bad, excellent, mediocre.
>
> B: crimson, square, electric, cause, city, marble, alive, sister, tall, unexpected.

Though I have not said what the words in each list have in common, most of us would guess correctly into which list most other words should go. 'Desirable', 'rational', 'blameworthy', and 'deserved', for example, should go in list A, and '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. 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 working in any of the natural or social sciences. I shall suggest later how we can make this vague definition more precise.

There are also 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', 'courageous', '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 depth to our normative thinking, the most important meta-normative questions are about the relations between the *thin,* purely normative concepts that are expressed by the words in list A, and the purely naturalistic concepts that are expressed by the words in list B.

According to Non-Cognitivists, normative claims should not be regarded as intended to be true, except in some minimal sense. Such views I shall discuss in Part Four. According to Cognitivists, most normative claims are intended to be true. Unless they are Nihilists, Cognitivists also believe that some of these claims *are* true, thereby stating normative facts.

There are three main kinds of Cognitivist theory. 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, there is no distinction between normative and naturalistic *claims*, but we can distinguish between normative and naturalistic ways of stating the same claim.

This view correctly describes some uses of normative words. For example, when I say

> My headache has got worse,

I might mean only

> My headache has become more painful.

These would then be different ways of stating the same claim, and the same natural fact. To give some other examples, when people compare the prices of goods that are for sale, buyers may use 'better' to mean 'lower' and sellers may use 'better' to mean 'higher'. And when people call some medicine 'the best', they might mean only that this medicine is the safest and the most effective. These uses of 'better' and 'best' state natural facts. Analytical Naturalism cannot be plausibly applied, however, to many other uses of 'worse', 'better', and 'best', or to some other normative words, such as 'unjust' or 'morally wrong'.

If some normative word, when used in some way, cannot be defined in non-normative terms, we can call this word, and the concept that this word expresses, *irreducibly normative*. Such words can be used to make irreducibly normative claims. We can similarly call some *fact* irreducibly normative if this fact cannot be restated in non-normative terms. According to *Non-Naturalist Cognitivists*, we make some irreducibly normative claims, and such claims, when they are true, state irreducibly normative facts.

According to *Non-Analytical Naturalists*, though we make some

irreducibly normative claims, such claims do not state irreducibly normative facts. When such claims are true, they state normative facts that could also be stated in non-normative, naturalistic terms. Such facts we can call 'natural' in the *reductive* sense.

These remark assume that, even when two sentences have different meanings, thereby stating different claims, these claims might state the same fact. That is true, in the sense that is relevant here, when these claims give us the same information.

To illustrate these kinds of theory, we can describe three versions of the view that

(1) some act is right

just when, and because,

(2) this act maximizes happiness. [614]

To save words, I shall here call this view *Utilitarianism*. 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 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 natural and normative.

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, that 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 natural and normative.

Similar claims apply to other Cognitivist moral theories, and to Cognitivist theories about non-moral normative concepts, claims, and facts. These theories can be either Analytically or Non-Analytically Naturalist, or be Non-Naturalist.

Before we try to decide between these three kinds of theory, we can introduce our main question, which is how we should understand normativity.

According to what I have called the *rule-involving* conception, normativity essentially involves rules, or requirements, which distinguish between what is *allowed* or *correct* and what is *disallowed* or *incorrect*. Some examples are laws, rules about spelling, grammar, and the meanings of words, the requirements of some code of honour, and the rules of etiquette. Such rules are often called *norms*.

According to the *reason-involving* conception, normativity essentially involves reasons or apparent reasons. To express this conception, we can call claims *normative* in the *reason-implying sense* when these claims assert or imply that someone did, or does, or might have some reason. This, I believe, is the best and most useful sense of 'normative'.

For some claim to be in this sense normative, it is not enough that this claim uses some normative word. If I say

> You shouldn't eat peas with a spoon,

or

> You shouldn't use 'refute' to mean 'deny',

these claims might be in this sense normative. But they would not be normative if I add that, since these rules are now so often broken, you have no reason not to act in these ways. I would then be using these claims to state what are merely natural facts. It is a sociological fact that, according to some rules of etiquette, we shouldn't eat peas with a spoon, and it is (or was) a linguistic fact that, according to the rules of English, we shouldn't use 'refute' to mean 'deny'. [615]

According to a third conception, normativity essentially involves *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? [616]

Elizabeth Anderson similarly writes

> any theory of the good must have normative force: we must be capable of being moved to action by the reasons it gives us. [617]

Many other writers make such claims. We ought, I shall argue, to reject this conception. Normative force is best conceived as wholly different from motivating force.

When Moore started the 20th Century debate about Naturalism, he was discussing the concept *good* and the property of being good. [618] Many other writers discuss Naturalist theories about morality. But I shall mainly discuss non-moral practical reasons and reason-

implying oughts. The questions raised by Naturalism here take simpler and clearer forms. And these are the most important questions if, as I believe, normativity is best understood as essentially involving reasons or apparent reasons. In the conflict between Naturalist and Non-Naturalist theories, reasons provide the decisive battlefield.

Most Naturalists claim or assume that our reasons for acting depend on certain facts about our desires, or about how we might be motivated to act. That is no coincidence. As Stephen Darwall writes:

> 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. . . [619]

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.

2 Analytical Subjectivism about Reasons

Of those who give Naturalist accounts of reasons, many are *Analytical Subjectivists*. According to Bernard Williams, for example, when we say that

> (A) someone has a reason to act in some way,

we often mean something like

> (B) this act would fulfil one of this person's present fully informed telic desires,

or

> (C) if this person knew the relevant facts, and deliberated rationally, this person would be motivated to act in this way. [620]

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 claims about reasons. David Falk, for example, defines a reason as a fact belief in which would motivate us. [621] Williams, Falk, and others give similar accounts of the reason-implying senses of 'should' and 'ought'. [622] 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 some way, or *should* or *ought* to act in some 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.[623]

This claim defines the *internal* senses of the words 'decisive reason', 'should', and 'ought'.

According to another view, which we can call

> *Analytical Externalism*: We often use words like 'reason', 'should', and 'ought' in irreducibly normative senses, which cannot be helpfully defined or explained in other terms.

These are the *external* senses of these words, which we can use to make claims about *external* reasons and oughts. [624]

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. [625] 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. 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 coincide, since facts about our reasons are facts about how we might be motivated to act. If we use these words in their *external* senses, these conceptions conflict, and we should accept the reason-involving conception. On this view, since facts about our reasons are not facts about our motivation, normative force is quite different from, and does not even include, motivating force. In *Early Death,* for example, your failure to be motivated 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' 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 disagree only in their beliefs about which facts give us reasons. But that, I believe, may not be true. I fully understand the concept of *an internal reason* as described by Williams 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. [626] I admit that, when I say that we have some reason, or that we should or ought to act in some way, what I mean cannot be helpfully explained in other terms. I could say that, when some fact gives us a reason, this fact *counts in favour* of our acting in some way. But this claim adds little, since 'counts in favour' means, roughly, 'gives us a reason'. Williams suggests that the phrase 'has a reason' does not have any such understandable, indefinable, and 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. [627] Several other writers make such claims.

I used to assume that, when Williams and I make claims about reasons, 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 support the view that I misunderstand my own thoughts, since I am wrong to believe that I use this phrase in such an indefinable external sense. I cannot believe that I really use this phrase in the internal sense. When I consider *Early Death*, I believe that you would have a decisive reason to take your medicine, though I know that you would have no internal reason to act in this way. I cannot see how I could be so deeply confused as to believe that it would be both true and false that you would have a decisive internal reason to act in this way. But I cannot exclude the possibility that, as Williams suggests, my use of the phrase 'has a reason' means nothing.

Suppose next that the phrase 'has a reason' has only one ordinary sense, which is the external sense. That would support the view that Williams misunderstands his own thoughts, since he does in fact use this phrase in this external sense. Williams might mistakenly deny that he uses this external sense because he has philosophical objections to the belief that we can understand and use any such irreducibly normative concept. Other philosophers have rejected

widely used concepts on similar grounds.

There are other possibilities. This phrase might have two ordinary senses, which are the internal and external senses, or each of these senses might be used by many people. We could then plausibly assume that Williams and I each mean what he thinks he means. We can now suppose that to be true, since it is worth asking whether we can have both kinds of reason.

Since it is clear that we can have 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 we have only internal reasons, we should revise our normative thinking, by coming to use the phrase 'has a reason' only 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 both what most of us mean, and the internal reasons that most of us truly believe that we have. 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 which are the questions that most people ask, but whether we have external reasons. If most people use the phrase 'has a reason' only in the internal sense, that might cast doubt on the view that we have external reasons, since it might seem unlikely that we have failed to recognize 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 people use the phrase 'has a reason' only in the external sense. Externalists could again claim that theirs is the more important view.

Internalists might reject this claim in a similar way. These people might say that, if they are not discussing the questions about reasons that most people 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 would be important.

3 The Unimportance of Internal Reasons

Some claim is

> *conceptual* or *linguistic* when this claim, if true, would state some fact about our concepts, or the meaning of our words or sentences,

and

> *substantive* when this claim, if true, would state some non-conceptual fact.

This distinction is rough, and needs to be qualified in various ways,[628] but we can ignore these qualifications here. 'Substantive' does not mean 'important', since there are many trivial substantive facts, such as the fact that I was born on a Friday. And though some conceptual or linguistic facts are trivial, others have great importance. Two examples would be facts about what is implied

by the concepts of *truth* and *knowledge*.

Let us next use 'desires' as short for 'fully informed telic desires' and 'ideal' as short for 'fully informed and procedurally rational'. 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 in the decisive-reason-implying senses,
>
> 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 the simplest form of

> *Analytical Internalism*: (A) and (B) mean the same.

If these claims meant the same, Subjectivism about Reasons would not be a substantive normative view. This view would be a concealed tautology, which might seem to give us non-linguistic information, but which really told us only that we could say the same thing in two different ways. To make this fact clearer, we could restate this view as

> *SR2*: Some possible act is
>
> > (A) what we have most reason to do, and what we should or 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.

This claim would add nothing to Analytical Internalism. So Analytical Internalists should claim only that, when we call some act

> what we have most reason to do, or what we should or ought to do,

we mean something like:

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

There is no point in 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 believe that they are describing and defending a substantive normative view. For example, Darwall claims to be describing 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. [629]

On Darwall's assumptions, however, (C) is not a substantive claim. Darwall assumes that, 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. [630]

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.

We could all accept this trivial claim.

I have argued that

> (E) If we used normative words in the senses that Analytical Internalists describe, Subjectivism about Reasons would not be a substantive normative view.

Someone might suggest:

> (F) Subjectivism about Reasons *is* a substantive normative view.

Therefore

> We do not use these words in these senses.

But this second argument would fail. For the reasons that I have just given, Analytical Internalists should reject premise (F). And some of these Internalists may have already seen that they should reject (F). Williams, for example, never claims that his Internalist version of Subjectivism is a substantive normative view.

There is, however, a better argument. We can claim:

> (1) If we used normative words in the senses that Analytical

> Internalists describe, though we might have some substantive normative beliefs, we could not have such beliefs about what we have reasons to do, or what we should or ought to do.
>
> (2) We do have such beliefs.

Therefore

> (3) We do not use these words in these senses.

Though I am inclined to believe that this argument is sound, I cannot show that (2) is true, since it is an empirical question what most people believe. But I shall now try to show that (1) is true. I also believe that

> (4) we can use these words in external senses, and we can thereby have substantive beliefs, some of them true, about what we have reasons to do, and what we should or ought to do.

(1) and (4), I believe, are more important than (2) and (3).

Williams seems to accept (2), since he writes:

> It is essential to any adequate account of 'A has reason to do X' that it should be normative. . . [631]

But Williams denies (1), since he argues that his Internalist account of claims about reasons explains the normativity of these claims. 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'. [632]

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 concept or notion 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 someone 'You have a reason to do X', we often intend to be giving this person advice. It would seldom be advice to say 'You want to do X', since few people need to be told what they want to do. But it would 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 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 normative claims need not use such a concept. Though the claims just mentioned 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 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. These facts may seem enough to show that Williams's Internalist account, according to which (G) means something like (H), 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.

This defence of Williams's account does not, I believe, succeed. To see why, we can first distinguish between facts that are *normative* and facts that have *normative force* or *importance,* in the sense that these facts give us reasons. [633] 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 all cases. Whenever some natural fact gives us a reason, there is also the normative fact *that* this natural fact gives us this reason. [634]

It is easy to overlook such normative facts. That mistake is

especially likely if, rather than saying that some natural fact *gives* us a reason, we say that this fact *is* a reason. If we claim that certain natural facts are reasons for us 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, I believe, whether we can understand and think about normative facts.

Return next to Williams's suggestion that, when we say 'You ought to do X', we 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.

As I have said, such claims are in one way like:

> You would enjoy this book,
>
> This is the sharpest knife,
>
> Your wine is poisoned,
>
> There's an angry bull in the next field.

We can add:

> This act would bring you great happiness,
>
> Since this hotel is on fire, jumping into the canal is your only way to save your life.

These claims might tell you facts that would give you reasons for acting, so such claims might be used to give advice. But, to intend to use such claims in this way, we must have the concept of *advice*.[635] We must be able to understand the thought that

> (L) such claims might state facts that would give someone reasons to act in some way, and might make this act something that this person ought to do.

We can now ask whether, if we used normative words in the senses that Williams and other Internalists 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 he discusses this example, Williams 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, since we cannot be advising people when we tell them to do what we believe that they have no reason to do.

Suppose next that, in

> *Revenge*, after considering the relevant facts in an ideal, procedurally rational way, you would choose to kill some enemy who has insulted you. This act would also best fulfil your present fully informed telic desires. You know, however, that you would be caught, and punished with hard labour for the rest of your life.

As I explain in Chapter 3, since Internalists and other Subjectivists appeal only to claims about procedural rationality, these people cannot claim that such a case is inconceivable. Nor can these people claim that, when we ask whether we ought to accept their theories, we can ignore such imaginary cases, or give them less weight.

According to Analytical Internalists, you have decisive internal reasons to kill your enemy, so this is what in the internal sense you ought to do. As before, these Internalists could not honestly advise you *not* to kill your enemy. It is even clearer that we cannot be *advising* people when 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 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 would choose 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*, as I have said, 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. Most of these people are Analytical Internalists, who assume that, when we claim that

> (A) we have decisive reasons to act in some way, or that we should or ought to act in some way,

we mean something like

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

Though (O) uses the normative words 'best' and 'ideal', these Naturalists believe that we can restate (O) in non-normative terms. On this view, we could give naturalistic accounts of what would best fulfil our desires, and of what would be involved in ideal, procedurally rational deliberation. We could then claim that (A) means something like

> (P) 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 should or ought to act in this way, these facts are natural facts of the kinds described by (P).

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. [636]

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. [637]

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. [638]

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. [639]

When Darwall discovers that he does not have this reason, what he discovers is the natural fact that he is not motivated to act in some way. [640]

If these Naturalist Internalists correctly described our normative concepts, we could not, I believe, have normative thoughts. To illustrate this point, suppose first that I am outside your *Burning Hotel*, and I can see you at some window. According to these Internalists, if I think

> You ought to jump into the canal,

I would be thinking something like

> (Q) Jumping would do most to fulfil your present fully informed telic desires,

or

> (R) After deliberating in certain naturalistically describable ways, you would choose to jump.

These are not normative thoughts. (Q) is merely a causal claim, and (R) 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, thereby 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

> (S) in the Naturalist internal sense, you ought to kill your enemy.

This fact, I have claimed, is not normative. If I said that, in this sense, you ought to kill your enemy, I would not be advising you to act in this way. I would mean only that this act would do most to fulfil your present fully informed desires, or is what, after deliberating in certain ways, you would choose to do.

These Internalists might next reply that, in making this causal claim, or this psychological prediction, I *would* be advising you to kill your enemy. I would be telling you what you ought to do. If this act would do most to fulfil these desires, and is what after such deliberation you would choose to do, that's *what it is* for killing your

enemy to be what you ought to do.

This reply assumes that

> (T) words like 'reason', 'should', and 'ought' have no external, irreducibly normative senses.

If (T) were true, normative claims might have to be such psychological predictions, or causal claims. There might be nothing else for normativity to be. We might have to admit, as the price of making such claims, that you have no reason to take your medicine, and that you ought to kill your enemy.

I believe that

> (U) these words can be intelligibly used in such external senses,

and that

> (V) since we can use these words in these senses, these are the only uses of these words that are truly normative.

Judged by this standard, psychological predictions 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 you would choose to do, I can still believe that you ought, and have decisive reasons, *not* to act in this way.

Some writers, such as Darwall, defend a different form of Internalism. These people also claim that, when we say that

> (A) we have decisive reasons to act in some way, or should or ought to act in this way,

we mean something like

> (W) 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 Internalists, however, (W) cannot be restated in non-normative and naturalistic terms. [641] (W) describes 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 senses, we could have irreducibly 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. These beliefs would not be *relevantly*

normative.

To illustrate this distinction, we can first imagine a new, stipulated sense of the word 'ought'. When we claim that some act is

> what someone 'ought to do' in this *unjust-world* sense, we 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 concept that we could express with this sense of 'ought'. But this concept is only partly normative, and this concept's normative part is not about what people ought to do. Suppose that, in *Revenge*, you have chosen to kill your enemy. Since I believe that the world is unjust, I would then agree that

> (X) in the unjust-world sense, you ought to kill your enemy.

Though this claim uses the word 'ought', (X) is merely another way of saying that you have made this choice in an unjust world. Such claims are irreducibly normative because they imply that the world is unjust. But these are not substantive 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 internal sense of 'ought' that Darwall and others use. If we know that

> (Y) you have chosen to kill your enemy after fully informed and procedurally rational deliberation,

we could truly claim that

> (Z) in this internal sense, you ought to kill your enemy.

But, though this claim is irreducibly normative, (Z) is not relevantly normative. (Z) merely tells us in a different way that you have made your choice after such ideal deliberation. Since (Y) is not a claim about what you ought to do, and (Z) adds nothing to (Y), (Z) cannot be a substantive normative claim about what you ought to do.

Here is another way to make this point. Though I would believe that (Z) is true, I would also believe that you have no reason to kill your enemy, and decisive reasons *not* to act in this way. Since these beliefs would not conflict, (Z) would not be a normative belief.

If we used the word 'ought' only in this internal sense, we could ask

> Q1: Which ways of deliberating are procedurally rational, and in other ways ideal?

In answering this question, we could have irreducibly normative beliefs, some of which might be true. And when considering some person's choice between different possible acts, we could ask

> Q2: After such a process of ideal deliberation, what would this person choose to do?

We could also ask

> Q3: Which act would best fulfil this person's present fully informed telic desires?

But we could not ask, as a further, independent question:

> Q4: What ought this person to do?

If we used 'ought' in this internal sense, this would be merely another way of asking Q2 or Q3. In other words:

> (5) If we used this sense of 'ought', we could have substantive normative beliefs about which ways of deliberating are ideal, and about what would best fulfil some set of desires. But we could not have any such distinct normative beliefs about what we or other people ought to do.

I shall now restate my main conclusions. 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 in the decisive-reason-implying senses,
>
> 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 one form of

> Analytical Internalism: (A) and (B) mean the same.

If these claims meant the same, Subjectivism about Reasons would not be a substantive normative view. This view would be a concealed tautology, one of whose open forms would be

> (C) What would best fulfil these present desires is what would best fulfil these desires, and what after such deliberation we would choose is what after such deliberation we would choose.

Analytical Internalists might say that, when they describe the internal senses of the words 'reason', 'should', and 'ought', they are not intending to state a substantive view. These Internalists may be intending only to give a true account of our normative concepts. 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 normative words and concepts to have and to state substantive normative beliefs.

This claim, I have argued, is not true. If we used these words in what I have called their *Naturalist* internal senses, the concepts that these words express would not even be normative. If our conceptual scheme took this impoverished and cruder 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 Internalists claim that, since the phrases 'best fulfil' and 'ideal deliberation' are irreducibly normative, so are the internal senses of 'reason', 'should', and 'ought'. These senses, I agree, express normative concepts. As I have argued, however, these concepts are not relevantly normative. If we used these concepts, we could have substantive normative beliefs about which acts would best fulfil our desires, and about which ways of deliberating are procedurally rational. But we could not have such beliefs about what we have reasons to do or what we ought to do.

Some Internalists claim that, since the phrases 'best fulfil' and 'ideal deliberation' are irreducibly normative, so are the internal senses of 'reason', 'should', and 'ought'. These senses, I agree, express normative concepts. As I have argued, however, these concepts are not relevantly normative. If we used these concepts, we could have some substantive normative beliefs, but we could not have such beliefs about what we have reasons to do, or what we should or ought to do.

4 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 ought in the external sense 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 Internalism, this theory is a substantive normative view. Though I have objections to such subjective theories, which I present in Chapter 3, these are not relevant here.

It may seem that, if we use 'reason' and 'ought' in their external senses, we cannot accept a desire-based or choice-based subjective theory. But that is not so. It is true that, if we use these words in their external senses, we could coherently deny that there any desire-based or choice-based reasons, thereby rejecting all subjective theories. But that is precisely why, if we instead defend such a 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 decisive external reasons to act in some way just when, and because,
>
> > (1) we have decisive *internal* reasons 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). Internalists might next suggest that there are two kinds of external reason, since subjective theories appeal to *internal* external reasons, and objective theories appeal to *external* external reasons. But that would not be worth saying. As I have argued, the concept of an *internal reason* does no useful work. We should use the phrase 'a reason' only in its external, irreducibly normative sense. Since *all* true reasons are external reasons, we need not call such reasons 'external'. And, when we are discussing substantive theories about reasons, we need not call these theories 'Externalist'.

5 Normative Beliefs

To be able to understand and accept substantive theories, we must use words like 'reason', 'should', and 'ought' in their indefinable, irreducibly normative senses. We can now briefly consider whether these words have such senses.

Falk makes several relevant remarks. When we believe that we ought to do something, Falk claims, we are often believing that, if we reflected on the relevant facts, we would want to do this thing.

