CLIMBING THE MOUNTAIN

SUMMARY

CHAPTER 1 REASONS

1 Practical 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. Events can be good or bad for particular people, or *impersonally* good or bad, in reason-involving senses.

2 Reason-Giving Facts

According to *desire-based* or *aim-based* theories, reasons for acting are all provided by facts about how we could fulfil or achieve our present desires or aims. There are, I shall argue, no such reasons. As *value-based* theories claim, reasons for acting are all provided by facts about what is relevantly good, or worth achieving. These facts give us reasons both to want to achieve certain aims, and to try to achieve them. Since it is only these facts that give us reasons, why do so many people accept desire-based or aim-based theories? There are several possible explanations.

3 Value-based Theories

We can respond to our reasons for acting in a direct and voluntary way. We can also respond directly to our reasons for having some belief or desire; but, in most cases, these responses are not voluntary.

Our reasons to have some desire are provided by facts about this desire's *object*, or what we want. It is often claimed that we have reasons to have some desire when and because our *having* this desire would be, in some way, good. There are, I suggest, no such reasons. Nor do we have such *pragmatic* reasons to have particular beliefs.

What we want is always some event, in the wide sense that covers acts and states of affairs. When we know the relevant facts, we ought rationally to want those events that we have most reason to want. If we want some event whose features give us strongly decisive reasons *not* to want it, our desire is contrary to reason, and irrational. It would be irrational, for example, to prefer to have one hour of agony tomorrow rather than five minutes of slight pain later today.

4 Desire-based Theories

Desire-based theories cannot make such claims. According to these theories, we can have reasons to want something as a *means* to something else that we want. But we cannot have reasons to want anything as an *end*, or for its own sake. We cannot have such reasons to want to avoid agony, or to be happy, or to have any other aim. On these theories, *nothing matters.* We should reject the arguments for this bleak view.

CHAPTER 2 RATIONALITY

5 Rational Desires

Our desires are rational, many people claim, just when they causally depend on rational beliefs. That is not true. Most of our desires are rational when they depend on beliefs whose truth would give us reasons to have these desires. It is irrelevant whether these beliefs are rational. Nor does the rationality of our desires normatively depend, as many people claim, on how we came to have these desires, or on whether these desires are inconsistent, or on whether our having these desires has good effects. Special claims apply to the relations between our desires and some of our normative beliefs.

6 Sidgwick's Dualism

When we are trying to decide what we have most reason to do, we can rationally ask this question, Sidgwick assumes, either from our actual personal point of view, or from an imagined impartial point of view.

From our personal point of view, Sidgwick claims, we have most reason to do whatever would be best for ourselves. From an impartial point of view, we have most reason to do whatever would be impartially best. 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 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, I shall say, 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 roughly. As *wide value-based* 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 knew the facts that gave us such reasons, either act would be rational.

CHAPTER 3 MORALITY

7 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 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, before we can decide whether and when we might have either sufficient or decisive reasons to act wrongly, we must know more about which acts are wrong, and what makes them wrong.

8 Moral Concepts

The words 'ought morally' and 'wrong' can be used in several senses.

By distinguishing and using these senses, we can avoid some unnecessary disagreements.

It is a difficult question whether, as I believe, there are some irreducibly normative truths, some of which are moral truths. This *meta-ethical* question will be easier to answer when we have made more progress in answering questions about what we have reasons to want and do, and about what we ought morally to do. Rather than proposing a new moral theory, this book tries to develop and combine existing theories of three kinds: Kantian, Contractualist and Consequentialist.

CHAPTER 4 POSSIBLE CONSENT

9 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 it is always wrong to coerce or deceive people, since these may seem to be acts whose nature makes consent impossible. But that is not relevantly true.

10 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 they *cannot actually* give or refuse consent, because we have not given them 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 they knew the relevant facts, and we gave them the power to choose how we treat them. This principle might be true, and is more likely to be what Kant means.

Kant's claims about consent give 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, if they knew the facts, they could rationally consent. And this might be how everyone ought always to act.

11 Reasons to Give Consent

Whether we could achieve Kant's ideal depends on which are the acts to which people could rationally consent. If we ought to accept either

some desire-based 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 theory, as I believe, the Consent Principle may succeed. As some examples suggest, there may always be at least one possible act to which, if they knew the facts, 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.

12 A Superfluous Principle?

According to some writers, even if the Consent Principle is true, this principle would add 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. It is also worth asking whether we could achieve Kant's ideal.

13 Actual Consent

It is wrong to treat people in certain ways if they 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 extreme view, it is wrong to treat people in any way to which they refuse consent. That is clearly false. It may be objected 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.

14 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. That is not, as it may seem, an objection to this principle. The Consent Principle, we can argue, could never require us to act wrongly.

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