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. [642]

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 be unalterable '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 suggests another objection to his view. On Falk's account, when I say that you ought to do something, I mean that there are facts belief in which would give you decisive reasons, by moving you to act in this way. This use of 'ought', Falk writes, is in one way puzzling. The claim 'You ought' does not itself give you a reason. Since that is so, this claim

> seems. . . logically redundant. . . One persuades by rational methods when one gives reasons. . . What else but another reason could add persuasive force to the reasons already given? But 'you ought to' is said after everything to count as a reason has been enumerated. It seems persuasive, and like adducing a reason, and yet is not. It seems both to belong to persuasion by rational methods, and not to be part of it.

Falk may here have seen that, if such uses of 'ought' were merely psychological predictions, as his account implies, these claims would not be normative. [643] He then asks whether we could expand this sense of 'ought' in some way that would make such claims 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, to make these 'ought'-statements normative, 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?' [644] 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. So this attempt to make 'ought'-statements normative fails.

Falk then suggests that, when I say 'You ought to do X', I am not merely claiming that you have reasons to do X, in the sense that there are facts belief in which would motivate you. I am also claiming that these facts are *good* or *valid* reasons for you to do X. [645] On Falk's account, however, this second claim would merely repeat my psychological prediction. I would mean only that, if you knew these facts, your belief in these facts would motivate you. In Falk's words, we want 'the hearer to have the benefit of *experiencing* what we claim'. [646] [647] 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 make 'ought'-statements normative also fails.

It may seem surprising that, when Falk worries that his internal sense of 'ought' is not normative, his first response is to treat the meaning of 'ought' as 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 other expressive utterances, such as 'Hurray!' or 'Boo!' Commands, I believe, cannot be normative, in part because they cannot be true. Falk briefly mentions the view that we might 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'. [648]

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* 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. Perhaps, when I use these words, I mean nothing, and am thinking nothing.

I am not aware of any arguments or facts that give us decisive reasons to believe that words like 'reason', 'should', and 'ought' cannot be intelligibly used in such irreducibly normative senses. It is, I admit, unclear how we come to understand such words, and the concepts they express. It is also unclear how we can recognize such irreducibly normative truths. But these unclarities do not, I believe, give us decisive reasons to conclude that we have no such concepts, or that there are no such truths. It is not surprising that, like some other fundamental concepts such as *time*, *space*, *possibility*, and *reality*, these fundamental normative concepts cannot be helpfully explained in other terms.

It is worth repeating here that, if we had decisive reasons to believe that there are no such irreducibly normative truths, the fact that we had these reasons would itself have to be one such truth. So we could not have decisive reasons to believe that we have no such reasons. That 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 would go, and how blank this sceptical state of mind would be.

I used to assume that most people have such irreducibly normative beliefs about reasons. But, given what some people say and write, that assumption may be false. Darwall, for example, 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. [649]

If Darwall had my concept of a reason, he would not make such claims. When I believe that I have decisive reasons 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 these reasons, by doing this thing. Just as giraffes were naturally selected for their long necks, and cheetahs were selected for their speed, we were selected for our ability to respond to reasons. Since we are partly rational beings, it is not mysterious how, when we are aware of facts that give us decisive reasons, we can respond to these reasons.

PART TWO

6 Non-Analytical 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. I have argued that, if we used these normative words in the senses that these Naturalists describe, we could not have 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, in the sense that we can state these facts by making non-normative and purely naturalistic claims. But some natural 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 impressive and plausible ways.

Most of these people make claims that are not about reasons, but about morality. Though I believe that reasons provide the decisive battlefield, we can start by considering what these people claim about morality.

These Naturalists mostly assume that, if there are 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.' [650] Richard Boyd even writes that 'goodness is probably a physical property'. [651] [652]

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 just when, and because, these acts have a certain natural property. [653]

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. [654]

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, and because, 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 more complicated set of natural properties.

This argument does not, I believe, succeed. When we consider pairs of 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, relevantly true. Consider first the arithmetical concepts expressed by these phrases:

> *the only even prime number*,
>
> *the positive square root of 4*.

These 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 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. [655]

(B) is also false, I believe, when applied to pairs of concepts of which one is naturalistic but the other is normative. If (B) is false, this argument does not show that some form of Moral Naturalism must be true. If acts were right just when and because they maximize happiness, that would not imply that being an act that maximizes happiness is 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, for example, that given certain moral assumptions, Utilitarianism might coherently and 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 that most of us find more plausible.

According to this form of Utilitarianism,

> (D) though the concept *right* is different from the concept of *maximizing happiness*, these concepts both refer to the same property. Being an act that maximizes happiness is the same as being right.

Such a claim, as I have said, 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,

(C) acts are right just when, and because, they maximize happiness.

It would then be true that

(F) maximizing happiness is the property that makes acts right.

This claim would use an irreducibly normative concept to refer to the natural property of maximizing happiness. [656]

(F)'s truth would not, however, support Moral Naturalism. Though (F) would use a normative concept that referred to a natural property, (F) would be merely another way of stating the normative claim that is stated by (C). 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. The claim that

(G) some natural property is the property that *makes* acts right,

differs from the claim that

(H) this natural property is the property of *being* right.

To explain this distinction, we can first note that, when some act has some property that makes it right, this act's having this property does not *cause* it to be right. Though there are many 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 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. As before, having such greater energy does not cause this thing to be hotter, but is *what it is* for this thing 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 difference between the claims that mental states *are* or *consist in* states of the brain.

Return now to the relation between *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 has one of several natural properties, such as being an act that either saves someone's life, or benefits someone and harms no one, 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 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 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 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

> (F) maximizing happiness is the natural property that makes acts right.

this truth would not support Moral Naturalism. (F) does not imply that

> (I) being an act that maximizes happiness is the same as being right.

(F) could be claimed to imply that

> (J) when acts maximize happiness, that makes these acts right by giving them the different, irreducibly normative property of being right. [657]

This, for example, is Sidgwick's view.

These remarks do not refute Moral Naturalism. Naturalists might still argue that moral rightness is, or consists in, one or more natural properties. But 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 has 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 has 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 has 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. We can see 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 6. 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 is harder to defend. Though social scientists can discover facts about the moral practice of judging people's acts to be blameworthy, these are not, I believe, 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. [658]

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 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 answer:

> 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 the *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 the 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 now say:

> You have not seen how deeply you and I disagree. Though 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 understand it. Ross and I both believe that, when acts maximize happiness, this fact might make these acts have the different property of being right. If there are not two different properties here, Ross and I would be seriously mistaken. And our mistake would not be about what it is for acts to maximize happiness. Our mistake would be about what it is for acts to be right. [659]

> As you say, Ross and I both asked what it is for acts to be right, and which acts have this property. Your view, you claim, answers these questions. That would be true only in a bleak and Nihilistic way. Ross and I both know that some acts maximize 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. There could not be any substantive moral facts about which acts are right or wrong. That would be how, in trying to decide which acts are right or wrong, 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 some 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 respond to this response, I shall use a crude and only partial analogy. [660] Suppose that I believe in God, and I have spent much of my life trying to decide which religious texts and theologians give the truest account 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.

7 Non-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 more usefully add another definition. When we call some normative fact

> 'natural' in the *reductive* sense, we mean that such facts could in principle be restated or redescribed by making non-normative and naturalistic claims.

Facts are *not* in this sense natural if they are irreducibly normative, in the sense that such facts could not possibly be restated or redescribed 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 in principle be restated or redescribed by making claims that are not normative. If the answer is No, this normative fact is not a natural fact. 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.

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.

With these definitions, we can use Scanlon's phrase 'irreducibly normative and, hence, non-natural'.

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. [661] On the contrary, he explicitly claims that normative facts might be able to restated in non-normative terms. Sturgeon illustrates this claim in a familiar way. Though he is not a Utilitarian, Sturgeon claims that if Hedonistic 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. [662]

Though Sturgeon does not reject my reductive definitions of 'natural' and 'Naturalism', we can imagine Non-Reductive Naturalists who did that. According to what we can call

> *Wide Naturalism*: Normative facts are natural facts even if such facts are irreducibly normative, because it would be in principle impossible to restate these facts in non-normative, naturalistic terms.

Normative *properties* are natural properties, these Naturalists would similarly claim, even if it would impossible to refer to or redescribe such properties by using non-normative, naturalistic concepts.

Since Wide Naturalists admit that certain properties and facts may 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 people who were working in one of the social sciences. But this definition would not be helpful here. We would have to ask whether, if some social scientist made claims about certain irreducibly normative facts, such as facts about which acts are wrong, this person would be making these claims *as a social scientist*. [663] It would be difficult for Wide Naturalists to defend the answer Yes except by arguing that these irreducibly normative facts would also be natural facts. So these remarks would not help to explain some sense in which these would be natural facts.

Sturgeon suggests another sense of 'natural' to which Wide Naturalists might appeal. According to

> *the Causal Criterion*: Some fact is natural if such facts can be appealed to in good causal explanations of what are clearly natural facts. [664]

Since there are many kinds of causal explanation, this criterion raises questions that I shall not try to discuss here. [665] It will be enough to give some reasons why, as I believe, we need not ask whether normative facts meet this criterion. First, this criterion is too strong, since there are many kinds of natural fact that could not be appealed to in any good causal explanations. Second, we cannot assume that, when normative facts provide such explanations, that would make them, in a 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 all or most 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 sentient or conscious animals, it is hard to believe that the actual Universe, or world, is the best possible 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, that 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. [666]

There is another, simpler reason why we need not discuss Wide Naturalism. We are asking whether we ought to accept some form of Non-Naturalist Cognitivism. When I and other people claim that some normative facts are not natural facts, we mean that these normative facts differ in several important ways from what are most clearly natural facts. The most fundamental normative facts are not, we believe, contingent, empirically discoverable facts about the actual Universe. These facts are necessary truths, which would be true in all possible words. It could not have been true, for example, that undeserved suffering was not bad. And when we claim that such facts are irreducibly normative, we mean that these facts are in a distinctive, autonomous category, which cannot be redescribed in other terms, or reduced to non-normative facts. Since Wide Naturalists accept these claims, they do not reject Non-Naturalist Cognitivism. [667] So, in considering 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 also be claimed to be natural facts.

It is worth mentioning one such wider sense. Wide Naturalists might say that, when they claim that irreducibly normative facts are also natural facts, they mean that such facts, or our belief in such facts, are 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 or epistemic reasons. Scientists make progress by responding to epistemic reasons and appealing to such facts about these reasons.

8

. . . . [A section to be added here about those natural facts that can be most plausibly claimed to be normative, thick normative concepts, and arguments from 'is' to 'ought'.] . . .

9 The Fact Stating Argument

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.

We can first look more closely at one of the main assumptions that leads people to accept such views. 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. [668]

Many other people make such claims. These people assume that

> (A) irreducibly normative words and concepts might refer to natural properties,

so that

> (B) we might use these words and concepts to make irreducibly normative claims which are about natural properties, and state natural facts.

(A), we have seen, is true, and the inference to (B) 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) does not support (B).

Consider first these phrases:

> (C) 'the largest planet',
>
> (D) 'being the largest planet'.

Despite their similarity, (C) refers to Jupiter, and (D) 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 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 some other property.

When stated so abstractly, this second distinction is confusing. But examples may make it clear. Return to the use of 'heat' which means

> the property, whichever it is, that has 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 has 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

> (E) maximizing happiness is the property that makes acts right.

If (E) were true, this claim would use an irreducibly normative concept to refer to the natural property of maximizing happiness. So (E) might seem to be the kind of claim for which Naturalists are looking: an irreducibly normative claim which might state a natural fact. But, as I have said, (E) is not such a claim. (E) could be more fully stated as

> (F) 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. [669] Since (E) and (F) both claim that a certain natural property has a certain normative

property, these are claims which, if they were true, would state a normative fact. So this example does not support Moral 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 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 'this apple is red', refers to redness *implicitly*, by describing this apple as having this property.

Return now to the phrase

> (G) 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 property. As we have seen, however, (G) 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 (G) would *also refer implicitly* to this other, normative property. And (G) 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 some fact that was natural in the reductive sense. [670]

Similar remarks apply, I believe, to all irreducibly normative 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

> (H) irreducibly normative concepts all refer, either explicitly or implicitly, to some normative property.

This is why, though it is true that

> (A) irreducibly normative words and concepts might refer to natural properties,

this truth does not support Naturalism. As we have seen, Gibbard takes (A) to imply that it is only concepts, not properties. that are distinctly or irreducibly normative. That, I have argued, is not so. Since these normative words and concepts would refer to natural

properties only by *also* referring to such normative properties, (A) does not help to show that there are no irreducibly normative properties. And we have no reason to expect that, as many Naturalists assume,

> (B) we could use these words and concepts to make irreducibly normative claims which might state natural facts.

Since such claims would refer to normative properties, they would, when they are true, state normative facts. So this common argument for Naturalism fails.

Naturalists might give other arguments. In considering other possibilities, we can distinguish three kinds of irreducibly normative concept. Such a concept might be

> (J) definable in some way that shows how this concept might refer to some natural property,
>
> (K) definable in some way that shows, or gives us some reason to believe, that this concept could *not* refer to some natural property,

or

> (L) indefinable.

We have just been discussing one concept of type (J): 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 (K), I gave the concept *blameworthy*. Other examples are the concepts expressed by these phrases:

> *unjustifiable to others,*
>
> *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 and resentment.*

It would be harder for Naturalists to argue that these concepts refer to one or more natural properties. These people would have to claim, for example, either that

> (M) the concept *unjustifiable to others* does not refer to the property of being unjustifiable to others,

or that

> (N) 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 (K) did refer to some natural property, Naturalists would have to argue that these concepts did not *also* refer to some normative property. That would be even harder to show.

The most important normative concepts, however, are of type (L). 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', or by the indefinable sense of 'wrong' which I express with the phrase 'mustn't-be-done'.

According to Non-Analytical Naturalists, though these concepts are irreducibly normative, they refer only to natural properties. Since these concepts are indefinable, it is in one way easier for Naturalists to argue that these concepts do not also refer to irreducibly normative properties. When concepts are indefinable, that leaves it in one way open to which properties these concepts refer. And some Naturalists claim that, though these normative concepts are not explicitly definable in non-normative terms, we can describe the role or function that these concepts have in our thinking, and we might then be able to argue that, for these concepts to play this role, they must refer 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.

No such argument, I believe, could succeed. We can give this counter-argument:

(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 these facts were natural, they could also be stated by some other non-normative, naturalistic claims.

(4) Any such true normative claim would then 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) The non-normative claim could not state a normative fact.

Therefore

If these claims stated the same fact, by giving us the same information, the 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. Facts are natural in this sense just when they could be stated by some non-normative, naturalistic claim.

These three premises imply (4). If such true normative claims stated natural facts, any such normative claim would state the same fact as some other, non-normative claim.

Premise (5) needs some explanation and defence. There are different senses in which different claims can state the same fact. It will be enough here to consider two such senses. Different claims state the same fact in what we can call

> 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:

> (O) Shakespeare is Shakespeare.
>
> (P) Shakespeare and the writer of *Hamlet* are one and the same person.
>
> (Q) Shakespeare wrote *Hamlet*.

In the referential sense, (O) and (P) state the same fact, since both claims refer to Shakespeare and tell us that Shakespeare has the property of being one and the same person as---or numerically identical to---himself. [671] In the informational sense, however, (O) and (P) state different facts, since only (P) tells us that Shakespeare wrote Hamlet. In this informational sense, it is (P) and (Q) which state the same fact.

Consider next

> (R) water is H2O,
>
> (S) water is water.

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 (R) states an important scientific discovery, since this fact would be the same as the trivial fact stated by (S). To explain how (R) was an important discovery, we must claim that (R) and (S) give us different information, thereby stating different facts.

Similar remarks apply to

(T) heat is molecular kinetic energy

and

(U) heat is heat.

When scientists discovered that (T) is true, they were not merely rediscovering the trivial fact stated by (U). (T) was an important discovery because these claims give us different information, thereby stating different facts.

Similar remarks apply to normative facts. Many Naturalists claim that, just as we have discovered that water is H2O and that heat is molecular kinetic energy, we might discover, or be able to show, that

(V) moral rightness is the same as some natural property.

In the referential sense, however, (V) would state the same fact as

(W) this natural property is the same as this natural property,

and this fact would be trivial. To be able to claim that (V) would state an important truth, these Naturalists must say that, since (V) and (W) would give us different information, these claims would state different facts.

As these remarks show, two claims state the same fact, in the sense that is relevant here, just when these claims give us the same information. So, as the Fact Stating Argument assumes,

(5) If some normative claim stated the same fact as some other, non-normative claim, these two claims would give us the same information.

I also believe that

(6) The non-normative claim could not state a normative fact.

If (6) is true, we ought to accept this argument's conclusion. If the non-normative claim could not state a normative fact, and these two claims stated the same fact, the normative claim could not state a normative fact. So such claims could not, as these Naturalists believe, state facts that are both normative and natural.

To illustrate this argument, and help us to ask whether (6) is true, we can return to claims about practical reasons and reason-implying oughts. As I have said, the relevant questions here take simpler and clearer forms.

10 The Normativity and Triviality Objections

Most Non-Analytical Naturalists accept some form of Subjectivism about Reasons. It will be enough here to consider the view that

> (A) we have decisive reasons to act in some way, and should or 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 those who accept this view, some believe that, when we make claims like (A), we often mean something like (B). Other people defend this view in other ways.

According to some Naturalists, though (A) and (B) are irreducibly normative claims, such claims, when they are true, state facts that are both normative and natural.

For these facts to be natural, in the relevant reductive sense, they must be able to restated by some other, non-normative, naturalistic claim. We can take this claim to be

> (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. (C) is not a normative claim because (C) does not use any normative word. According to these Naturalist Subjectivists, true claims like (A) and (B), though irreducibly normative, state facts that could also be stated by claims like (C).

This form of Naturalism could not, I believe, be true. We can argue:

> (7) Since (C) is a non-normative claim, this claim could not state a normative fact.
>
> (5) If different claims state the same fact, these claims give us the same information.

Therefore

> If (A) and (B) stated the same fact as (C), (A) and (B) could not state a normative fact.

Therefore

> Normative claims like (A) and (B) could not, as these Naturalists believe, state facts that are both normative and natural.

Naturalists could not, I have argued, reject premise (5). But they might reject (7). These Naturalists might claim

> (8) The fact stated by (C) *is* normative, because this fact could also be stated by the normative claims (A) and (B).

To illustrate this disagreement, remember that, in *Burning Hotel*, you will die unless you jump into the canal. According to these Naturalists, the fact stated by

> (D) you ought to jump,

is the same as the fact stated by

> (E) jumping would best fulfil your present fully informed telic desires, or is what, after fully informed and procedurally rational deliberation, you would choose to do,

which is the same as the fact stated by

> (F) 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.

On this view, this fact is normative because it can be stated by the normative claims (D) and (E), but this fact is also natural because it can also be stated by the non-normative, naturalistic claim (F).

Given the difference in their meaning, these 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 be soon be overcome by smoke. You might then believe, and tell yourself, that you have *decisive reasons* to jump, that you *should, ought,* 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, such as the causal and psychological facts stated by (F).

These Naturalists might reply that claims like (F) do not state *merely* natural facts, since these facts are also normative. My argument, Naturalists might say, shows nothing, since premise (7) merely assumes that non-normative claims cannot state normative facts, thereby assuming that Naturalism is false.

This reply does not, I believe, succeed, since we can defend premise (7). We can first remember the distinction between facts that are normative and facts that have normative importance. If it were true that

(G) eating walnuts would kill me,

this fact might have normative importance, since this fact might give me a decisive reason not to eat walnuts. But (G) would not state a normative fact.

There is another relevant distinction. Some normative claim is

> *conceptual* or *linguistic* when this claim, if true, would state some fact about our normative concepts, or the meaning of normative words or claims.

One example is the conceptual normative fact that, when we say that we ought morally to act in some way, we mean that all of the other possible acts would be wrong. Some normative claim is

> *substantive* when this claim, if true, states the fact that something has some normative property.

One example would be the substantive normative fact that we ought to act in a certain way, since all of the other possible acts would be wrong. To illustrate this distinction, we can note that even Moral Nihilists believe that there are conceptual moral facts, such as the fact just mentioned about how the concept *ought* is related to the concept *wrong*. These people are Nihilists because they believe that there are no substantive moral facts, since there is nothing that we ought to do, and no acts are wrong.

We can next compare these claims:

(H) You drove at 100 miles an hour.

(I) You drove at 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 there was no distinct property of being illegal. Only that would make it true that (I) would not give us any further information. If you *were* acting illegally, so that (I) *would* give us further information, (H) and (I) would not state the same fact, in the relevant, informational sense. In other words, if (H) and (I) stated the same fact, (I) could not be a substantive legal claim, since it is not a substantive legal fact that you drove at 100 miles an hour.

Similar remarks apply to

> (J) This act is what, after some process of deliberation that had certain natural properties, we would choose to do,

and

> (K) This act is what, after fully informed and procedurally rational deliberation, we would choose to do.

If these claims stated the same fact, that would have to be because there was no distinct property of being procedurally rational. Only that would make it true that claims like (J) and (K) gave us the same information, thereby stating the same fact. These Naturalists must therefore admit that, on their view,

> (L) if some process of deliberation would have these natural properties, that's *what it would be* for this deliberation to be procedurally rational.

If (L) were true, because there was no distinct normative property of being procedurally rational, (K) could not be a substantive normative claim. If some process of deliberation has certain natural properties, that is not a substantive normative fact.

Similar remarks apply to

> (A) We ought to act in some way,

and

> (C) 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.

According to these Naturalists, the facts that are stated by claims like (A) are the same as, or consist in, the facts stated by claims like (C). If this were true, that would have to be because there was no distinct normative property of being something that we ought to do. Only that would make it true that claims like (A) and (C) gave us the same information, thereby stating the same facts. So these Naturalists must admit that, on their view,

> (M) if some act would do most to fulfil our present fully informed telic desires, or is what after deliberating in such ways we would choose to do, that is *what it would be* for this act to be what we ought to do.

We can draw a similar conclusion. If M) were true, because there was no distinct normative property of being something that we ought to do, claims like (A) could not state substantive normative facts.

In applying the Fact Stating Argument to this form of Non-Analytical Naturalism, I have now made two overlapping claims. I have claimed that

> (N) if we ought to act in some way, this normative fact could not be the same as some merely natural fact, such as the facts stated by (C), since such facts are not normative.

We can call this *the Normativity Objection*. Though this objection seems to me decisive, it has the disadvantage that the word 'normative' is used in different senses, or to express different concepts. My sense of 'normative' cannot be explained except by appealing to some other indefinable normative words, such as 'ought' or 'reason'. Many people use the word 'normative' in other senses. Some of these people explain their sense of 'normative' by appealing to rules, or norms, which distinguish between what is correct or incorrect. Others appeal to certain claims about motivation. When people use 'normative' in such other senses, they may believe that certain natural facts *are* normative.

I have also claimed that

> (O) if we ought to act in some way, this would be a *substantive* normative fact. This fact could not be the same as some merely natural fact, such as the facts stated by (C), since these are not substantive normative facts.

This objection is in one way stronger than (N). Compared with the question whether certain natural facts could be in *some* sense normative, it is clearer to ask whether these natural facts could be substantive normative facts. If, as I shall argue, the answer is No, we could then claim:

> If Naturalism were true, there could not be any substantive normative facts.
>
> There might be such facts.

Therefore

> Naturalism cannot be true.

If there were no substantive normative facts, normative claims would be trivial. So we can call this *the Triviality Objection*.

To illustrate these objections, remember that, in *Revenge*, if you killed the enemy who has insulted you, you would be caught and punished with a life of hard labour. We have supposed that

> (P) this act would do most to fulfil your present fully informed telic desires, and is what, after deliberating in certain naturalistically describable ways, you would choose to do.

These Naturalist Subjectivists must agree that, on their view,

> (Q) you ought to kill your enemy.

This implication may seem implausible. But, as I have said, these Naturalists could reply that

> (R) since killing your enemy would do most to fulfil these desires, and is what after such deliberation you would choose to do, that's *what it is* for killing your enemy to be what you ought to do.

If (R) were true, (Q) would *not* be implausible, since (Q) would merely restate the natural facts that are stated by (P). (Q) would not state a substantive normative fact. If killing your enemy would have the natural properties described by (P), and that were also *what it would be* for this act to be what you ought to do, we would have no reason to deny that killing your enemy is what you ought to do. This claim would add nothing, and would be in this way trivial.

Consider next Mark Schroeder's version of Subjectivism. Schroeder claims that

> (S) when some fact helps to explain why some act would fulfil one of our present desires, that's what it is for this fact to be a reason for us to act in this way. [672]

Schroeder's view takes this form because he distinguishes between the facts which are reasons for acting and the facts about desire-fulfilment which make these facts reasons. On Schroeder's view, for example, if

> (1) you want to stay alive,

and

> (2) jumping into the canal would save your life,

the fact stated by (2) is a reason for you to jump because this fact explains why this act would fulfil your desire.

Schroeder points out that, on his view, we might have a reason to act in some crazy way, such as trying to eat our car, since we might have some 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. [673]

If we accepted Schroeder's view, however, we should not think it *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. On Schroeder's view,

> (3) if certain facts explain how acting in these crazy ways would fulfil one of our present desires, that's *what it is* for these facts to be reasons for us to act in these ways.

If that's *what it would be* for these facts to to be reasons for us to act in these ways, it would be easy to defend the claim that these facts would be such reasons. These facts *would* explain how these acts would fulfil one of our desires. In saying that these facts would provide this explanation, we would not be making a substantive normative claim, so this claim could not conflict with anyone's normative intuitions. So Schroeder need not argue that such 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* substantive normative implications.

Why does Schroeder believe that his view *does* have such implications? The answer may in part be that, 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'. When we consider Schroeder's view, we can ask

> Q1: If some fact explains how some act would fulfil one of our present desires, would this fact count in favour of this act?

We may here be asking

> Q2: If some fact has the property of explaining how some act would fulfil one of our present desires, would this fact thereby have the different property of counting in favour of this act?

We may assume that, on Schroeder's view, the answer to Q2 is Yes. And when we consider Schroeder's imagined cases, this answer may seem intuitively 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. And when Schroeder discusses such cases, he seems to assume that, in asking Q1, we would be asking Q2. When he supposes that some fact would explain how some crazy act would fulfil one of our desires, it seems to him implausible to claim that this fact would also have the property of counting in favour of this act.

On Schroeder's view, however, when some fact explains how some act would fulfil such a desire, that's what it is for this fact to count in favour of this act. There is only one property here. Being a fact that provides this explanation is the same as counting in favour of this act. On this view, when we ask Q1, we should take our question to be

> Q3: If some fact has the property of explaining how some act would fulfil one of our present desires, would this fact thereby have this property?

And to this question, the answer is clearly Yes. Acts that have this property have this property.

Schroeder's view *is* intuitively implausible. But what is implausible is not, as Schroeder assumes, some of this view's substantive implications. What is implausible is the way in which Schroeder's view does not have any substantive implications. If this view were true, claims about reasons would be normatively trivial, since such claims could not state substantive normative facts. Schroeder's view here conflicts with our *meta*-normative or meta-ethical intuitions. Many of us believe that, when we make true claims about reasons, and about what we should or ought to do, these claims *do* state----or, at least, *might* state---substantive normative facts.

Schroeder's view *is* intuitively implausible. But what is implausible is not, as Schroeder assumes, some of this view's substantive implications. It is the fact that, if this view were true, claims about reasons would be normatively trivial, since such claims could not state substantive normative facts. Schroeder's view here conflicts with our *meta*-normative or meta-ethical intuitions. Many of us believe that, when we make true claims about reasons, and about what we should or ought to do, these claims *do* state----or, at least, *might* state---substantive normative facts.

Schroeder himself seems to have this belief, since 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 also notes, however, his view 'analyzes reasons. . . in wholly non-normative terms'. [674] Since non-normative claims could not state substantive normative facts, Schroeder should admit that, on his view, there are no such facts.

These remarks do not refute Schroeder's view. Schroeder might be right to claim that there are no substantive normative facts. As we shall see, some Naturalists would accept this conclusion.

If Subjectivist Naturalists cannot accept this conclusion, they might give up their Naturalism. Darwall, for example, claims that, when we say that

> (A) we ought to act in some way,

we often mean that

> (K) this act is what, after fully informed and procedurally

rational deliberation, we would choose to do.

Darwall might now add that

> (T) since this claim is irreducibly normative, such claims, when they are true, state irreducibly normative and hence non-natural facts.

Since these Subjectivists would cease to be Naturalists, they would avoid the main objections that we have been discussing. But, as I argued earlier, if we used (A) to mean (K), we could have substantive normative beliefs about which ways of deliberating are procedurally rational, but we could not have such beliefs about what we ought to do. These Subjectivists would do better to give up their Analytical Subjectivism, by starting to use 'ought' in its indefinable decisive-reason-implying sense. These people could then have substantive Subjectivist beliefs about what we ought to do, such as the belief that we ought to do what, after such procedurally rational deliberation, we would choose to do.

Other Naturalists might respond to these objections in a different way. These people might reject Subjectivism about Reasons, and defend some Objectivist view. For Naturalists, however, that is hard to do. It is not surprising that most Naturalists are Subjectivists, since these views can be plausibly combined. It is fairly easy to believe that, when we have decisive reasons to act in some way, and we should and ought to act in this way, this fact is the same as, or consists in, some fact about what would fulfil our present informed desires, or about what, after some kind of ideal deliberation, we would choose. 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 be least implausible if we assume some form of Rational Egoism. Naturalists might then claim that, whenever we have a reason to act in some way, this fact would be the same as, or would consist in, the fact that this act would promote our own well-being, on some Naturalist account of well-being. But most of us believe that various other kinds of fact can give us reasons for acting. 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 fact that gives us this reason, or some other natural fact. Suppose that

> (U) if I acted in certain ways, I would relieve some stranger's pain, discover the cure for some disease, and help to save Venice from being destroyed.

We may believe that

> (V) these facts would give me reasons to act in these ways.

The normative facts that are stated by (V) cannot be plausibly

claimed to be the same as, or to consist in, the natural facts that are stated by (U). Of the features of Subjectivism that make this view appealing, one is the way in which subjective theories provide unified accounts of how a great variety of facts can give us reasons. On these theories, the facts stated by (U) 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 are acts that, after some process of deliberation, I would be motivated to do, or would choose to do. These claims give a plausible Subjectivist account of how a great variety of facts can give us reasons. But if Naturalists are not Subjectivists, there is no such set of natural facts with which they could plausibly identify our having reasons of these and various other kinds.

Even if Naturalists could defend some form of Objectivism about Reasons, my main objection would still apply. On such views, when some act would have certain natural properties, that's *what it would be* for us to have reasons to act in this way, or for this act to be what we ought to do. As I have argued, such views would imply that there are no substantive normative facts about what we have reasons to do, or what we ought to do.

When we compare conflicting meta-normative theories, I have claimed, it is reasons that provide the decisive battlefield. If Naturalism fails here, as I have argued that it does, it will also fail elsewhere. That would be clearly true if Naturalists have to admit that

> (W) when we have decisive reasons to act in certain ways, and we should and ought to act in these ways, these facts would be irreducibly normative, and would not be natural facts.

This claim would contradict the main thesis of Normative Naturalism.

There is another possibility. Rather than merely retreating from the battlefield of reasons, Naturalists might adopt a scorched earth strategy. Such facts about our reasons, Naturalists might claim, are *not* normative. If the fact that you ought to kill your enemy is not normative, we could accept that you ought to act in this way.

I believe that we do have decisive reasons, that you ought not to kill your enemy, and that such facts *are* normative. If Naturalists reject these beliefs, their view would be close to Nihilism.

PART THREE

11 Moral Naturalism

We can now turn from reasons to morality. Suppose that, in *Revenge*, you kill your enemy, and we believe that

(A) this act was wrong.

According to Non-Analytical Naturalists, though the concept *wrong* is irreducibly normative, this concept refers to a natural property, and (A), if true, states a natural fact.

Which fact does (A) state? Some Naturalists might say: 'There is no significantly different way to state this fact. We can say only that (A) states the natural fact that this act was wrong.'

This wide form of Naturalism, as I have argued, is not worth discussing. If these Naturalists believe that moral facts are irreducibly normative, they are accepting the main belief of Non-Naturalist Cognitivists. For Naturalism to be worth discussing, Naturalists must claim that moral facts are, in the reductive sense, natural facts. On this view, true claims like (A) state facts that could also be stated by some non-normative and naturalistic claim.

According to Utilitarian Naturalists, (A) would state the fact that

(B) this act failed to maximize happiness.

Most Naturalists would reject this claim, since they are not Utilitarians. But few Naturalists propose and defend some other, particular moral view. Most of these writers aim only to show that the best moral theory might take some Naturalist form.

Since these writers do not defend any particular moral view, we have to discuss this form of Naturalism in general terms. These people agree that the concepts *right* and *wrong* are irreducibly normative. We must ask whether these concepts might refer, not even in part to some irreducibly normative property, but only to some natural property, or set of properties. Only then could irreducibly normative claims state natural facts. As before, I believe, no such view could be true. If some act is wrong, this fact could not be the same as, or consist in, some natural fact, such as some causal or psychological fact.

These people might reply that, in rejecting Naturalism in this sweeping way, I am appealing to an overly simple view about how concepts refer to properties, and about how the meanings of our claims are related to the facts that our claims may state. These

people agree that, when we claim that some act is wrong, what we *mean* is quite different from some naturalistic claim. But claims with different meanings might state the same fact. For example, when people in earlier centuries said

> (C) The water in this bath is hotter,

they did not mean

> (D) The H2O in this bath has greater molecular kinetic energy,

but these claims can be held to state the same fact. These Naturalists might therefore say that, even though normative and naturalistic claims have different meanings, we cannot exclude the possibility that

> (E) some normative and naturalistic claims state the same facts.

If we use 'normative' in its narrow, reason-implying sense, we can, I believe, exclude this possibility. We can first remember that, in the relevant informational sense, (C) and (D) do *not* state the same fact. That is how, in discovering facts like the one stated by (D), scientists were not merely discovering facts like the one stated by (C).

Second, though normative and naturalistic *concepts* might refer to the same property, they do that in different ways, so that claims that use such concepts thereby give us different information. As I argued in Section 9, normative concepts refer to natural properties only by *also* referring to normative properties. So we have no reason to believe that, when we use these concepts to make normative claims, such claims might state facts that could also be stated by non-normative, naturalistic claims.

Given the differences between these kinds of claim, they could not, I believe, state the same facts. If we use 'normative' in the reason-implying sense and 'natural' in the reductive sense, we can claim that

> (F) normative and natural facts are in two quite different, non-overlapping categories.

Some reductive views can, defensibly, cross categories. Persisting things, for example, are in a different category from processes, or events. But the continued existence of at least some persisting things can be plausibly claimed to consist in the occurrence of certain processes or events. The continued existence of a river, for example, consists in a continuous flowing of water in a certain pattern. Other cross-categorial reductions are more controversial. Thus some people claim, while others deny, that experiences are the same as, or consist in, physical events in some brain. This disagreement is about whether conscious 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. [675] In the same way, I believe, when there is something that we have decisive reasons to do, and should and ought to do, this fact could not be the same as, or consist in, some psychological or causal fact.

In making these claims, I am appealing to what I mean by the words 'reason', 'should', and 'ought'. Some Naturalists would repeat that, when they 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 writers might again cite the discoveries that water is H2O and that heat is molecular kinetic energy. When scientists made these discoveries, many Naturalists say, they were not appealing to the pre-scientific meanings of the words 'water' and 'heat'.

This defence of Naturalism does not, I believe, succeed. Though these discoveries were not implied by the pre-scientific meanings of these words, these scientists *did* appeal to these meanings. That is why their 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. When people claim that their pain has got worse, for example, it may be unclear whether these people are using 'worse' merely to mean 'more painful'. And some words have complex senses that are partly normative and partly naturalistic. Some examples are 'dishonest', 'cruel', 'cowardly', and 'mean'.

For Naturalism to succeed, however, even the claims that are most

purely normative must, if they are true, state natural facts. And no such claims, I believe, could state such facts.

If, as I believe, normative facts are in a unique, distinctive category, there is no close analogy for their irreducibility to natural facts. The best comparison may be with some 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.

There are other distinctive categories of facts, such as physical or legal facts. And it is often clear that such categories could not overlap. For example, it could not be a physical or legal fact that 5 + 3 = 8, 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. These claims are not controversial, because we nearly all believe that these facts are in three quite different, non-overlapping categories, and we understand the differences between them.

It is harder to defend the claim that normative and natural facts are in quite different categories. All Naturalists believe that there could not be any irreducibly normative facts. And many Naturalists find the idea of such facts mysterious, or unintelligible.

There is another problem. Some normative words and concepts are, I have said, indefinable, in the sense that they cannot be helpfully explained in other terms. That makes it in one way hard to argue that claims that use such words could not state natural facts. When Non-Analytical Naturalists use words like 'reason', 'ought', and 'wrong', they may not mean what I mean. It seems to me likely that they don't. If these people *do* mean what I mean, I do not see how they could believe that irreducibly normative claims---such as the claims that some act is wrong, or that we have reasons to want to avoid agony---might state natural facts, such as some causal or psychological fact.

If these Naturalists and I use these words in different senses, we might be able to resolve our disagreements. If I understood what these people use these words to mean, I might agree that these people's normative claims might state natural facts, so that Naturalism would be true when applied to these people's claims. I give examples of such facts in Section 8. And if these people understood what I use these words to mean, they might agree that *my* normative claims could *not* state natural facts. These people would then agree that, when applied to my claims, Naturalism would be false.

To reach agreement in this second way, I would have to get these

people to understand my concepts and my claims. Since these normative concepts cannot be helpfully defined in other terms, I would have to use some other method. I might try to explain these concepts by getting these Naturalists to imagine cases in which they would have normative thoughts like mine. This method may fail. If Falk, for example, imagined being in your *Burning Hotel*, and thinking that he ought to jump, he might imagine feeling irresistibly impelled to jump. That thought is not in my sense normative. But it might help if Falk imagined being outside your hotel, and thinking that *you* ought to jump. He would be less likely to imagine that *you* were irresistibly impelled to jump. And he might find himself thinking that you have decisive reasons to jump, and that if you don't jump you would be making a terrible mistake. Falk would then be having thoughts that are not about your motivation, but are of a quite different kind.

This argument against Naturalism, we can now add, need not assume that there are some irreducibly normative facts. Nor need the argument even assume that irreducibly normative claims would, if they were true, state such facts. This argument's claim could instead be that

> (G) normative claims could not state natural facts, because such claims are in a unique, distinctive category.

Of those who are *metaphysical* naturalists, in the sense that they believe that all facts are natural facts, many would accept (G). 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. But 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. On such views, though there are no normative facts, we can justifiably make normative claims. These claims do not state beliefs, but express certain kinds of attitude. Though most Non-Cognitivists believe that all facts are natural facts, they share my belief that, since normative claims are in a unique, distinctive category, Normative Naturalism could not be true.

12 Substantive Normative Facts

There are other arguments for this conclusion. We can first return to the way in which some Moral Naturalists defend their view. These people claim that, though the central moral concepts are both irreducibly normative and not helpfully definable in other terms, we can describe the roles that these concepts play in our moral thinking. We might then be able to argue that, if these concepts refer to natural properties, that would best explain how these concepts play their roles. Similar arguments might appeal, not to the roles that these concepts play, but to various other features of our best moral thinking.

As before, we can ask whether some such argument might support some Naturalist version of Utilitarianism. Our conclusions could then be transferred to the more complex moral views that most of us find more plausible. So we can first suppose that our best moral thinking supports the claim that

(A) acts are right just when they maximize happiness.

This claim's truth would be best explained, Naturalists might say, if it were true that

(B) being an act that maximizes happiness is the same as being right.

As I have said, however, there is another possibility. (A)'s truth would also be explained if it were true that

(C) being an act that maximizes happiness is the property that *makes* acts right.

To defend their view, these Naturalists must claim that, to explain (A), we should appeal to (B) *rather than (C).*

This claim, I believe, would be hard to defend. Some Naturalists would claim that, if we appealed to (B), our theory would be simpler, and in that way better. While (C) makes a claim about the relation between two different properties, (B) refers to only one property. Though this difference might give us some reason to appeal to (B) rather than (C), this reason would not, I believe, be strong. And there are some features of our moral thinking that would be harder to explain if rightness were the same as maximizing happiness, or were the same as, or consisted in, any other natural property, or set of properties.

One such feature provides what I have called the *Triviality Objection.* As I claimed in Section 10, we can argue:

(D) If Naturalism were true, there could not be any substantive normative facts.

(E) There might be such facts.

Therefore

Naturalism cannot be true.

Naturalists would accept (E). I shall here continue my defence of (D).

According to all Naturalists, normative and naturalistic concepts might refer to the same properties, and be used to state the same facts. As I have said, Naturalists can claim that

(F) when we learn that different concepts refer to the same

> property, we may thereby learn some important substantive fact.

One example is the claim that

> (G) heat is molecular kinetic energy.

Though this claim uses two concepts that refer to the same property, (G) does not tell us only that this property is identical to itself. (G) also tells us how this property is related to various other properties. (G) can give us this information, as I have said, because the concept of *heat* has a gap that is waiting to be filled. 'Heat' means 'the property, whichever it is, that has certain effects, such as those of melting solids, turning gases into liquids, etc.' (G) could be restated as

> (H) molecular kinetic energy is the property that has these effects.

This claim gives us substantive empirical information.

Return next to the similar claim that

> (C) being an act that maximizes happiness is the same as having the natural property that makes acts right.

If (C) were true, this claim would use two concepts that refer to the same property, but it would also tell us how this property is related to another property. (C) could be restated as

> (J) when acts have the property of maximizing happiness, that makes these acts right by giving them the different property of being right.

(J) would state a substantive normative fact. As we have seen, however, this fact would not support Naturalism, since (J) would give us no reason to believe that rightness is a natural property, or that an act's being right is a natural fact.

To defend their view that rightness is a natural property, Naturalists must appeal to claims that use, not the concept *makes right*, but some version of the concept *right*. If they were Utilitarians, these Naturalists might try to show that

> (B) being an act that maximizes happiness is the same as being right.

If *this* claim were true, Naturalists could say, rightness *would* be a natural property, and an act's being right would be a natural fact.

(B) could not, I believe, be true. When these writers suggest how we might be able to defend some claim like (B), most of their suggested arguments would support, not (B), but (C) and (J). But we can try to suppose that (B) is true, and we can then ask whether,

as many Naturalists would claim, (B) would state a natural fact that was also a substantive normative fact.

The answer, I believe, is No. Suppose first that, when they discuss claims like (B), Naturalists use the indefinable version of the concept *right*. If (B) were true, this claim would then tell us that this version of the concept *right* and the concept *maximizes happiness* refer to the same property. Unlike (G) and (C), however, (B) would not also tell us how this property is related to any *other* property. Unlike the concept *makes right*, the indefinable concept *right* does not have an explicit gap that is waiting to be filled, so (B) does not refer to any other property. Since (B) refers only to a single property, (B) could not state a substantive normative fact.

These Naturalists might reply that, given the role that the concept *right* plays in our moral thinking, (B)'s truth would *indirectly* give us substantive information. For example, (B) might indirectly tell us that

> (K) when some act maximizes happiness, or is right, that is the same as this act's being justifiable to others, praiseworthy, and something that we have strong reasons to do.

It is even clearer, I believe, that (K) could not be true. When some act maximizes happiness, that could not be *what it it* is for this act to be justifiable to others, praiseworthy, and something that we have strong reasons to do. But if we can manage to conceive that these phrases all refer to the same property, we should conclude that, like (B), (K) would not state a substantive normative fact. If these phrases all referred to the same property, (K) would not tell us how this property is related to any other property. Only that would be substantive information.

These Naturalists might instead suggest that, given the role of the concept *right* in our moral thinking, (B) would indirectly tell us that

> (L) when some act maximizes happiness, or is right, this fact would make this act justifiable to others, praiseworthy, and something that we have strong reasons to do.

In a different version of this reply, these Naturalists might suppose that (B) uses 'right', not in its indefinable sense, but in a complex definable sense, which includes the *justifiabilist*, *praiseworthiness*, and *strong-reason-giving* senses. (B) might then directly imply (L). If (L) were true, this claim would state a substantive normative fact, since (L) would tell us how the property of maximizing happiness, or being right, is related to the different properties of being justifiable to others, praiseworthy, and something that we have strong reasons to do. My objection, Naturalists might say, would then be answered.

This reply, I believe, fails. For this reply to support Naturalism, Naturalists would have to claim that, of the normative facts stated by (L), at least one is also a natural fact. Since these people are

Non-Analytical Naturalists, they would agree that the concepts *justifiable*, *praiseworthy*, and *what we have strong reasons to do* are irreducibly normative. To defend their claim that (L) states a fact that is both normative and natural, these people would have to argue that at least one of these concepts both refers to some natural property, and does that in some way that does not also refer to some irreducibly normative property. This would be no easier than arguing that the concept *right* refers to some natural property without also referring to some such normative property. So even if (B) would imply (L), this fact would not support Naturalism.

We can now restate the Triviality Objection. We are supposing that, by appealing to various features of our moral thinking, Utilitarians can defend the claim that

(A) acts are right just when they maximize happiness.

We can call this the *fundamental* Utilitarian claim. This claim can be understood in two ways. On Sidgwick's view, (A) is a substantive normative claim, which can be stated more clearly if we add that

(C) when acts maximize happiness, that makes them right.

If Utilitarians are Naturalists, they might reject (C) and claim instead that

(B) maximizing happiness is the same as being right.

But if we add (B), as I have just argued, (A) would *not* be a substantive normative claim. Similar points apply to other, more complex moral theories. According to the fundamental claim of any moral theory,

(M) acts are right just when they have a certain natural property, or one of several natural properties.

If Naturalism were true, because rightness was the same as, or consisted in, one or more natural properties, these fundamental moral claims would not be substantive normative claims. So Naturalists must argue that, if these fundamental claims *were* substantive normative claims, various features of our moral thinking would be hard to explain.

The *opposite*, I believe, is true. When we consider acts that have some natural property, and we ask whether these acts are right or wrong, most of us take this to be a substantive question. We assume that, if acts have this natural property, that might give these acts the *different* properties of being right or wrong. According to Naturalists, there are no such different properties. To defend this view, Naturalists would have to adopt some error theory, claiming that most of us misunderstand morality. On this view, as I have said, Sidgwick, Ross, I, and others would have wasted much of our

intellectual lives.

If these Naturalists were Utilitarian, they might now reply:

> It is true that, on my view, there are not two different properties: maximizing happiness and being right. But that does not show that my view cannot answer your questions. When we ask which acts are right, why do we need to be asking about the relation between two different properties? Why isn't it enough to learn that acts can be right, since acts can maximize happiness, and that is the *same* as being right, or is *what it is* for an act to be right?

If this reply seems plausible, this plausibility may depend in part on the standard Naturalist analogy with certain scientific discoveries. If we ask about the nature of heat, it *is* enough to learn that, when the molecules in some object move more energetically, that is the *same* as this object's being hotter, or is *what it is* for this object to be hotter. But this analogy, I have argued, fails. Though this claim uses two concepts that refer to the same property, it tells us how this property is related to other, *different* properties.

This plausibility may also depend on a misleading feature of the question we are now discussing. We are asking whether, *if* Naturalism were true, fundamental moral claims would be substantive. But Naturalism, I believe, could not be true. And when some view could not possibly be true, it can be hard to judge what would follow if, *impossibly*, this view were true. We can sometimes argue against some view by *supposing* that this view is true, and claiming that this view would then have unacceptable implications. But, in supposing that this view is true, we may also be supposing that this view's implications would *not* be unacceptable. And this view's defenders could reply that, if this view were true, it would be true. These points could not, however, help to show that this view is true.

To illustrate these points, we can return to our crude and partial analogy with conflicting views about God. Suppose some theologian claims that

> (N) God exists, because some people love others, and God is such love.

We can object that, if (N) uses the word 'God' in anything close its ordinary senses, this claim could not possibly be true. God could not be the love that some people have for others. This theologian might reply that, if (N) were true, (N) would be true. God would exist, and be the love that some people have for others. But this could not help to show that (N) is true.

Similar claims apply to Naturalism. I have argued that, if Naturalism were true, there would be no substantive normative facts. That, I have claimed is an unacceptable implication. These Naturalists might reply that, if Naturalism were true, this

implication would *not* be unacceptable, since there would be not be any substantive normative facts. But this reply could not help to show that Naturalism is true.

Most of us believe that, if acts of some kind are wrong, that *would* be a substantive normative fact. If Naturalism implies that there could not be such facts, Naturalists would have to admit that their view conflicts with some of our most deeply held beliefs. And if there were no such facts, morality would have no importance.

13 Soft Naturalism

Though I believe that Naturalism could not be true, it is worth again 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. These Naturalists all believe that 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'. [676] 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. . [677]

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'. [678] Darwall is another Soft Naturalist. On Darwall's view, claims about reasons and reason-implying oughts are irreducibly normative. We have strong reasons, Darwall assumes, to make such claims, even though, when

they are true, such claims 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 normative claims.

If (A) is true, however, Soft Naturalism fails, since Naturalism is the view that

> (B) all normative facts are also, in the reductive sense, natural facts.

(A)'s truth would make (B) false. If instead (B) is true and (A) is false, Soft Naturalism again fails. If all normative facts were also 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 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 apply it to 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 all normative properties 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 quite different concepts. As we have seen, that is true of the concepts *heat* and *molecular kinetic energy*. It would have similar importance, Soft Naturalists might claim, if certain normative and naturalistic concepts both referred to the same property. That would be why, even if all properties and facts are natural, we would need to use some irreducibly normative concepts.

Though this reply is plausible, it does not, I believe, succeed. These Naturalists are right to claim that, when we learn that two concepts refer to the same property, that might give us important information. That is true of the claims that

> (C) heat is molecular kinetic energy,

and that

> (D) being an act that maximizes happiness is the same as

having the property that makes acts right.

These are substantive claims. Return next to the claim that

(E) being an act that maximizes happiness is the same as being right.

Though I believe that (E) could not possibly be true, I can try to suppose that (E) is true, and ask whether, as Soft Naturalists claim, (E)'s truth would then be important. These Naturalists might say that, if (E) were true, this claim would not merely tell us that the concepts *right* and *maximizes happiness* refer to the same property. Given the difference between these concepts, (E) would give us further information. That is how (E) would differ from the trivial claim that

(F) being an act that maximizes happiness is the same as being an act that maximizes happiness.

(E) would give us further information because, unlike (F), (E) uses the normative word 'right'. There are now two possibilities. It might be true that

(G) the further information given by (E)'s use of 'right' is irreducibly normative.

If (G) were true, Naturalism would be false, since (E) would state an irreducibly normative fact. It might instead be true that

(H) this further information consists in one or more natural facts.

If (H) were true, Soft Naturalism would fail when applied to such claims. In stating this information, we would not have to use an irreducibly normative sense of 'right'. Rather, as Hard Naturalists believe, the fact that (E) states could be restated with some non-normative and naturalistic claim. [679] So, on both alternatives, Soft Naturalism fails. We can call this the *Further Information* version of the Soft Naturalist's Dilemma. [680]

This argument's conclusion is not 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, that would make it clearer that Sidgwick, Ross, I, and others would have wasted much of our lives. We have asked what matters, which acts are right or wrong, and what we have reasons to believe, or to want, or to do. If Naturalism were true, we wouldn't even need to ask such questions, since such questions could not have substantive answers.

These remarks do not imply that, if Naturalism were false, *Naturalists* would have wasted much of their intellectual lives. When Naturalists develop theories about *what it is* for acts to be right or wrong, we can often revise these theories, so that they instead make claims about what *makes* acts right or wrong. When so revised, some of these theories make original and plausible claims. And some other Naturalist theories could be revised so that they make original and plausible claims about which are the natural facts that can make events good or bad for people, or good or bad in the impartial reason-involving sense.

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.

14 Hard Naturalism

Some Naturalists would agree that their view is close to Nihilism. According to these people, we don'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 Williams and Falk, 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 them even after full reflection on the relevant facts: or what Brandt calls *cognitive psychotherapy*. Similarly, *we* are rational 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 is also, Brandt proposes, '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 value-based objective theories about 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 Brandt claims, we would lose nothing if we replaced our 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 view with these other views, we should turn to cases in which their implications differ.

As one example, we can imagine some young woman who is afflicted with *anorexia nervosa*.[681] Though this woman knows that she could live a full 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, that 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, *why* I ought to do this thing. 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'. But that is not 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. . . For there is no definitional connection between being good. . . and either action or desire.

Brandt's 'similar' question would be

> Q4: Why should I want only those things that are good or worthwhile?

This is a similarly difficult question, Brandt claims, 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

> Q5: 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 what 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 which 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. Such a trivial claim 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.

If this claim were true, that would have great importance. Brandt's 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. [682]

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 knows 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 knows he would enjoy less. But, on the ordinary view, this claim would be obviously false. To compare Brandt's proposed sense of 'better' with what he calls the ordinary sense, we must consider cases in which these senses conflict. One example is the case of our anorexic woman. Compared with a miserable and early death, it would be clearly better in the ordinary sense for this woman to prefer a full and rewarding life.

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 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 adds, 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 such other views, but merely writes that, since it is unclear what kind of fact such truths might be, 'I shall ignore this possibility'. We cannot defend some view by ignoring the alternatives.

Brandt continues:

> 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, that 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 alternatives to his view. Brandt here supposes 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, this would not be merely an absurd 'conditioned response' to the word 'best'. We would be guided, not by mere words, but by our awareness of the facts that gave us most reason 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 in this sense better, but some other choice would be in Brandt's sense rational. So Brandt's conclusion is that, in such cases, we would have less reason to make the choice that we would have more reason to make.

Since Brandt is an excellent philosopher, why does he make these mistakes? The answer seems to be that, even when he claims to be supposing that one of two choices would be, in the ordinary sense, better, he is not really doing that. It seems to Brandt inconceivable that there could be such irreducibly normative truths. As 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. Brandt would have to claim, for example, 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 Brandt's proposed senses of 'rational' and 'best', his claims would mean only that, in starving herself to death, this woman would be doing what, even after cognitive psychotherapy, she would most want. These claims would be true.

These claims would also be trivial. That is the objection to Brandt's view. When Brandt claims that we ought rationally to do what would best fulfil our desires after cognitive psychotherapy, Brandt means that, in doing what would would best fulfil these desires, we would be doing what would best fulfil these desires. When we use such naturalistic substitutes for normative concepts, our claims would never be absurd because they would be empty, claiming nothing. We could not significantly claim, or think, that this anorexic woman should not starve herself to death.

Brandt's claims illustrate another point. Though some 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. [683] 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'. [684]

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. (1) 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. (2) 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.

Though (1) and (2) are plausible, 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. [685]

Though Jackson's use of 'hedonic value' is not normative, this is another normative claim. If we didn't need normative concepts, as Jackson believes, Jackson would be able to restate this claim without using words like '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 maximizing such value 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 claims could not state natural facts. But there is another way in

which normative claims might be compatible with a wholly naturalistic view.

PART FOUR

15 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 half right. Though Moore saw that normative claims could not be 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 writers, 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 practical reasoning, 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 certain kinds of desire, sentiment, or other *conative attitude*.

This view can take Cognitivist forms. According to

> *Moral Subjectivists*: When we claim that some act is wrong, we mean that we have some disapproving attitude towards this act. [686]

But this view is clearly false. As Sidgwick points out, if Subjectivism 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

> *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 way in which apples can be red or green, jokes can be tedious or funny, 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. [687]

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 Moral Intersubjectivists, I might similarly truly claim that some act is wrong, because such acts are disapproved by most people, though I myself approve such acts. That is not how we think about morality. If we ourselves approve some act, we do not 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 disapprove such acts. This view, though more plausible, also misdescribes most people's moral beliefs and claims. When we claim that some act is wrong, we may believe that everyone, under ideal conditions, would disapprove such acts. But that is not what we mean.

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 make subtler and 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 use of 'true', these writers say, is merely another way of expressing the same attitude as you.

There are two main arguments for Moral 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 someone unless this belief was combined with some independent or pre-existing desire.

These people claim that, if moral convictions were beliefs, it would at least make sense to suppose that we might believe some act to be our duty, without having the independent 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. [688]

As Nagel and others claim, we can reject the Humean Theory. When we come to have some belief, this might cause us to have some new desire. In response to this objection, Humeans might revise their view. Whenever we act in some voluntary way, Humeans might say, we must have wanted to act as we did. But in some cases, we could reply, our having such desires would consist only in our being motivated by some belief, such as the belief that we ought to act in some way. In either of these ways, some belief might motivate us without the help of any independent, pre-existing desire. [689]

Humeans might again revise their view. No belief, they might claim, 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 revised 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 reason to be able to motivate us, we must be beings of a kind whom reason is able to motivate. It is no objection to Kant's view that pure reason could not motivate a snail, or a stone.

Even when so revised, however, the Humean Theory may sufficiently support premise (B). We might have to admit that, if moral convictions are beliefs, it would be at least conceivable that we might have some moral belief without being disposed to be motivated by this belief.

If we have to accept premise (B), we could reject this argument 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 is wrong but was not in the slightest motivated to refrain from acting in this way. If we ask instead whether this person might *know* that this act is wrong, the answer would be Yes. And, in knowing that this act is wrong, this person must in one sense believe that this act is wrong. If we revise premise (A) so that it refers to moral beliefs rather than what we call 'sincere moral convictions', this premise ceases to be true, and the Humean Argument fails. [690]

We have other normative beliefs, such as beliefs about what we should or ought to do in the decisive-reason-implying sense. 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 some act 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 premise (A) is true only because (A) uses the phrase 'sincerely convinced' rather than the word 'believe'.

Another argument for Non-Cognitivism starts as follows:

> (C) Moral claims cannot be explained or restated in non-normative and wholly naturalistic terms.

Therefore

> (D) If these claims were true, they would state facts that were not natural but irreducibly normative.
>
> (E) There are no such facts.

Therefore

> (F) Moral claims could not state facts.

I believe that, though (C) and (D) are true, (E) is either false or at least not clearly true, so this argument does not force us to accept (F).

When other people are convinced by this argument, and accept (F), they have two alternatives. According to Nihilists or Error Theorists, since moral claims are intended to state facts, all positive moral claims are false. According to Non-Cognitivists, we should regard moral claims as not intended to state facts.

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 Error Theory. Ayer's first comment was 'That's what I should have said'. Ayer happily turned to the view that most people misunderstand morality, since most people mistakenly believe that there are objective moral truths. Some later Non-Cognitivists, however, 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 truths but to express certain attitudes.

Some of these writers make a surprising further claim. According to these Non-Cognitivists, though we do not intend our moral claims to state objective moral truths, these claims do, in a way, state such truths. Two such writers are Simon Blackburn and Allan Gibbard, who defend partly overlapping Expressivist theories. [691] By asking what these original and impressive theories achieve, we can reach some conclusions that apply to all forms of Non-Cognitivism.

16 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.' [692]

Expressivist Non-Cognitivists, I shall argue, cannot explain how we can have moral disagreements. Expressivism is more plausible than Subjectivism, since Expressivist theories do not imply that two apparently conflicting moral claims might both be true. But Expressivists cannot explain how, when two moral claims conflict, one of these claims must be false, or be in some other way mistaken. Nor can Expressivists explain what it would *be* for some moral claim to be false, or mistaken.

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. [693]

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. [694]

It is misleading, I believe, to describe such people as *disagreeing*. When two people have conflicting desires, they cannot both get what they want. They may oppose each other, and may even fight. But fights are not disagreements. For people to disagree, they must have conflicting beliefs.

Gibbard claims that we can disagree with people's preferences and acts. [695] 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 normative belief.

Gibbard also claims, more plausibly, that we can disagree with people's decisions. [696] And Gibbard might say that we can disagree with people's acts by disagreeing with their decisions to act as they do.

Though Gibbard discusses our moral beliefs, his main claims are about rationality, and about what we ought to do in the decisive-reason-implying sense. To explain 'what *ought* assertions mean', Gibbard writes, we can say:

> the concept of ought just *is* the concept of what to do. [697]

He also writes:

> The hypothesis of this book is easy to state: *Thinking what I ought to do is thinking what to do*. [698]

Gibbard's phrase 'thinking what to do' is ambiguous. If I said that I was trying to decide what to do, I might mean that I was trying to decide what I *ought* to do. But that is not what Gibbard means, since that would make his hypothesis trivial. Gibbard means:

> Thinking about what I ought to do is trying to decide what I *shall* do.

Gibbard also writes:

> If we understand concluding what to do, then we understand concluding what a person ought to do. [699]
>
> When I speak of concluding 'what to do', understand this to mean coming to a choice. [700]

On Gibbard's suggested view, concluding that we ought to do something is the same as choosing, or deciding, to do it.

This view, I believe, is seriously mistaken. When some of us 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, he could not decide which bale to eat, since he had no reason to prefer either bale to the other. So he starved to death.

Suppose next that, to escape from the fire in your *Burning Hotel*, you must jump into the canal. 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 is not so. Jumping through the other window would, as you know, be just as good. But though you don't believe that you ought to jump through one particular window, you must still decide through which window you will jump.

In most cases, our decisions could be based on normative beliefs. But that does not show that, when we believe we ought to do something, that is the same as 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'. [701] But, even when considering people who are always practically consistent in the sense that they always decide to do whatever 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. On his view, it is *by* ignoring this distinction 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. [702]

We can best explain the concept *ought*, Gibbard suggests, by describing what is involved in making plans, and in changing or 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 merely imaginary cases. We decide what we would do if we knew that we were going to be in someone else's position, and that we would be relevantly like this other person. Suppose you tell me that, if you

were offered some job, you would accept. I might decide that, if I were in your position and were in other ways like you, I would refuse this offer. On Gibbard's account, our plans would then disagree, and we would thereby disagree about what you ought to do.

This account, as Gibbard notes, can be challenged. We can object 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. [703] If such a difference between people's plans is not a disagreement, that would undermine Gibbard's explanation of our normative concepts and beliefs.

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. [704]

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. [705] 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. That change of plan may involve no disagreement with my earlier normative beliefs.

To answer this objection, Gibbard's claim must be qualified in a second way. [706] He might say that, when we change some plan because our preferences change as a result of some change in our normative beliefs, this change of plan involves a disagreement with one our earlier normative beliefs. Though Gibbard's claim would then be true, it would not support his suggestion that, to explain our normative concepts and beliefs, we can appeal to the simpler idea of adopting and changing plans.

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. [707] But this fact does not, I believe, 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 doing what we ought to do. We may know that some other plan would be just as good. To be motivated to jump through one of your two windows, you would not need to believe that this is the window through which you ought to jump. (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.)

I have rejected Gibbard's claim that, when we change some plan because our preferences change without any change in our factual beliefs, we must be disagreeing with one of our earlier beliefs about what we ought to do. When Gibbard turns 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.' [708]

As before, I believe, this claim could not support Gibbard's view. Gibbard is trying to explain normative disagreements by appealing to the simpler idea of disagreements between such plans. Gibbard's suggested explanation assumes that people who have such different plans thereby disagree. Gibbard concedes that such people may not disagree. And 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 only be 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. [709]

17 Can Non-Cognitivists Explain Normative Mistakes?

Even if we can 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. . .? [710]

If such judgments cannot be either correct or mistaken, and merely involve arbitrary choices, there would be no point in trying to decide what we ought to do, since we could not make better decisions. We might as well act on impulse, consult our horoscopes, or toss coins.

Gibbard and Blackburn both argue that, though our normative

judgments express desires, decisions, or other conative attitudes, these judgments and attitudes *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 Realists or Cognitivists 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. [711]

For Gibbard and Blackburn to defend these claims, they must explain what it would *be* for our conative attitudes to be correct or mistaken. Suppose I claim that lying is always wrong, thereby expressing my negative attitude toward such acts. In what sense might my attitude and moral judgment be mistaken?

According to Cognitivists, normative judgments express beliefs which can be true or false. When two people's judgments conflict, at least one of these judgments must be false, since contradictory beliefs cannot both be true.

Non-Cognitivists, Gibbard concedes, cannot make such claims. [712] 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, that 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.

When Blackburn discusses practical conflicts, he writes:

> if our attitudes are inconsistent, in that what we recommend as policies or practices cannot all be implemented together, then something is wrong. [713]

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 mistaken. Such attitudes, Blackburn claims, are one kind of desire, and when two desires cannot both be fulfilled, this kind of inconsistency does not imply that one of these desires must be mistaken. We have many rational desires that cannot all be fulfilled. As Blackburn himself writes, 'desires can be faultlessly inconsistent'. [714]

Blackburn suggests two other ways in which, if we are Non-Cognitivists, we can explain what it would be for people's attitudes

and moral judgments to be mistaken. He first remarks:

> Of course there is no problem in thinking that *other* people may be mistaken. [715]

There *is*, I believe, a problem here. Blackburn's remark suggests that, to explain a sense in which other people may be mistaken, it is enough to point out that we may disagree with these other people. But that is not enough. On Blackburn's account, we disagree with other people when we and they have different conative attitudes, which clash. 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. [716]

But suppose I reject this claim, since I believe that people's headaches can be correct or mistaken. To explain this strange view, it would not be enough for me to say that other people's headaches are 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 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'. [717]

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 claim that some standpoint would be improved in the sense that, if people had this standpoint, their 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 claim 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 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 present 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 mistaken. But this claim would not explain what it would be for my headache to be 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.[718]

As before, in claiming to know that other people have bad opinions, Blackburn 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. [719] As this remark implies, 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. [720]

Blackburn might reply that, on any view, we cannot avoid giving some 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. [721]

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 we can easily 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 such an explanation. Our present conative attitudes cannot be false. And, as I have said, it would not help to

claim that such 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 claims are, I believe, decisive. Andrew Egan adds a more particular objection. [722] 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 suggestion, 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. Blackburn claims that 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 Blackburn's 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 mistaken.

Blackburn might point out that, on his view, 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'. [723] 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. For our beliefs to be knowledge, they must at least be true. And Blackburn cannot explain what it would *be*, on his view, for our judgments to be true.

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 merely expresses a similar attitude. If you said 'Milk chocolate is disgusting', I might say 'That's true'. Suppose that some Cognitivist 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. [724] This Cognitivist 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 plausibly 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. [725]

As Blackburn admits, however, there are some meta-ethical questions that cannot be regarded as internal moral questions. And that is true, I believe, of the central question that we are now discussing. 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 Expressivist 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, on these Non-Cognitivist theories, this question has no answer, since there is *nothing* that it could be for any moral judgment to be true or false, correct or mistaken.

Blackburn tries to avoid this conclusion. Making his 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. [726]

This claim, however, is false. 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 Realists or Cognitivists appeal. This, moreover, is Blackburn's view. Blackburn himself 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. . [727]

> . . . 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. [728]

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. [729]

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 legitimately 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. [730]

Blackburn also claims that, as a Quasi-Realist Expressivist, he doesn't need to give an informative account of moral error, or of what it would be for some conative attitude or judgment to be false. In Blackburn's words:

> If some theorist. . . asks me what my account of moral error itself is, then I am not very forthcoming. Why should I give one? . . 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 I recognize to the question 'what is it for happiness to be good?' is happiness being good. [731]

As we have seen, however, 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. . . [732]

As Blackburn saw, Quasi-Realists need to *earn this right*. On Blackburn's view, though our moral judgments fundamentally express certain kinds of desire or 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, he persuasively defends our right to call some of these beliefs true.

In the longer passage just quoted, however, Blackburn merely asserts that we have such a 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 say '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. I might similarly say that we judge truly that some headache is mistaken if this headache is mistaken. For Expressivists to *earn* their right to talk of moral truth, they must explain what it would be, on their view, for us to judge truly that some act is wrong. That is why I could not defend Quasi-Realism about my headache judgments. I could not earn a right to call these judgments true, because I could not explain what it would be for us to judge truly 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'. [733] Blackburn's explanation, I have argued, fails. Though Expressivists can mimic what Realists *say*, they cannot think or believe what Realists think or believe.

We can draw a wider conclusion. Suppose that, as Moral Sentimentalists believe, morality essentially involves certain 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 the simplest view of this kind, moral claims are true when the attitudes they express are correct, and false when the attitudes they express are mistaken. We can reject Realism only if these attitudes cannot be correct or mistaken. Only then should we believe that moral claims cannot be true or false, and merely express such conative attitudes. Quasi-Realist Expressivists therefore face a dilemma. To defend their Quasi-Realism, these people must claim that our conative attitudes can be correct or mistaken. To defend their Expressivism, these people must claim that these attitudes cannot be correct or mistaken. These people must therefore claim that these attitudes both *can* be, and *cannot* be, correct or mistaken. Since that is impossible, no such view could be true.

18 Expressivism

Blackburn and Gibbard might give a different reply. There is a reason, Blackburn writes, why Expressivism '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'. [734] To provide answers to practical questions, normative judgments cannot be beliefs about such normative facts, but must be conative attitudes.

Gibbard similarly claims that, when applied to the judgments with which we make decisions, 'expressivism has to be right'. [735]

According to *Non-Naturalists* like Sidgwick, asking what we *ought* to do is not the same as asking *what to do*. Gibbard writes that, if these were different questions, asking what we ought to do could not help us to decide what to do. Non-Naturalists

> just change the subject. We ask what to do, and they hand us analyses of a different question. [736]

Like Blackburn, Gibbard here claims that normative facts could not answer practical questions.

This 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. [737] Though we might have true beliefs about what we ought to do, such beliefs would add nothing.

This view, I believe, is seriously mistaken. If your room is in the top storey of your hotel, and you are terrified of heights, you may need to believe that you ought to jump into the canal, and that if you don't jump, you would be making a terrible mistake. It may be only such beliefs that would enable you to jump, and thereby save your life. And in many other cases, we would make better decisions, and act in better ways, if we try to answer questions about what we should or ought to do, in the decisive-reason-implying senses.

Blackburn and Gibbard may make these claims because they assume that there could not be any irreducibly normative truths. Blackburn writes that an Expressivist

> affirms *all that could ever properly be meant* by saying that there are real obligations. [738]

Cognitivist moral claims, Blackburn suggests, are incoherent, or unintelligible. [739] When Gibbard discusses Sidgwick and Moore, he calls their views 'Platonistic or intuitionistic'. Since Sidgwick and Moore were not Platonists, we can call them *Rational Intuitionists*. On such theories, Gibbard writes,

> 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. [740]

This description is misleading. When Gibbard writes that, on such views, facts about what is rational are 'among the facts of the world', this remark suggests that, according to these Intuitionists, the rationality of some desire or act is a contingent feature of the spatio-temporal world. And, when Gibbard talks of a faculty analogous to sense perception, that suggests that, according to these Intuitionists, 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.

Rational Intuitionists need not hold such views. As Gibbard elsewhere notes, Sidgwick and others claim that we can understand and recognize some irreducibly normative truths, such as truths about practical or epistemic reasons, in something like the way in which we can understand and recognize some mathematical or logical truths. These two kinds of necessary truth are not part of 'the fabric of the world', with which we causally interact. And, in claiming that we can recognize such truths intuitively, Sidgwick is not appealing to any special faculty that is like sense perception. He means only that there are certain beliefs that we are justified in having, not because they are implied by other beliefs, or because we have evidence for them, but because of their content, or *what* we are believing. Such beliefs, Sidgwick adds, may be mistaken.

Not only are normative truths, on this view, partly like logical or mathematical truths. Logical and mathematical reasoning both involve awareness of such normative truths. We can recognize, for example, that if it is true both that *P*, and that *If P, then Q*, these facts give us a decisive reason to believe *Q*. Besides being able to understand and recognize such truths about what we have reasons to believe, we can understand and recognize some truths about what we have reasons to want and to do. One example is the truth that we have reasons to want to avoid being in agony.

Gibbard sees no need to argue against Sidgwick's view. He 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 and these powers of non-sensory apprehension. [741]

To non-philosophers, Gibbard adds, such claims 'sound fantastic'. In several other recent books, views like Sidgwick's are dismissed in a paragraph or two. Jackson thinks it worth explaining why he even bothers to discuss such views. [742]

Such dismissals are too brisk. Most of us would hesitate to claim that we have non-sensory powers of apprehending non-natural

facts. But, as I have said, Gibbard misdescribes the Intuitionist account of how we can recognize such facts. Most of us believe that we can have reasons to have certain beliefs and desires, and to act in certain ways. Nor is it fantastic to suggest that such beliefs can be both irreducibly normative, and in a strong sense true.

I shall not try here to defend this kind of view. Blackburn writes

> there is precious little surprising left about morality: its meta-theory seems to me pretty well exhaustively understood. [743]

This meta-theory seems to me very far from being understood. There are, I believe, several relevant and fundamental questions that we haven't answered, and there are some such questions, I assume, that we haven't even asked. Before we understand these questions better, and have made more progress in trying to answer them, we should not have firm beliefs about whether views like Sidgwick's are true. [744]

My aim, in this appendix, is only to argue that views like Sidgwick's should at least be taken seriously. Naturalism and Non-Cognitivism are both, I am arguing, close to Nihilism. Normativity is either an illusion, or involves irreducibly normative truths.

I have mentioned Gibbard's claim that, even if acts could have the non-natural property of being what we ought to do, that could not help us to decide what to do. Since Gibbard is an excellent philosopher, who follows his arguments wherever they lead, he also makes some other, strikingly different claims. For example:

> Anyone who reasons what to do, I argue, is committed to something very much like facts of what to do. Reasoning what to do commits us to thinking in terms of conclusive to-be-doneness. It commits us to this thinking's being very much like thinking of properties. In our reasoning to decisions, we must think very much as if there were properties that include, in some queer way, conclusive to-be-doneness. [745]

Though Gibbard starts by claiming that the concept *ought* just is the concept of what to do, he ends by claiming that, in deciding what to do, we must think in terms of facts about what is *to-be-done*, or what we ought to do. We must use the 'non-naturalistic' normative concepts that people like Sidgwick use. In Gibbard's words, 'in a sense I end up as a non-naturalist about *oughts'*. [746]

When Gibbard first gave his account of rationality, he made some other relevant claims. I ought, for a different reason, to discuss this account. 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 am assuming that we need to know what it would be for such judgments to be 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's main question, he writes, is '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 giving some other description of this property. 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. [747]

Since Gibbard's account takes this indirect, Expressivist form, it is in one way harder to decide whether this account succeeds. Gibbard gives us no definition that we could try to assess. We can ask only whether Gibbard gives a good description of our state of mind when we judge that something is rational.

In considering this question, we can start with some remarks about expressivist accounts of meaning. 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. When we are trying to explain some belief, nothing is gained by switching to an expressivist account. 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 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'. [748] Imperatives are commands, like 'Do not be angry with such people!', 'Keep your promises!' or 'Never lie!' Imperatives 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 some way. (It is no objection to this claim that, when some legitimate authority commands us to act in some way, we may 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 'concluding *what to do*'. [749]

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.' [750]

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.' [751]

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? [752]

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. [753]

This reply, I believe, fails. Like many great philosophers, Gibbard tries to have things both ways. On Gibbard's view, acts cannot truly be rational. And 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 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 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. [754]

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, this 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. [755]

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. [756]

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.' [757] 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 also 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.' [758] 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 writes, '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 can 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.' [759]

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 says that this claim does not leave 'normative language defective, or second rate'. [760] 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 concedes that his view does not preserve some of our beliefs. In his words, it may 'end up missing something in our ordinary claims to objectivity.' But he continues:

> Norm expressivism is meant to capture whatever there is to ordinary notions of rationality if Platonism is excluded. . . My hope, then, is to save what is clear in ordinary thought about rationality. . . with one exception: our wavering penchant for Platonism. [761]

We should remember here that, by 'Platonism', Gibbard is not referring to some extravagant metaphysical view. He is referring to Sidgwick's view that our beliefs about reasons and *oughts* can be both irreducibly normative and in a strong sense true. As I have said, Gibbard does not argue against this view. He thinks it enough to call it 'fantastic', and a view that is not required by 'a plausible, naturalistic picture of our place in the universe'.

Remember next that, on Gibbard's naturalistic picture, beliefs about reasons and oughts are replaced by internalized imperatives. We share such imperatives with other social mammals, such as wild dogs, but because we use language we have one further 'biological adaptation': we 'avow' such imperatives in 'normative discussion'. And there are no truths about we have reasons to want or do. This view is not fantastic. But it is bleak. Gibbard says he 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 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. . ? [762]

On Gibbard's view, I have argued, there is nothing to discover. We could never be mistaken in our judgments about how it would be best to live, since this *is* just 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 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.

These remarks do not refute Gibbard's view. Perhaps the truth *is* bleak. But we would need more reason to accept this view than Gibbard gives us when dismissing Sidgwick's view. It is not enough to say: 'If this is what anyone seriously believes, then I simply want to debunk it.'

19 Hare on What Matters

Hare would deny that, on his version of Non-Cognitivism, the truth is bleak. A young Swiss guest of 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?' [763]

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, and 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, nothing can be truly described 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.'[764]

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, objectivists believe that some 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 subjectivists 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 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 true or false.

Hare might give a different reply. He might concede that, when objectivists claim that some act is wrong, 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. [765]

As this passage shows, Hare ignores the possibility that there might be normative truths. If moral judgments were capable of being true, or stating facts, they could not, Hare claims, guide actions. But, if we judged that we ought to do something, that judgment could guide our acts. So Hare must assume that judgments like 'I ought to do that' could not conceivably be true.

20 Normativity and Truth

Many other writers ignore the possibility that there might be normative truths. Patrick Nowell-Smith, for example, 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'. [766] 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 would not be answers to my question that conflict with mine. 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 it be, as Nowell-Smith claims, 'the fundamental question of ethics'. The fundamental question is: 'What *should* I

do?', or 'What *ought* I to do?' Since that question *is* normative, it might have answers that moral philosophy, or other people, could give, or help us to find. 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 ought to do. In answering moral questions, we would then be answering what Nowell-Smith calls our practical question, by deciding what to do.

Rather than merely assuming that there could not be normative truths, many Non-Cognitivists make the bolder claim that, even if there were such truths, they could not answer normative questions.

This claim is often made 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?[767]

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"'. But 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. [768]

Suppose that we knew another fact about the Universe, 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?' [769] 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. [770]

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 we should do something, it is not an open question whether we should do this thing.

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*. [771] 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. [772]

Practical reasoning, Korsgaard 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. ' [773] On their view, 'our relation to reasons is one of seeing that they are there or knowing truths about them.' [774] Realism fails, Korsgaard claims, because no knowledge of truths about reasons could answer normative questions.

In this and other passages, Korsgaard's objections to normative realism seem to be these:

> (A) Realists discuss the wrong question.
>
> (B) Realists may not be able to convince us that some answer to our question is really true.
>
> (C) Even if our question had some true answer, that would not solve our problem.
>
> (D) 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. [775] 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 could 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. [776]

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 here suggests that, if realism were true, we might have no reason to want, and do, what 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 could 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.

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 what makes this person's normative belief *motivate* her. As Korsgaard next writes

> We must explain how these reasons *get a grip* on the agent. [777]

If Korsgaard is using 'normative for her' to mean '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 decisively or 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 here 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 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 might be motivated or moved to act. On this view, as I have said, normativity is wholly different from, and doesn't even 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. [778]

Korsgaard here assumes that, in judging something to be desirable, we are judging that this thing is what we want most. If that is 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 most 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 sufficient reasons to have our aims, we would not, as Korsgaard claims, run out of answers.

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 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 quite unnecessary.

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, '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 be ended only 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?' [779]

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 it is the content of certain beliefs that provide the answers to 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.

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 he will torture me for ten minutes unless, one minute 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. [780]

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*, he will torture me unless, one minute 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 he will torture me 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 would 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 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, in the hour that this man has generously given me, 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 belief, that would cause us to have this 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 goodness of having 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 for ten minutes. I might ask whether, compared with being tortured for ten minutes, 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.

State-Given Theorists might now 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. [781] 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. Theoretical reasoning is a voluntary activity, in which we often engage for practical reasons. And such reasoning involves many more particular voluntary mental acts. 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, to confirm the results of some earlier calculation. 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 come to believe, that, since something is true, something else must be true. But coming to have such a belief is not a voluntary mental act.

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. Suppose that, when my hotel is on fire, I could save my life only by jumping from my bedroom window into some canal. This fact would give me a decisive reason to jump, and a decisive reason to believe that I 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 he will torture me for ten minutes unless, one minute 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 minutes. 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. [782] I tried to refute that argument. [783] Since Gauthier was not convinced, I shall try again. [784]

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. [785]

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. [786] 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 if she broke her promise. Either she could break it secretly, or the damage to her reputation would be outweighed by what she gains. 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. [787] 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. [788]

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. [789]

I challenged (A) as follows. [790] 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. [791]

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 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. [792] 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'. [793]

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. [794] 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. [795]

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? [796]

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. [797] If I could lose this disposition, it would be irrational 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. [798]

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. [799]

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. [800]

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. [801] 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. [802]

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. [803] 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. [804]

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 is rational for us to do does not depend, in this way, on what is causally possible. [805] 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. [806]

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 kill my children, he will not induce me to unlock my safe. That will give him reasons to give up and leave, which will be much better for me. [807] 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. [808]

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. [809]

This view is very 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. [810] That is hard to believe.

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*. [811]

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. [812]

When we cannot maximize at both levels, it will be better for us if we have maximizing dispositions. The good effects of these dispositions will outweigh the bad effects of our acts. [813]

Gauthier claims that, given this fact, it will be better for us if we accept not E but G. [814] 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.[815] 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.[816] 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. [817]

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. [818]

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. [819] 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. [820]

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. [821]

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. It might 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. [822]

Suppose that, for these or other reasons, it would 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. [823]

Shelly Kagan suggested the following objection. [824] 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. [825]

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'. [826] 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)? [827] 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. [828]

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. [829]

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.' [830]

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. [831]

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. [832] 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.' [833]

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 argue that, even in self-interested terms, we are rationally required never to act wrongly. No other moral theory, Gauthier claims, achieves this aim. [834] If Gauthier's argument fails, as I have claimed, we lose our main reason to accept Gauthier's minimal morality.

APPENDIX D 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. [835]

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. [836]

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. [837]

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. [838]

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'. [839] 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. [840]

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 most 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'. [841] 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. [842]

2

We can now turn to some of Kant's arguments for his Formula of Universal Law, which Kant sometimes calls his *Formal Principle*.

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.[843]

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. [844]

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 is both invalid and unsound. 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. [845]

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. . . [846]

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. [847] 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 even 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 substantive 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 is 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.[848]

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. [849]

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 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. [850]

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.[851]

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. [852]

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. . . [853]

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.[854]

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.[855]

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'. [856] 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.[857]

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. [858]

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. . .[859]

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. [860]

The *Second Critique* contains another version of Kant's 'refutation'. [861] 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. . .[862]

In this passage, Kant argues:

(1) 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.

(2) 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.

(3) 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.

(4) The principle of obeying God's will also cannot move us to act except through the expectation of our own happiness.

Therefore

(5) These principles are material, and are the only possible material principles.

(6) Material principles cannot be moral laws.

Therefore

(7) Kant's Formula is the only moral law.

This argument is both invalid and unsound. 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 might have defensibly, though not undeniably, claimed to be the supreme moral law.

APPENDIX E 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*).[863]

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. [864]

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. . .[865]

As Terence Irwin notes, Kant misunderstands this Stoic claim. [866] 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. [867]

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. [868]

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. [869]

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. [870]

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. [871]

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. . . [872]

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 F 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**. . .[873]

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 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* some principle to ourselves *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. [874]

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		
and either	only if and because we have willed that to be so	or	whether or not we have willed that to be so
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. [875]
>
> reason alone. . . gives the law. . . [876]
>
> 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. . . [877]

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. . .' [878] 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. [879]

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. [880] 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. . . [881]

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.' [882] 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. [883]

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. [884]

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'. [885] 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? [886]

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'. [887] 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. [888]

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.

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. [889] 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. [890]

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. [891]

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 G.

APPENDIX G 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'. [892]

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. [893] And Kant later writes that hypothetical imperatives say 'what one must do in order to attain some end'. [894] 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. . .[895]

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 did 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. . .[896]

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] [897], 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. . . [898]

Kant defines a 'determining ground' as 'the motivating cause' of an act. [899] 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'. [900]

(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. [901]

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. [902]

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. [903]

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*. [904]

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? [905]

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.[906]

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 motivates 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. [907]

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), according to which 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. [908]

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. [909]

> Everyone does in fact appraise actions as morally good or evil by this rule. [910]

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. [911]

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 this 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. [912]

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. . .[913]

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. [914]

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. [915]

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. [916]

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. [917]

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. [918]

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.[919]

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. [920] But this cannot be all that Kant means. When Kant asks 'whether the will has only empirical or also pure determining grounds', [921] 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. [922]

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. [923]

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. [924]

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. [925]

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. [926]

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*. [927]

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. . . [928]

> 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. [929]

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. [930] 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. . . [931]

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.[932]

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. . .[933]

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. [934]

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. . .[935]

> 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. . . . [936]

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. [937]

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. [938]

> 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). [939]

> 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. [940]

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. [941]

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 reading of Kant's Formal Principle, acts are wrong unless they are permitted by principles whose universal acceptance everyone could rationally will. This claim implies that

> (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*. [942]

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? [943]

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. [944]

> . . . 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.[945]

> 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. [946]

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.[947]

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.[948]

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.[949]

Though Kant describes necessitation as the relation which is expressed by 'ought', this remark treats this relation as the bringing of an act about. 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. [950]

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. [951]
>
> duty. . . lies. . . in the idea of a reason determining the will by means of a priori grounds. [952]
>
> Practical good. . . is that which determines the will by means of representations of reason. . . [953]
>
> The concepts of *good* and *evil*. . . are. . . modi of a single category, namely that of causality. . .[954]

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. [955]

In some passages, Kant seems to forget this warning. But as I have said, Kant should not have claimed that consistency is the

greatest obligation of a philosopher. [956] It is more important to have ideas that take us closer to the truth.

243,176 words

---

[1] In these opinions I follow C. D. Broad, *Five Types of Ethical Theory* (Littlefield, Adams, and Co, 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] *Henry Sidgwick*, *A Memoir*, by A.S. and E. M. S henceforth *M* (Macmillan, 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), and Bart Schultz's fascinating *Henry Sidgwick: Eye of the Universe, an Intellectual Biography* (Cambridge University Press, 2004).

[7] M 396.

[8] M 92.

[9] M 170-1.

[10] 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'.

[11] ME 298-9 (my italics).

[12] 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".'

[13] ME 507 note.

[14] 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.'

[15] *Essays on Ethics* 118.

[16] ME 295.

[17] ME 437.

[18] ME 248 note.

[19] ME 490.

[20] ME 284.

[21]

[22] M 74.

[23] 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'.

[24] 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.'

[25] M 421.

[26] Hastings Rashdall, in his review of Sidgwick's *Elements of Politics*, in the *Economic Review* 2, 1892.

[27] Broad *Five Types*, *op.cit*. 14.

[28] 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).

[29] Onora O'Neill *Constructions of Reason* (Cambridge University Press, 1989) 126. (She is discussing only Kant's books on practical philosophy.)

[30] Norman Kemp Smith 'Kant's Method of Composing the Critique of Pure Reason', *The Philosophical Review* 1915, 531.

[31] *Religion within the Bounds of Bare Reason*, 72.

[32] 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). Because philosophy makes progress, we can now see mistakes made by earlier philosophers who were much smarter than us.

[33] Kemp Smith *op. cit*. 527. Though this is a remark about Kant's *First Critique*, it also applies, I believe, to Kant's books on ethics.

[34] *Second Critique* 24.

[35] *Lectures* 127 and 148.

[36] *Lectures* 369.

[37] *Lectures* 44.

[38] Christine Korsgaard, *Creating the Kingdom of Ends*, henceforth *CKE* (Cambridge University Press, 1996) 126.

[39] MM 429-30.

[40] M 177.

[41] 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' (John Ruskin, *The Works*, edited by E.T.Cook and Alexander Wedderburn (London, 1903), xi. 381.)

[42] 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'.

[43] *Lectures* 18.

[44] See Nagel's wonderful *The View from Nowhere* (Oxford University Press, 1986), especially Chapter VIII, and his *The Last Word* (Oxford University Press, 1997).

[45] If we ask 'Is that true?', either answer would be astonishing. There are not many questions of which that could be claimed.

[46] I follow Thomas Scanlon, *What We Owe to Each Other*, henceforth WWO, (Harvard University Press, 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.

[47] 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.

[48] Reasons can be related in more complicated ways. Some facts give us reasons, for example, to ignore some other kinds of reason. And some facts give us reasons, not separately, but only when combined with certain other facts. I shall mainly be discussing simpler, more fundamental questions.

[49] Like the concept of *a reason*, the concept expressed by these senses of 'should' and 'ought' cannot, I believe, be helpfully defined in other terms. 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.

[50] 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.

[51] To be fully rational, we may also need to respond to certain rational requirements, even in cases that do not involve reasons or apparent reasons. In this book I shall say little about such requirements. Some people claim that, to be rational, we don't need to respond to reasons, or apparent reasons, since it is enough to avoid certain kinds of inconsistency between our mental states. According to another widely held view, the rationality of our desires and acts depends on the rationality of our beliefs. I shall question these views in Chapter 4.

[52] 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.

[53] 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.

[54] P. G. Wodehouse *Pigs Have Wings* (Ballentine Books, 1952) 93.

[55] WWO 97. Scanlon calls this the *buck-passing view*.

[56] Though Scanlon claims that goodness and badness are not reason-giving properties, he sometimes mentions what I am calling derivative reasons. He 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' (WWO 181).

[57] 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.

[58] CKE 225.

[59] In these remarks I partly follow Christine Korsgaard's, 'Two Distinctions in Goodness', in *Creating the Kingdom of Ends*, henceforth *CKE* (Cambridge University Press, 1996). Gilbert Harman draws further relevant distinctions in his *Explaining Value* (Oxford University Press, 2000) Chapter 8.

[60] We are here appealing to the normative reasons which have become our motivating or belief-producing reasons. I follow Scanlon's claims in WWO 18-22.

[61] 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 could be claimed to have desire-based state-given reasons to have some desire when, and because, we want to have it.)

[62] For a different view, see Stuart Rachels '"Is Unpleasantness Intrinsic to Unpleasant Experiences?" *Philosophical Studies*, Vol. 99, No. 2, May (II) 2000.

[63] I discuss these and other attitudes to time in Sections 62-70 of my book *Reasons and Persons* henceforth *RP* (OUP 1984-7). (In that rather tortuous discussion I failed to make it clear that, in my view, the most rational attitude is temporal neutrality.)

[64] 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, David Chan 'Are There Extrinsic Desires?' *Nous* 38: 2 2004.

[65] References to Michael Smith, Julia Markovits, and others.

[66] Given the meaning of 'procedurally rational', we cannot sharply define this ordinary, thin sense. The distinction between these kinds of theory is, in a way, only a matter of degree. There is a much clearer distinction between theories that assert or deny that we have object-given reasons.

[67] As before, I follow Scanlon, WWO Chapter 1. See also Joseph Raz, *Engaging Reason* (Oxford University Press, 199) Chapter 2.

[68] I follow Scanlon, [reference]. See also Dennis Stampe [reference] and Stephen Schiffer [reference].

[69] 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

[70] Reference to Bratman.

[71] Kant's *Lectures*, 58-9 (Prussian edition 27: 264-5), and Sidgwick's ME 74-5.

[72] For an excellent account of such reason-giving facts, see Niko Kolodny, 'Love as Valuing a Relationship', *The Philosophical Review*, 2003.

[73] I could refer to the new appendix here.

[74] Williams draws this distinction in his 'Internal Reasons and the Obscurity of Blame' in his *Making Sense of Humanity* (Cambridge University Press, 1995) 36-7.

[75] *Political Liberalism*, (Columbia University Press, 1996) 49. Rawls writes 'rational' because, as an Internalist, he takes rationality to be purely procedural. In his words, 'There is no way to get beyond deliberative rationality' (TJ 560).

[76] *Philosophy as a Humanistic* Discipline, 111.

[77] This claim is made, for example, by Richard Hare and Richard Brandt.

[78] See Guy Kahane, [reference to his thesis].

[79] 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.

[80] Even Hume claimed only that such desires or preferences would not be *contrary* to reason.

[81] 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. . .

[82] For similar objections to theories of this kind, see David Enoch 'Why Idealize?' *Ethics* 115 (July 2005): 317-345.

[83] CKE 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.

[84] Here is another, similar point. According to many desire-based subjective theories, we have most reason to fulfil the desires that we would now have, or would want ourselves to have, if we had carefully considered all of the *relevant* facts. According to some writers, any fact is relevant if our awareness of this fact would cause us to have, or to lose, some desire. Other writers object that, on this assumption, desire-based theories would have implausible implications. If we carefully considered certain facts, such as facts about appalling atrocities in previous centuries, that might cause us to lose many of our desires. If we vividly imagined what was taking place in the innards of our fellow diners, we might lose our appetite. These writers claim that, when we apply desire-based theories, we should ignore such facts. We should count, as *relevant*, only facts about the intrinsic features of the possible outcomes of our acts. But, to explain why these are the relevant facts, we must claim that only these facts give us reasons to want to produce, or prevent, these outcomes, if we can. And Subjectivists about Reasons cannot make that claim.

[85] 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.

[86] John Rawls *A Theory of Justice*, (Harvard University Press, 1971) henceforth TJ, 395.

[87] TJ 417.

[88] As Sidgwick notes in *The Methods of Ethics*, henceforth *ME* (Macmillan and Hackett, various dates) 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?

[89] TJ 408 (my italics).

[90] TJ 184-5, RE 161.

[91] TJ 432.

[92] TJ 401. Cf 111 and 451.

[93] Rawls's thin theory of the good is strongly subjective, because it defines the best life for someone as depending entirely on a single, present choice. A temporally-neutral desire-based definition is less subjective, because it appeals to all of someone's desires throughout this person's whole life. There are, I have claimed, three kinds of *substantive* theory about well-being. Such theories can be hedonistic, or appeal to desire-fulfilment, or appeal to a list of substantive goods, which will include happiness and the avoidance of pain and suffering, but add such other things as mutual love, moral goodness, some kinds of achievement, or knowledge. These might be called *Objective List Theories* of well-being. Hedonism is best regarded, I believe, not as a subjective theory, but as an Objective List Theory with a very short list. Of these three kinds of theory, it is temporally neutral desire-fulfilment theories that are most subjective. But, as *Blue's Choice* shows, these theories are much less subjective, and much closer to the other two kinds of theory, than Rawls's *present choice-based* theory.

[94] Rawls makes some remarks which suggest that he accepts, not his present-choice-based theory of the good, but the hedonistic and temporally neutral desire-based theories. He writes, for example, 'future aims may not be discounted solely in virtue of being future' (TJ 420'). This suggests that the best life for someone is not the life that he would now choose with full deliberative rationality, but the life in which all of his aims at all times would, on the whole, be best fulfilled. Rawls also writes 'one feature of a rational plan is that in carrying it out the individual does not change his mind and wish that he had done something else instead' (421). It is not clear how to include this feature in Rawls's theory of the good. MORE.

[95] In Harry Frankfurt's words, we 'need to understand what is *important*, or, rather, what is *important to us*' (*The Importance of What We Care About* (Cambridge University Press, 1988) 81, and 91 note 3.

[96] 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 (Robert Nozick, *The Nature of Rationality* (Princeton, 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.

[97] This argument is suggested, for example, by Williams's remarks in 'Internal and External Reasons', 102 and 106-7, and in 'Internal Reasons and the Obscurity of Blame', 39. For a longer discussion of such arguments, see my 'Reasons and Motivation', *Proceedings of the Aristotelian Society, Supplementary Volume*, 1997.

[98] Stephen Darwall, Allan Gibbard, and Peter Railton 'Toward Fin de Siecle Ethics: Some Trends, *The Philosophical Review*, January 1992, 176-7.

[99] 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 cannot be irreducibly normative. These people believe that, to preserve the normativity our normative claims, we should not regard these claims as intended to state *truths*. For a discussion of these views, see Appendix A.)

[100] 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.

[101] In these remarks, I follow Thomas Nagel, *The View from Nowhere*, henceforth *VFN,* (Oxford University Press, 1986) 141, and his *The Last Word* (Oxford University Press, 1997).

[102] 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.

[103] 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.

[104] Hume, for example, 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.

[105] It might be suggested that we should not distinguish between these two kinds of reason and rationality. Rather than claiming that my smoking is *epistemically* irrational, we would then merely claim that this act is irrational. But this claim would still be misleading, since it is not in *smoking* that I am failing to respond to my reasons to *believe* that smoking will be likely to damage my health.

[106] 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.)

[107] As Scanlon argues, in WWO, Chapter 1.

[108] WWO 29-31.

[109] WWO 29.

[110] I have supposed that Scarlet has a strange theory about his *reasons* to promote his future well-being, since that is what is relevant here. When Scanlon supposes that Scarlet has 'some strange theory of well-being', he may be supposing that, on Scarlet's theory, agony won't hurt if it is felt on a Tuesday, or something of the kind. I would agree that, if *that* is why Scarlet has his preference, this preference would not be practically irrational. It is not irrational to prefer mild pain to what we believe won't hurt at all.

[111] 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, see my *Reasons and Persons*, sections 63 to 70, 102 to 105, and Appendix F.

[112] WWO 25-30. Add comments on Broome.

[113] For some examples, see Appendices B and C.

[114] Comment on contrary claims and arguments made by Temkin and Rachels.

[115] 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).

[116] ME, Concluding Chapter. This is only part of Sidgwick's view. Sidgwick makes other claims, to which I shall turn in Section 16.

[117] 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).

[118] John Findlay, *Values and Intentions* (George Allen and Unwin, 1961) p 294. Compare Rawls's claim: 'Utilitarianism does not take seriously the distinction between persons' (TJ 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.

[119] In Thomas Nagel, *The View from Nowhere* henceforth *VFN* (Oxford University Press, 1986) especially chapters VIII and IX, and *Equality and Partiality* (Oxford University Press, 1991) Chapter 2.

[120] 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 by Jerome Schneewind, *Sidgwick's Ethics and Victorian Moral Philosophy* (Oxford University Press, 1977) 369.) Nagel calls 'the transcendence of one's own point of view. . . the most important creative force in ethics' (VFN, 8).

[121] 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).

[122] ME 508.

[123] For a discussion of these reasons, see Niko Kolodny, 'Love as Valuing a Relationship', *The Philosophical Review*, 2003.

[124] VFN 160.

[125] Some people would add 'unless this person deserves to be in pain'.

[126] 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).

[127] VFN 161.

[128] 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 (*Moral Luck*, 45). But if some act is generous and fine, that gives us *some* reason to act in this way.

[129] 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.

[130] ME 386 note 4.

[131] *The Works of Thomas Reid* (Georg Olms Verlag, 1983) 598.

[132] 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.

[133] 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'.

[134] TJ 575.

[135] This is forcefully argued, for example, by Niko Kolodny in 'Why be Rational?', Mind, 2005 Volume 114.

[136] See again Nagel's *The Last Word*.

[137] 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.

[138] Some Act Consequentialists claim, for example, 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.

[139] See, for example, Thomas Nagel's 'Moral Luck' in his *Mortal Questions* (Cambridge University Press, 1979).

[140] There is a quite 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.

[141] See Jonathan Bennett, 'The Conscience of Huckleberry Finn' *Philosophy*, Vol.49, No 188. (April, 1974) 123-134.

[142] 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').

[143] 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.

[144] ME 207-8.

[145] 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.)

[146] 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.

[147] ME 382-3.

[148] ME 200, 403.

[149] 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.

[150] 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' *Essays on Ethics*, 607. In this passage, Sidgwick's (1) seems to suggest Impartial-Reason Consequentialism.

[151] 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.)

[152] 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-involving 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.)

[153] WWO, 97.

[154] Rawls, TJ 52; Nagel, *Other Minds* (Oxford University Press, 1995) 182.

[155] *The Groundwork*, henceforth *G* 392. Page references are to the page numbers of the Prussian Academy edition, which are given in most English translations.

[156] 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).

[157] G 430.

[158] CKE 139.

[159] Onora O'Neill, *Constructions of Reason*, henceforth *CR*, (Cambridge University Press, 1989), 111.

[160] CKE 140.

[161] I here follow CKE 295-6.

[162] 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.

[163] CR 110.

[164] 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' (CKE 138-9).

[165] Other writers have assumed or claimed that this is what Kant means. See, for example, Thomas Hill, *Dignity and Practical Reason*, henceforth *DPR*, (Cornell University Press, 1992) 45.

[166] That seems often true, for example, when Kant claims that we could not will that some maxim be a universal law.

[167] G 429-30, my italics.

[168] John Rawls, *Lectures on the History of Moral Philosophy*, henceforth *Lectures*, edited by Barbara Herman (Harvard University Press, 2000) 100-91. A similar claim is made by Hill in DPR 45.

[169] G 436.

[170] Rawls, *Lectures*, 191 and 182-3.

[171] *Critique of Practical Reason*, henceforth *Second Critique*, note on p.8.

[172] Barbara Herman, *The Practice of Moral Judgment*, henceforth *PMJ* (Harvard University Press, 1993) vii.

[173] The Consent Principle would also imply that it would be wrong for me not to save Blue's life, since Blue could not rationally consent to my failing to save her life. So this principle would mistakenly imply that I cannot avoid acting wrongly.

[174] Things might be different if Blue was old and Grey was a young professional dancer. Blue might then have sufficient reasons to consent to our saving Grey's leg rather than Blue's life, since Blue's loss might here be no greater than Grey's. This is the kind of morally relevant further fact that, in considering my examples, we should suppose would not obtain.

[175] This argument was suggested to me by Ingmar Persson.

[176] 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.

[177] 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.

[178] 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.

[179] *Metaphysics of Morals*, henceforth *MM* 454. See Allen Wood, *Kant's Ethical Thought*, henceforth *KET* (Cambridge University Press, 1999) 5-8, from whom I take this and the next quotation.

[180] *Lectures on Ethics*, edited by Peter Heath and J.B. Schneewind, henceforth *Lectures* (Cambridge University Press, 1997) 179 (Prussian Edition, 27: 416).

[181] This may be the most important moral question that most rich people face. For three excellent discussions, see Liam Murphy *Moral Demands in Nonideal Theory* (Oxford University Press, 2000), Timothy Mulgan, *The Demands of Consequentialism* (Oxford University Press, 2001) and Garrett Cullity *The Moral Demands of Affluence* (Oxford University Press, 2004).

[182] G 392.

[183] 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.

[184] Kamm gave me this objection in discussion. In her *Intricate Ethics* (OUP, 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 we 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'.

[185] G 423. Kant discusses someone for whom 'things are going well', and who 'contributes nothing' to those who are in need.

[186] MM 443.

[187] 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*.

[188] G 429.

[189] I am not assuming here that, whatever our motives, it cannot be wrong for us to save someone's life. As Marcia Baron suggests, [following Mill?] if some sadist saves someone's life so that he could then kill this person in a more painful way, his act may be wrong as the first part of an intended series of acts that are wrong. But no such claim applies to saving someone's life, in some way that we know will benefit this person, in the hope of getting some reward.

[190] For a further defence of these claims, see pages 000 below.

[191] This is claimed, for example, by Robert Nozick, in *Anarchy, State and Utopia* (Blackwell, 1974) 31, and by Frances Kamm, in *Intricate Ethics*, op.cit. 000.

[192] 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).

[193] See page 00 above.

[194] CR 111 and 114.

[195] CKE 347. Korsgaard may be intending only to describe Kant's view.

[196] CKE 142.

[197] CKE 93.

[198] CR 138.

[199] TJ, 111. and 184

[200] John Rawls *Collected Papers*, edited by Samuel Freeman (Harvard University Press 2001) 355.

[201] Since Rawls makes no use of these proposed senses of 'right' and 'true', my remarks are no objection to his moral theory.

[202] 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).

[203] 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.

[204] Judith Thomson, *The Realm of Rights* (Harvard University Press, 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!'

[205] ibid. 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.

[206] We might claim, however, that it would be wrong for this gangster to save his *own* life in this way.

[207] G 428.

[208] KET 152-5.

[209] KET 117.

[210] MM 462-8.

[211] KET 141, and Shelly Kagan, 'Kantianism for Consequentialists', in Allen Wood's translation of the *Groundwork* (Yale University Press, 2002) 000.

[212] KET 155.

[213] KET 139. (This longest book is *the Metaphysics of Morals*.)

[214] 'The Supreme Principle of Morality', in *The Cambridge Companion to Kant and Modern Philosophy*, edited by Paul Guyer (Cambridge University Press, 2006) 346.

[215] MM 444 and 392.

[216] MM 423-5.

[217] MM 429-30.

[218] KET 154, and 371, note 32.

[219] TJ 31, note 16.

[220] KET 141.

[221] Herman, PMJ 208, 153.

[222] I here follow Scanlon, WWO Chapters 1 and 2.

[223] 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.

[224] *Principia Ethica* 171. (At the end of this paragraph he seems to contradict this claim.)

[225] WWO 99.

[226] 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.)

[227] WWO 104.

[228] WWO 105.

[229] 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.

[230] G 435-6.

[231] Herman writes, 'the domain of the good is rational activity and agency, that is willing' (PMJ 213).

[232] G 396-7.

[233] 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.

[234] 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 (*Second Critique*, 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.

[235] For references, see the notes near the start of Section 32.

[236] *Second Critique* 119.

[237] G 428.

[238] *The Critique of Judgment* 442-3.

[239] PMJ 238.

[240] KET 133.

[241] Herman, PMJ 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' (KET 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.

[242] PMJ 213.

[243] PMJ 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 (DPR 392).

[244] G 435.

[245] Cardinal John Henry Newman, *Certain Difficulties Felt by Anglicans in Catholic Teaching*, (London, 1885) Vol I, 204. [Ross, with less excuse, makes a similar claim.]

[246] DPR 50-57.

[247] G 435.

[248] 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 (CKE 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. Add a reference to Richard Dean's book.

[249] MM 427.

[250] PMJ 215.

[251] PMJ 210.

[252] See, for example, the *Second Critique* 20.

[253] As I have said, there are other kinds of value which are not kinds of goodness, such as economic value. That is irrelevant here.

[254] PMJ 129.

[255] 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'.)

[256] *Second Critique* 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' (*Second Critique* 122), and 'it is our duty to realize the Greatest Good to the utmost of our capacity' (*Second Critique* 143 note).

[257] *Second Critique* 129.

[258] 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).

[259] For example, Stephen Engstrom writes that, on Kant's view, the achievement of such proportionality would be 'the next best thing' ('The Concept of the Highest Good in Kant's Moral Theory', *Philosophy and Phenomenological Researc*h, 1992, 769).

[260] 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' (*Second Critique* 61).

[261] In Moore's words '"right". . . does and can mean nothing but "cause of a good result"(*Principia Ethica* 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.)

[262] 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).

[263] *Second Critique*, 63-4.

[264] G 413. Explain why the word 'ought' is a mistranslation.

[265] G 412.

[266] 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.

[267] *Second Critique* 64.

[268] *Religion* 72.

[269] Lectures 440-1 (27:717). This 'highest end' is the Greatest Good.

[270] *Religion Within the Boundaries of Mere Reason*, 6: 8.

[271] or, more precisely, expectably-best, in the sense defined in Section 17. I shall often ignore this more precise formulation.

[272] Provided, Moore adds, that these rules are both 'generally useful and generally practiced' (G.E.Moore *Principia Ethica* (Cambridge University Press 1903) 211-13). Moore denied that it would be best if there was most happiness; but this point is irrelevant here.

[273] *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.

[274] 'On a supposed right to lie from philanthropy' (8: 425-30).

[275] 8: 426.

[276] *Lectures* 388 (27:651).

[277] *Metaphysik* L1,28:337, From lectures given around 1778, cited in Paul Guyer *Kant on Freedom, Law, and Happiness* (Cambridge University Press, 2000) 94.

[278] 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.

[279] See, for example, the quotations in note 188 above.

[280] As Rawls writes: 'There is nothing in the CI-procedure that can generate precepts requiring us to proportion happiness to virtue' (*Lectures*, 316.)

[281] *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' (*Second Critique* 37).

[282] In Kant's words, 'he must also assume freedom of the will in acting, without which there would be no morals.'

[283] 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.

[284] 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).

[285] 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.'

[286] 'Review of Schultz', 8: 13.

[287] Also in his 'Review of Schultz', 8:13

[288] *The Second Critique*, 5:95.

[289] *Religion Within the Boundaries of Mere Reason*, 6: 44.

[290] *Nicomachean Ethics*, 1114a19;cf.1114b30 *seq*.

[291] See Thomas Nagel's *The View from Nowhere* (Oxford University Press, 1986) Chapter 7. In my statement of this argument, I partly follow Galen Strawson, who gives excellent versions of this argument in his 'The Impossibility of Moral Responsibility' *Phil.Studies* 75, 5-24, and his 'Free Will' in the *Routledge Encyclopaedia of Philosophy*, edited by E.Craig (London, Routledge).

[292] For discussions of the many questions raised by the belief that no one can deserve to suffer, see Sidgwick, *The Methods*, Chapter V, especially section 4, and Derk Pereboom, *Living Without Free Will* (Cambridge University Press) 2001, Chapters 5 to 7.

[293] PMJ vii.

[294] *Second Critique* 27.

[295] *Second Critique*, 19.

[296] G 423.

[297] *The Second Critique* 34.

[298] 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.

[299] MM 453. Kant also refers to the universality of a law that everyone *could* act in certain ways (G 422, my emphasis).

[300] 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.

[301] See, for example, O'Neill, CR, 157. (O'Neill's view has since changed. See, for example, *Towards Justice and Virtue*, 59.)

[302] This is most clearly shown in Kant's discussion of lying promises in G 422.

[303] PMJ 118-119.

[304] PMJ 119.

[305] CKE 136.

[306] *Lectures* 232-3 (29:609).

[307] *Second Critique*, 19.

[308] G 402-3, and 422.

[309] G 422.

[310] Rawls *Lectures* 169.

[311] 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.)

[312] 'On a supposed right to lie from philanthropy' *8*, 425-30.

[313] 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.

[314] CKE 95.

[315] 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.

[316] CR, 133 and 215 and elsewhere.

[317] CR 138-9.

[318] CR 215-6.

[319] CR 102-3.

[320] CKE 92-3.

[321] This is Herman's example (PMJ 138-9).

[322] I take this example from Simon Blackburn *Ruling Passions* (OUP, 1998) 218.

[323] Herman again (PMJ 141).

[324] 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).

[325] PMJ 123.

[326] 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 *Second Critique* 69-70.

[327] 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' (*Second Critique* 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, 000, Wood, KET 80, and Herman, PMJ 120-1.

[328] 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' (*Second Critique* 69). (As before, the emphases are mine.)

[329] 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.

[330] G 403.

[331] Rawls, in his *Lectures*, 166-70, who attributes this point to Herman.

[332] 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.

[333] CR 85.

[334] See Wood's excellent discussion, KET 103-5.

[335] Lectures, 187

[336] MM 455-7.

[337] *The Second Critique* 34.

[338] 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.

[339] 'The Supreme Principle of Morality', in *The Cambridge Companion to Kant and Modern Philosophy*, edited by Paul Guyer (Cambridge University Press, 2006) 345; and 'What is Kantian Ethics?' in *Groundwork for the Metaphysics of Morals*, translated by Allen Wood, (Yale University Press, 2002) 172.

[340] PMJ 104, 132.

[341] Onora O'Neill, *Acting on Principle*, henceforth AOP (Columbia University Press, 1975) 129, 125. See also CR 130

[342] *Human Welfare and Moral Worth*, 122.

[343] PMJ 117.

[344] CR 86, 98, 103. O'Neill is here appealing to Kant's claim that. .

[345] G 403.

[346] G 404, 424.

[347] *Second Critique*, 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.'

[348] O'Neill , Herman, Pogge, and Shelly Kagan all make or discuss proposals of this kind. (CR 87, 130-1; PMJ 147-8; Pogge 'Parfit on What's Wrong', the *Harvard Review of Philosophy*, Spring 2004, 56-58; and Kagan's 'Kantianism for Consequentialists', in Allen Wood's translation of the *Groundwork* (Yale University Press, 2002) 122-127.

[349] ME 202 note. Sidgwick claims that, though this nihilist'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.

[350] There are some exceptions. We might claim, for example, that in driving recklessly, someone caused an accident and thereby killed some some other people.

[351] 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.

[352] AOP 112.

[353] 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' (CR 132).

[354] CR 87.

[355] AOP 112-117, and 124-143, and CR 130. Herman makes similar claims in PMJ Chapters 4 and 10.

[356] Comment on particularism.

[357] It might be claimed that an act's moral worth does depend on the agent's maxim, but that an act can have such worth even if the agent's maxim could not be rationally willed to be universal. That would allow Kant's telling of the truth to have moral worth, even though his maxim could not be rationally willed to be universal. But in our account of moral worth, we don't need to appeal to the agent's maxim. If Kant told the truth (1) because he believed this act to be his duty, (2) at great cost to himself, and (3) this act was indeed his duty, that is enough to give his act moral worth. (It might be enough that (1) is true, since some acts may have moral worth though they cost the agent nothing, or even if the act in question is not, as the agent non-culpably believes, his duty.)

[358] 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.

[359] 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 my *Reasons and Persons*,

(Oxford University Press, 1984), and in my 'Comments' in *Ethics*, Summer 1986.

[360] 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.

[361] 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 she herself never gave such help. This person would then avoid the costs to herself. And whether she received such help would very seldom depend on whether she 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.

[362] 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.

[363] 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 herself or her M-related people, since she would be helping to bring it about that everyone rather than no one did what would be better for herself or for these other people. Since this promise requires unanimity, each person would know that, if she 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.

[364] MM 393.

[365] In a different way, however, this solution may be indirectly collectively self-defeating. See page 000 below.

[366] For a suggestion about when that might be morally permissible, see my *Reasons and Persons*, 100-1. [Refer also to Murphy.]

[367] I take this example from Thomas Pogge, 'The Categorical Imperative', in Kant's *Groundwork of the Metaphysics of Morals: Critical Essays*, edited by Paul Guyer (Rowman and Littlefield, 1998) page 190.

[368] 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.

[369] 'The Categorical Imperative', *op.cit*., 190. Pogge is here following an unpublished lecture given by Scanlon in 1983.

[370] 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." (Manfred Kuehn, *Kant*, (Cambridge University Press, 2001) page 169.)

[371] 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.

[372] CKE 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.

[373] Latest collection, 66.

[374] 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 her part in the best possible pattern of acts.

[375] 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.

[376] This version of RC is open to another objection which I discuss on p 000. But this objection is irrelevant here.

[377] This rule would not in fact be ideal, for reasons that I describe in Section 54, but this point is irrelevant here.

[378] For the best recent statement and defence of Rule Consequentialism, see Bradford Hooker, *Ideal Code, Real World* (Oxford University Press, 2000).

[379] I am partly following some of Kagan's suggestions in his 'Kantianism for Consequentialists', in *Groundwork for the Metaphysics of Morals*, Immanuel Kant, edited and translated by Allen Wood (Yale University Press, 2002), and also Kagan's *Normative Ethics*, 231-5.

[380] 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's paper 'Introducing Variable Rate Rule-Utilitarianism', *The Philosophical Quarterly* (April 2006) 242-253.)

[381] See Hooker's discussion of this question in *Ideal Code, Real World*, *op.cit*.

[382] As Herman notes, PMJ Chapter 7.

[383] 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*.

[384] MM 451. I have changed 'benevolent' to 'beneficent', since that must be what Kant means.

[385] The ancient Near East, India, and China. Add references.

[386] G 430 note.

[387] G. 423 (my italics).

[388] Thomas Nagel, *The Possibility of Altruism* (Oxford University Press, 000) 000, and Equality and Partiality 000-000.

[389] R.M.Hare, *Freedom and Reason*, 000.

[390] TJ, *passim*.

[391] As Leibniz pointed out. See *Leibniz: Political Writings* 2nd edition translated by Patrick Riley (Cambridge University Press, 1988) 56. (I owe this reference to D. D. Raphael *Concepts of Justice* (Oxford University Press: 2001) 84-5.)

[392] MM 450-1.

[393] 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).

[394] CR 94.

[395] TJ, section 30.

[396] G 424.

[397] See Allen Wood's KET op.cit. 3 and 7.

[398] See, for example, G422.

[399] CKE, 101.

[400] Thomas Nagel, *Equality and Partiality* (Oxford University Press, 1991) 42-3.

[401] Kant does write 'every rational being. . . must always take his maxims from the point of view of himself, and likewise every rational being' (G438). 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.

[402] G 423, (my emphases).

[403] 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. . .' *Lectures* 175.

[404] Quote and discuss the passage from the *Second Critique* to which Rawls appeals.

[405] T.C.Williams, *The Concept of the Categorical Imperative*, (OUP, 1968), 123-131.

[406] Thomas Scanlon, WWO 170-1, and in unpublished summaries of lectures.

[407] G 402.

[408] 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.

[409] 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.'

[410] 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?'

[411] WWO 171.

[412] KET 172, PMJ 104 and 132, AOP 125 and 129.

[413] or something similar, such as steadily increasing penalties for failure to agree.

[414] David Gauthier, *Morals by Agreement* (Oxford University Press, 1986) 133.

[415] See, for example, Robert Sugden 'Contractarianism and Norms', *Ethics* 100, 1990.

[416] See Brian Barry,

[417] TJ 134, Revised Edition, henceforth RE, 116.

[418] TJ sections 18-9.

[419] One example is Rawls's appeal to the arbitrariness of the natural lottery. I am here following several writers, especially Thomas Nagel, in his 'Rawls on Justice', *Philosophical Review* April, 1973, reprinted in *Reading Rawls*, ed. Norman Daniels (Blackwell, 1975) , and Brian Barry, in his *Theories of Justice*, Volume 1 (Harvester-Wheatsheaf , 1989), and *Justice as Impartiality* (Oxford University Press, 1995) both passim.

[420] TJ 569, RE 498.

[421] TJ 575, RE 503-4.

[422] *Political Liberalism*, (Columbia University Press, 1996) 49.

[423] TJ 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'. John Rawls, *Selected Papers*, edited by Samuel Freeman (Harvard University Press, 1999) 266.

[424] In appealing to his formula, Rawls writes, 'we have substituted for an ethical judgment a judgment about rational prudence' (TJ 44). When we are behind the veil of ignorance, we are 'assumed to take no interest in one another's interests' (TJ) 147. The people behind the veil of ignorance, he also writes, 'are prompted by their rational assessment of which alternative is most likely to advance their interests' (*Selected Papers*, 312). Rawls does *not* assume that, in the actual world, everyone is self-interested.

[425] TJ 142.

[426] TJ 140.

[427] 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' (TJ 148). Rawls's comparison here is with *impartial* benevolence, and, as he points out, the veil of ignorance makes *partiality* impossible.

[428] TJ 22.

[429] Add references to Brian Barry.

[430] He writes, for example, 'the utilitarian extends to society the principle of choice for one man'(TJ 28).

[431] TJ 165-6, RE, 143-4. Rawls might argue that, on this equal chance assumption, it would be rational to choose a principle that was more cautious than this Utilitarian Average Principle, by giving somewhat greater weight to the well-being of those who were worse off. But such a principle would not differ much from this utilitarian principle.

[432] TJ 168, RE 145.

[433] TJ 122 and 121, RE 105.

[434] John Rawls, *Selected Papers*, edited by Samuel Freeman (Harvard University Press, 1999) 335-6. See also TJ Section 40.

[435] As Rawls claims, TJ 397

[436] Add some remarks about G. A. Cohen's discussion of this question.

[437] *Selected Papers* op.cit 265.

[438] TJ 166, RE, 143.

[439] This objection to Rawls's argument I take from Nagel's 'Rawls on Justice', op.cit. 11.

[440] 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 (TJ 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.' 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.

[441] TJ 584, RE 512.

[442] 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*.

[443] TJ 29, RE 25-6.

[444] 'Distributive Justice: Some Addenda' 1968, republished in John Rawls, *Collected Papers*, edited by Samuel Freeman (Harvard University Press 2001), 174.

[445] 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' (*Justice as Fairness*, Harvard University Press, 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 (TJ 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.

[446] TJ 4, RE 3.

[447] 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.]

[448] 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.

[449] See Scanlon's discussion in WWO 333-342.

[450] WWO 4-5 (and elsewhere).

[451] WWO, 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.

[452] Scanlon appeals to this restriction (though not with this name) on WWO 4-5, 194, and 213-6.

[453] 'Contractualism and Utilitarianism', in *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (OUP, 1997) page 272.

[454] 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 Deontic Beliefs Restriction would exclude such appeals.

[455] WWO 215.

[456] 'Contractualism and Utilitarianism', 267.

[457] 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. . .

[458] This anxiety might not be rational, but that does not undermine these claims.

[459] 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. We might claim that, when considering *Transplant*, White could reasonably reject AU in a simpler way. White could claim that, since we cannot appeal to the burdens that groups of people would together bear, it is morally irrelevant that, if I secretly killed White, I could use her organs to save five people's lives. But such reasoning would also apply to a case like *Lifeboat*, in which I could save either White or five other people. White could claim that it is morally irrelevant that, if I don't save her, I shall be able to save five other people's lives. And Scanlon believes that this fact is not irrelevant, since he would claim that, in *Lifeboat*, we ought to save the five rather than White. Scanlon ought, I believe, to drop the Individualist Restriction, as I argue in my *Response* to Scanlon's *Commentary* below. In some of his more recent writings, Scanlon is less committed to this restriction [add references].

[460] Rawls calls this the *publicity condition*.

[461] 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.

[462] *Selected Papers* 344.

[463] 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' (*Selected Papers*, 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.

[464] *Selected Papers*, 351.

[465] Scanlon, 'Rawls on Justification', in *The Cambridge Companion to Rawls*, edited by Samuel Freeman, (Cambridge University Press, 2003) 149.

[466] I discuss this distinction further near the start of Appendix A.

[467] 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.

[468] As when he writes, 'Besides good and evil, or in other words, pain and pleasure. . . ' 439.

[469] The *Second Critique*, 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.

[470] On one interpretation, the Stoics were making the interesting claim that pain is not bad even in this non-moral sense. See for example, Terence Irwin, 'Kant's Criticisms of Eudaemonism', in *Aristotle, Kant, and the Stoics*, edited by Stephen Engstrom and Jennifer Whiting, (Cambridge University Press, 1996) 80. According to some other writers, the Stoics *were* merely claiming, like Kant, that pain is not morally bad.

[471] Sir David Ross, *Foundations of Ethics* (Oxford University Press, reprinted 2000) 272-284. (Though Ross makes these claims about pleasure, he intends them to apply to pain.)

[472] The *View from Nowhere*, 161.

[473] 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' (Judith Thomson *Goodness and Advice* (Princeton University Press, 2003) 19. In making this last claim, Thomson assumes too quickly that her question can't be answered.

[474] 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'.

[475] In the sense explained in Section 17 above.

[476] 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 test 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 test 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 Shelly Kagan's 'Evaluative Focal Points', in *Morality, Rules and Consequences*, edited by Brad Hooker, Elinor Mason, and Dale E. Miller (Edinburgh University Press, 2000) and in Kagan's *Normative Ethics* (Westview Press, 1998) Chapters 6 and 7.

[477] See, for example, Rawls, *Lectures*, 173-6 and 232-4.

[478] If these people themselves accept a 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-involving sense. But they might want things to go in the ways that would in fact be best in this sense.

[479] 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.

[480] For a partial defence of such a principle, see Frances Kamm's contribution to *Singer and his Critics*, ed D. Jamieson (Blackwell, 2000), or her 'The new problem of distance in morality', in *The Ethics of Assistance*, edited by Deen K. Chatterjee (Cambridge University Press, 2004).

[481] 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.

[482] 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 Brian Barry *Justice as Impartiality* (Oxford University Press, 1995) Chapters 1, 8, and 9.

[483] 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.

[484] Bernard Williams, *Moral Luck* (Cambridge University Press, 1981) 18.

[485] 125.

[486] *The Collected Works of W.B. Yeats, Vol III,* edited by Douglas Archibald and William O'Donnell, 246.

[487] This may not be the best description of what makes these acts permissible. For another account, see Frances Kamm *Intricate Ethics,* op.cit.

[488] A fuller version of this argument would need to consider other kinds of case, since we cannot assume that some fact always give us the same reason. But my claims would, I believe, apply to all other relevant cases.

[489] In the case of certain principles, there might be no such people. I discuss this possibility in my *Response* to Scanlon's *Commentary* below, p.000.

[490] ME, Book IV, Chapters III to V.

[491] Kagan suggests a similar argument in his 'Kantianism for Consequentialists', in *Groundwork for the Metaphysics of Morals*, Immanuel Kant, edited and translated by Allen Wood (Yale University Press, 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 his paper 'Could Kant Have Been a Utilitarian?', in R.M.Hare *Sorting Out Ethics* (Oxford University Press, 1997). These arguments differ in several ways from the arguments that we have been discussing. For another, even more different argument, see David Cummiskey, *Kantian Consequentialism*

(Oxford University Press, 1996). Kant's texts are inexhaustibly fertile, provoking in different people very different thoughts.

[492] 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.

[493] 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.

[494] That is mainly because, in asking which are the principles whose being universally followed wo best, we can ignore the various ways in which, when people try to make things go best, they ofte miscalculation, self-deception, and the like. We can also note that, on some versions of Rule Con appeal to the principles that are optimific in some community, or during some period. Kantian C take this form. If we ask which principles were or would be UF-optimific in the 20th and 21st Ce principles would be even closer to AC than they would be in most other centuries. Given extreme wealth and power, and great advances in technology, many Act Consequentialists would be able, do far more good than ever before. So ßAC might now be UF-Optimific. But AC was not opti centuries, and we could hope that things will change, so that AC would cease to be optimific in fut

[495] I discuss some of these questions in Sections 37 to 43 of my *Reasons and Persons*. And see again Kagan's 'Evaluative Focal Points', in *Morality, Rules and Consequences*, op.cit. and Kagan's *Normative Ethics* (Westview Press, 1998) Chapters 6 and 7.

[496] In Section 34.

[497] 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.

[498] *The First Critique*, A 851 B 879.

[499] *Metaphysik* L1,28:337, cited in Paul Guyer *Kant on Freedom, Law, and Happiness* (Cambridge University Press, 2000) 94.

[500] 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).

[501] WWO, 11.

[502] 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.

[503] Though Kantian Rule Consequentialism has different versions, which may conflict, these conflicts are not between the Kantian and Rule Consequentialist parts of this view.

[504] 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 beqk 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.

Notes to the Appendices

[505] 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.

[506] I have added the reference to harming people, which I assume that Wolf intends.

[507] G 432.

[508] ME, viii.

[509] Add some remarks about Thomson?

[510] Allen Wood *Kantian Ethics*, henceforth *KE* (Cambridge University Press, 2008), 00, See also his discussion in KET 000.

[511] 'The supreme principle of morality' in ed Paul Guyer *Kant and Modern Philosophy* (Cambridge University Press, 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.

[512] KE 00.

[513] KE, preface.

[514] G 431.

[515] KE 000.

[516] KE 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's KET 187-190.

[517] 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' (KE, 00).

[518] 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.

[519] 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.

[520] Reference.

[521] Commentary 000.

[522] KET 121.

[523] PMJ 210, 212.

[524] Just before this definition, Kant refers to 'the dignity of a rational being' (G 434).

[525] PMJ 238.

[526] KET 130.

[527] Kerstein 182.

[528] CKE 125.

[529] 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.' [529]

[530] Commentary 000.

[531] KE 000.

[532] KE 000.

[533] Reference to Korsgaard.

[534] KET 127.

[535] KET 129

[536] Commentary 000.

[537] G 435.

[538] Commentary 000.

[539] 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.' 86

[540] G 406.

[541] KET 120.

[542] 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 Schonecker)

> 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

[543] *Second Critique* 125 and 129.

[544] Commentary 000.

[545] KE 000. See also KET 000.

[546] KE 000.

[547] KE 000.

[548] PMJ 213.

[549] PMJ 214

[550] Commentary 000.

[551] PMJ 124.

[552] PMJ 124.

[553] PMJ 129. This sentence continues '(self) by another'. But I am not my agency.

[554] KET 116-7.

[555] KET 144.

[556] Second Critique 78.

[557] 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' (PMJ 147). See also PMJ 112 and 146.

[558] PMJ 104, 132.

[559] 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.'

[560] Reference.

[561] MM 219

[562] 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.

[563] G 403.

[564] G 404, 424.

[565] 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' (PMJ 89).

[566]566 PMJ 34.

[567] G 402

[568] 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.

[569] PMJ 95.

[570] PMJ 94

[571] PMJ 99.

[572] PMJ 118

[573] PMJ 120 (my italics).

[574] 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.' PMJ 194-5.

[575] G 423.

[576] PMJ 49.

[577] Nor, we can add, would it be enough to appeal to what people *prefer.*

[578] PMJ 52.

[579] PMJ 54 note 12.

[580] Herman suggests another argument for the view that, when we apply Kant's Formula, we should not appeal to prudential reasons. If that is how some agent applies Kant's Formula, she writes, that 'must lead him to think that the satisfaction of his desires is significant in deriving duties' (PMJ 50). This argument seems to me to have little force, since this bad effect would be produced only by a misunderstanding of Kant's Formula.

[581] 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.

[582] 'Contractualism and Utilitarianism', in *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (OUP, 1997) page 272.

[583] 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.

[584] 229.

[585] 212, and elsewhere. [Scanlon's claims about fairness do in a less direct way appeal to claims about well-being.]

[586] 'Contractualism and Utilitarianism', 267. He also says that he is one of those 'who look to contractualism specifically as a way of avoiding utilitarianism', 215.

[587] 235.

[588] 241.

[589] 230.

[590] 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.

[591] 240.

[592] 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. [592] 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.

[593] Scanlon writes: 'where the base line is equal, benefiting only Blue seems objectionable, because all have the same claim to some benefit' (ed Stratton-Lake, op.cit. 131).

[594] 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.

[595] 229.

[596] 'Contractualism and Utilitarianism', op. cit. 123.

[597] 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' (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.

[598] 'Thomson on Self-Defence', in *Fact and Value*, edited by Alex Byrne, Robert Stalnaker, and Ralph Wedgwood, (The MIT Press, 2001) 200.

[599] 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.

[600] Refer to Nagel, *Equality and Partiality*,

[601] references.

[602] 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' (ed Stratton Lake 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.'

[603] 222

[604] 182

[605] 219

[606] Comment on Scanlon's discussion of this claim.

[607] [Refer to my earlier more accurate statement of this view].

[608] Stratton Lake 133.

[609] TJ 25.

[610] Refer to Nagel's claims in *Equality and Partiality*.

[611] Note that Scanlon surprisingly rejects these claims?

[612] It might be claimed that, in *Case Eleven*, 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.

[613] Another example is provided by Liam Murphy's claims about the demandingness of morality. MORE.

[614] 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.

[615] There are also claims which use normative concepts, but are not in my 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.

[616] Christine Korsgaard, *The Sources of Normativity*, henceforth *Sources*, (Cambridge University Press, 1996) 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 to mean 'would motivate them'.

[617] 'Mill and Experiments in Living', *Ethics* October 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'.

[618] Though Moore himself did not distinguish between concepts and the properties to which they refer.

[619] Stephen Darwall 'Internalism and Agency' *Philosophical Perspectives, Vol. 6. Ethics* (1992) 168. Darwall's sentence continues 'perhaps when the agent's deliberative thinking is maximally improved by natural knowledge.'

[620] 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', henceforth *IROB*, in *Making Sense of Humanity* (Cambridge University Press, 1995), 40. (I have substituted 'do X' for 'phi'.) See also 'Internal and External Reasons', henceforth *IER*, reprinted in his *Moral Luck* (Cambridge University Press, 1982). These articles contain many similar remarks. In the next passage quoted in my text, 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. . . . '.

[621] 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. (Stephen Darwall, *Impartial Reason* (Cornell University Press, 1983) 128 (with 'A' replaced by 'X').

[622] 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' (*Moral Luck*, op.cit. 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' (*Moral Luck* 124). Falk discusses these senses of 'ought' and

'should' in many of the articles reprinted in W. D. Falk *Ought, Reasons, and Morality* henceforth *ORM* (Cornell University Press, 1986).

[623] 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' (IROB 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 can here ignore Williams's requirement 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.

[624] 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.

[625] IROB 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 worthwhile life. That further assumption would not alter Williams's view about this example.

[626] Some of Williams's other arguments I discuss near the end of Chapter 3, and in my 'Reasons and Motivation', *Proceedings of the Aristotelian Society, Supplementary Volume*, 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 claim that we use.

[627] 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. . . ' IROB 39-40. And Williams writes elsewhere that externalists do not 'offer any *content* for external reasons statements' (*World, Mind, and Ethics*, henceforth *WME*, edited by J.E.J.Altham and Ross Harrison (Cambridge, 1995) 191 (my italics).

Discussing another, similar example, Williams writes: 'What is gained, except perhaps rhetorically, by claiming that A has a reason to do a certain thing, when all one has left to say is that this is what. . . a decent person. . . would do?' (WME 215). This question assumes that, if our claim about A does not have the sense described by Analytical Internalists, there is nothing distinctive left for it to mean. We couldn't mean that, despite A's motivational state, A has a reason to do this thing. If we could mean that, there would be a simple answer to Williams's question. We might be saying something that was both distinctive and true.

[628] More needs to be said, for example, about those logical or mathematical claims whose truth in some sense follows from conceptual truths, though they give us substantive information.

[629] *Impartial Reason* 210-11.

[630] *Impartial Reason* 128.

[631] Bernard Williams, *Philosophy as Humanistic Discipline* (Princeton University Press, 2006) 110.

[632] IROB 36.

[633] Though Williams sometimes calls claims about reasons *normative*, he also writes that 'no account of 'A has reason to do X' can be adequate unless it has normative force. . .' (WME 191), and that 'on the internalist view, a statement of the form 'A has reason to do X' has normative force'(IROB 36). Delete?

[634] 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.

[635] I follow Scanlon WWO 20.

[636] Falk ORM 35, 184.

[637] *Impartial Reason* 134

[638] *Impartial Reason* 128.

[639] *Impartial Reason* 86. As we shall see, however, Darwall's final version of Analytical Internalism is not a form of Analytical Naturalism.

[640] Williams similarly describes a case in which 'ethical or prudential reasons. . . are not, *as it turns out*, the strongest reason for me, now; the strongest reason is tha t I desire very much to do something else' *Ethics and the Limits of Philosophy* (Fontana, 195) 19 (my italics).

[641] 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. . .

[642] W. D. Falk 'Morality and Nature', *The Australasian Journal of Philosophy*, 1950, 80.

[643] He may also see that the claim that you have these reasons does not give you a further reason.

[644] ORM 48 , 62-3.

[645] ORM 65.

[646] ORM 66.

[647] 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' (ORM 93).

[648] ORM 34.

[649] *Impartial Reason* 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' (*Impartial Reason* 57).

[650] Michael Smith, *The Moral Problem* (Blackwell, 1994) 57.

[651] Richard Boyd 'How to Be a Moral Realist?', in *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard, and Peter Railton (Oxford University Press, 1997) 119.

[652] 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.

[653] Remark about how this applies to particularism.

[654] For one version of this argument, see Frank Jackson *From Metaphysics to Ethics* (Oxford University Press, 1998) 122-129.

[655] 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*.

[656] 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, as its restatability as (C) shows, a normative claim.

[657] 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.

[658] For two such arguments, see Smith 1994 and Boyd 1997.

[659] I earlier wrote here: 'You claim that, since identity is a symmetrical relation, your view gives the same status to rightness and to maximizing happiness. But that is not so. On your view, if some act maximizes happiness, that's *what it is* for this act to be right. You could not state this view the other way round. You could not claim that, if some act is right, that's *what it is* for this act to maximize happiness. That claim would be absurd.' But the same could be said about heat and molecular kinetic energy. MORE.

[660] I take this analogy from Mark Schroeder, *Slaves of the Passions* (Oxford University Press, 2007) 75-8. Say why the analogy is only partial.

[661] Nicholas Sturgeon, 'Moral Explanations'. . . 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 criterion 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.

[662] 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.

[663] Sturgeon makes some relevant remarks, which I shall discuss in the unwritten Section 8.

[664] In a note only? 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'.

[665] Refer to Sturgeon, and to Cuneo's remarks about virtues.

[666] (As I argue elsewhere, this kind of explanation would not be wholly different from our most fundamental naturalistic explanations. See my 'Why Anything? Why This?'. . .)

[667] Some Wide Naturalists might accept only the second of these claims. But that is enough to support my conclusion.

[668] Gibbard 'Normative Properties', in *Metaethics after Moore*, edited by Terry Horgan and Mark Timmons (Oxford University Press, 2006) 323.

[669] As I have said, that is shown by the fact that both (L) and (F) are merely other ways of claiming that acts are right just when, and because, they maximize happiness.

[670] 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.

[671] I am assuming that, in (J), the phrase 'the writer of *Hamlet'* is a *rigid designator*, which refers to the person who in fact wrote *Hamlet*, and which would refer to this same person even in claims about other possible worlds, such as the claim that the writer of *Hamlet* might have died before he wrote *Hamlet*.

[672] Schroeder 59. Schroeder adds many qualifications to this claim, but these are irrelevant here.

[673] Schroeder 95-6.

[674] Schroeder 60.

[675] 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.

[676] 'Moral Explanations', in .

[677] Frank Jackson *From Metaphysics to Ethics* (Oxford University Press, 1998) 124-5. Jackson also writes: 'all there is to tell about moral nature can

be told in naturalistic terms' ('Critical Notice of Hurley' *Australasian Journal of Philosophy* vol. 70 1992, Section 4).

[678] Peter Railton, *Facts, Values, and Norms* (Cambridge University Press, 2003) xvii-xviii.

[679] Discuss the relation between Hard Naturalism and Analytical Naturalism.

[680] 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.

[681] I take this example from Gibbard 1990.

[682] *Morality, Utilitarianism, and Rights*, *op.cit*., 35-6

[683] *Morality, Utilitarianism, and Rights*, 29.

[684] Jackson 1998, 127.

[685] Jackson 1998 142.

[686] 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.)

[687] 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.'

[688] 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).

[689] Thomas Nagel, *The Possibility of Altruism* (Oxford University Press, 1970).

[690] For a fuller, partly similar response, see David Copp '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, 2001).

[691] 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 R.M.Hare *Moral Thinking* (Oxford University Press, 1981) and 'Could Kant Have Been a Utilitarian?' in *Sorting Ethics Out* (Oxford University Press, 1997).

[692] 194. Moore's main contribution, Gibbard also writes, was to ask 'What. . is at issue in moral disputes? What does the disagreement consist in?'

[693] Simon Blackburn *Ruling Passions*, henceforth *RP* (Oxford University Press, 1998) 49, 275, 90.

[694] RP 69.

[695] *Thinking How to Live*, henceforth *THTL* (Harvard University Press, 2003) 74.

[696] *THTL* 65.

[697] THTL 184.

[698] THTL ix-x.

[699] THTL x.

[700] THTL 10.

[701] From the *Phil Phen* Symposium.

[702] THTL 9.

[703] THTL 270.

[704] THTL 273, 271.

[705] THTL 54.

[706] Gibbard might reply that, when we change some plan because our preferences change, we shall know that our preferences have changed, and that will involve a change in our factual beliefs. Such cases are irrelevant, Gibbard might say, since his claim about normative disagreement applies only to those changes of plan that involve no change in our factual beliefs. But Gibbard could not then answer the earlier objection by restricting his claim to cases in which our preferences change.

[707] THTL 268-74.

[708] Gibbard 2002 in ed Cuneo Shafer-Landau 77. ??

[709] 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.

[710] THTL 17, x.

[711] Simon Blackburn *Spreading the Word* (Oxford University Press, 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).

[712] *Wise Choices, Apt Feelings*, henceforth *WCAF* (Oxford University Press 1990) 287.

[713] RP 309.

[714] RP 118 note 36.

[715] RP 318 (my italics).

[716] THTL 65.

[717] *Essays in Quasi-Realism*, henceforth *EQR* (Oxford University Press, 1993) 20.

[718] RP 318.

[719] RP 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' (EQR, 20).

[720] I follow Russ Shafer-Landau *Moral Realism* (Oxford University Press, 2003) 28-9.

[721] RP 313.

[722] Andrew Egan 'Quasi-Realism and Fundamental Moral Error', *Australasian Journal of Philosophy*, 85:2, 205-19.

[723] RP 318.

[724] 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)' RP 319).

[725] As Blackburn writes: 'The projectivist can say this vital thing: that it is not because of our responses. . that cruelty is wrong' (EQR 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' (EQR 177). But he could withdraw this claim.

[726] EQR 129.

[727] *Inquiry* 1999?

[728] EQR 173.

[729] RP 50.

[730] 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 nor true false.' Mark Timmons *Morality Without Foundations* (Oxford University Press, 1998)

[731] From the reply to Egan. Ask Simon for permission to quote from this.

[732] Simon Blackburn *Spreading the Word* (Oxford University Press, 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).

[733] EQR 4, 20.

[734] RP 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.

[735] THTL 98.

[736] THTL 15.

[737] THTL 16. Blackburn makes similar claims. See, for example, RP 87.

[738] EQR 157.

[739] See, for example, EQR 16, 157 note 9, and RP 305.

[740] WCAF 154.

[741] WCAF 154.

[742] *From Metaphysics to Ethics* (Oxford University Press, 1998) 128.

[743] EQR 163.

[744] As before, see Nagel's *The Last Word*.

[745] THTL 5.

[746] THTL xii.

[747] WCAF 8.

[748] WCAF 70, 46.

[749] Gibbard 2003, 9-10. More exactly, Gibbard says that 'ought' here adds nothing.

[750] WCAF 68-76.

[751] WCAF vii.

[752] WCAF 7.

[753] WCAF 9.

[754] WCAF 153.

[755] WCAF 172

[756] WCAF 173.

[757] WCAF 175.

[758] WCAF 177.

[759] WCAF 33.

[760] WCAF 8.

[761] WCAF 154-5.

[762] THTL 17, x.

[763] 'Nothing Matters', in R.M. Hare *Applications of Moral Philosophy* (Macmillan, 1972), 33-4.

[764] 'Nothing Matters', *op.cit.*, 40.

[765] R.M. Hare *The Language of Morals*, (Oxford University Press, 1952) 195.

[766] Patrick Nowell-Smith, *Ethics*, (Penguin, 1954) 319-20.

[767] Nowell-Smith *Ethics*, *op. cit.*, 61.

[768] Bernard Williams, 'Ought and Obligation', *Moral Luck*, *op.cit.* 122. (I have expanded some abbreviations.)

[769] R.M. Hare, *Moral Thinking*, (Oxford University Press, 1981) 217.

[770] 'Realism and Constructivism', *Philosophy in America at the Turn of the Century* APA Centennial Supplement, *Journal of Philosophical Research* (2003) 112.

[771] 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.

[772] *Sources*, 38.

[773] *Sources*, 44.

[774] *Sources* 41 note 68.

[775] For a longer discussion of Korsgaard's view, see my 'Normativity', in *Oxford Studies in Metaethics Vol 1*, edited by Russ Shafer-Landau (Oxford, 2006). Even there I say little about some of the most original and central features of Korsgaard's view, such as her claims about our practical identity.

[776] *NIR*, *op.cit*. 240.

[777] *NIR* 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?' ('Realism and Constructivism', op.cit. 110, my italics).

[778] *CKE*, 163-4.

[779] As Nagel writes, 'Only a justification can bring the request for justifications to an end'. Thomas Nagel, *The Last Word* (Oxford University Press, 1997) 1-6.

[780] Credit for such cases may be due to Gregory Kavka 'The Toxin Puzzle', *Analysis*, 43 (1986).

[781] For a similar appeal to the difference between such questions, see Pamela Hieronymi, 'The Wrong Kind of Reason', *The Journal of Philosophy*, 102 no 9 (September 2005). See also Hieronymi's 'Controlling Attitudes' , Pacific Philosophical Quarterly, 87, no 1 (March 2006). Hieronymi does not, however, conclude that there are no state-given reasons.

[782] 'Reason and Maximization' , *Canadian Journal of Philosophy* 4: 1975. This argument's fullest statement is in Gauthier's *Morals by Agreement* (Oxford University Press, 1986), henceforth *MA*.

[783] In an unpublished paper 'Rational Irrationality', and later in Sections 7-8 of my book *Reasons and Persons* (Oxford University Press, 1984).

[784] This appendix was written in 1994, in response to Gauthier's contribution to *Reading Parfit*, edited by Jonathan Dancy, (London: Routledge, 1997). I have not tried to take into account Gauthier's most recent work.

[785] 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 (*MA*, p. 6), most of his claims apply equally to Rational Egoism. And Gauthier often uses words, like 'benefit' and 'advantage', that refer more naturally to our interests rather than our present preferences. So we can here ignore the differences---though they are often great---between the Deliberative 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.

[786] 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.

[787] *Reasons and Persons*, Sections 7-8.

[788] Gauthier gave this reply in MA (especially, 173-4). In his contribution to *Reading Parfit*, 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.

[789] See, for example, MA, Chapter VI.

[790] In *Reasons and Persons*, Sections 7-8.

[791] 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.

[792] As he wrote (like Queen Victoria), 'We are unmoved' (MA, p. 185).

[793] Gauthier asserted (B)---which he calls his 'second level of commitment'---in *Reading Parfit*, p. 40. I discussed a similar claim, which I called '(G1)', in *Reasons and Persons* (p. 13). On Gauthier's second level of commitment, it is rational to act on a disposition 'so long as one reasonably expects past and prospective adherence to the disposition to be maximally beneficial'. This claim may seem to mean 'if one both reasonably believes that adherence to this disposition in the past has been beneficial, and reasonably expects that adherence to it in the future will be beneficial'. But this cannot be what Gauthier intends, since it would remove the difference between his second level of commitment and his first level (discussed below). Gauthier must mean: 'if one can reasonably believe that acquiring it was beneficial in one's life as a whole, taking the past and future together.'

Gauthier's move from (A) to (B), or from his third to his second level of commitment, hardly damages his defence of rational morality. On the view defended in MA, for morality's constraints to have rational force for us, accepting these constraints must have been expectably-best for us. On Gauthier's revised view, for these constraints to have rational force, they must also be known not to have been on the whole bad for us. Most of contractual morality's constraints would meet this second requirement.

[794] Perhaps I would have obeyed some order that would have proved fatal.

[795] 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' (*Reading Parfit*, p. 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.)

[796] 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' ('Deterrence, Maximization, and Rationality', *Ethics*, 94:1984 p, 489.) So

what? Her expectation may have depended on that willingness. But why should she remain faithful now?

[797] 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.

[798] 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.

[799] (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' (*Reading Parfit*, p. 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'.)

[800] My drug-induced insanity, Gauthier claims, is 'the rational disposition in such situations, and the actions to which it gives rise are rational actions' (*Reading Parfit*, p. 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.

[801] MA, (*passim).*

[802] 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.

[803] 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' (*Reading Parfit*, p. 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' (*Reading Parfit*, pp. 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.

[804] 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.' (MA,p. 182.)

[805] See pages 000.

[806] 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.

[807] 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.

[808] *Reading Parfit*, (p.37).

[809] 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.

[810] 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.

[811] 'Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices' (MA, p. 187).

[812] 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.

[813] 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.

[814] MA (p. 170).

[815] 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.

[816] 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.

[817] 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.

[818] 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.

[819] In 'Deterrence, Maximization, and Rationality', and in *The Security Gamble*, ed. Douglas MacLean (Totowa, NJ: Rowman & Allanheld, 1984).

[820] 'Afterthoughts', in *The Security Gamble* (pp.159-61).

[821] Cf. Edward McClennen, 'Constrained Maximization and Resolute Choice', *Social Philosophy and Public Policy*, *5: 1988.*

[822] 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.

[823] 'Constrained Maximization'.

[824] In a letter to me.

[825] See MA (p. 182) and *Reading Parfit*, (p. 31). (But see also MA, pp. 170 and 158.)

[826] *Reading Parfit*, (p. 36).

[827] 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)' (*Reading Parfit*, p. 40).

[828] 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.

[829] In *The Security Gamble*.

[830] 'Afterthoughts', in *The Security Gamble* pp. 159-61.

[831] *Reading Parfit*, p. 30.

[832] *Reading Parfit*,p. 36.

[833] *Reading Parfit*, p. 38.

[834] MA, 17

[835] G 414.

[836] G 427-8.

[837] 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.

[838] G 427-8.

[839] G 428.

[840] 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.)

[841] G 402.

[842] 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.

[843] G 414.

[844] G 420-1.

[845] Reference to Allison and others.

[846] G 399-402.

[847] G 398.

[848] G 399-400.

[849] G 400-402.

[850] *The Cambridge Companion to Kant*, edited by Paul Guyer, (Cambridge University Press, 1992) 325-6.

[851] Nelson Potter, 'The Argument of Kant's Groundwork', in *Kant's Groundwork of the Metaphysics of Ethics, Critical Essays*, edited by Paul Guyer, (Rowman and Littlefield, 1998) page 40.

[852] *Kant's Groundwork. . .Critical Essays,* ed.Guyer, *op. cit.* 318.

[853] Mary Gregor, *Laws of Freedom*, Oxford 1963, 78-9.

[854] CPR 29.

[855] CPR 27.

[856] 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.

[857] CPR, 33.

[858] G 443.

[859] G 444.

[860] 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.

[861] For an excellent discussion of both these arguments, see Samuel Kerstein *Kant's Search for the Supreme Principle of Morality* (Cambridge University Press, 2002) Chapter 7. There is much else in Kerstein's book which goes beyond, and may partly correct, my brief claims in this Appendix.

[862] CPR 39-41.

[863] CPR 59-60.

[864] CPR 59-60.

[865] CPR 60.

[866] Terence Irwin, 'Kant's Criticisms of Eudaemonism', in *Aristotle, Kant, and the Stoics*, edited by Stephen Engstrom and Jennifer Whiting, (Cambridge University Press, 1996) 80.

[867] CPR 58.

[868] CPR 58.

[869] CPR 60.

[870] CPR 61.

[871] CPR 61.

[872] CPR 62.

[873] G 432-3.

[874] G 441.

[875] G 408.

[876] G 457.

[877] CPR 82.

[878] G 448.

[879] *Metaphysics of Morals*, 418.

[880] See, for example, 'What is Orientation in Thinking', VIII, 145, 303-4, and G 448.

[881] *Metaphysics of Morals*, 438 and note.

[882] CPR 92.

[883] Thomas E. Hill *Dignity and Practical Reason* (Cornell University Press, 1992) 88.

[884] *CKE*, 22.

[885] G 444.

[886] G 449.

[887] G 449.

[888] In 'The Normativity of Instrumental Reason', in *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Gaut, (Oxford University Press, 1997)

[889] See, for example, CPR, 31.

[890] G 444.

[891] G 441.

[892] *G* 414.

[893] G 415.

[894] G 415.

[895] G 417.

[896] *Critique of Practical Reason*

[897] Kant writes, 'from the problematic and pragmatic', which are his names for the two forms of hypothetical imperative.

[898] *Lectures in Ethics*, henceforth *Lectures*, translated by Peter Heath, (Cambridge University Press, 1997) p. 256, Academy edition, 486.

[899] Lectures, 263 (Academy 493); see also 60 and 332 (Academy 268 and 582).

[900] 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.

[901] G 410-11.

[902] G 420.

[903] G 420.

[904] G 408.

[905] G 408.

[906] G 426-7.

[907] 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.'(Second Critique 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.

[908] G 417.

[909] G 402.

[910] CPR 69. 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?

[911] Reference to Darwall.

[912] CPR 62-3.

[913] CPR 64.

[914] Theory of Justice, (Harvard University Press, 1971) 30-3.

[915] CPR 63.

[916] CPR 63.

[917] CPR 63.

[918] CPR 63.

[919] CPR 58.

[920] 'Themes in Kant's Moral Philosophy', in E. Foerster (ed.) Kant's Transcendental Deductions (Stanford University Press, 1989), 109.

[921] CPR 63.

[922] CPR 62.

[923] CPR 22.

[924] MM. 378.

[925] G 444.

[926] CPR 41.

[927] CPR 62.

[928] CPR 93.

[929] CPR 25.

[930] CPR 34.

[931] G 460.

[932] CPR 64.

[933] CPR 24 (my italics).

[934] CPR 44.

[935] CPR 73.

[936]CPR 73.

[937] CPR 80.

[938] Second Critique 27.

[939] CPR 31.

[940] G 460, note.

[941] CPR 29.

[942] First Critique, A/319/B375.

[943] G 408.

[944] *Religion within the Limits of Reason Alone*, translated by T. Greene and H. Hudson, Harper 1960, 25.

[945] *Lectures*, 265 (Academy 497).

[946] *Lectures*, 229-30 (Academy 605).

[947] G 425.

[948] CPR, 31.

[949] *Lectures*, 257 (Academy 487).

[950] *Lectures*, 230 (Academy 606).

[951] *Lectures*, 256 (Academy 486).

[952] G 408.

[953] G 413.

[954] CPR 65.

[955] *Lectures*, 65 (Academy 274).

[956] *Second Critique* 24.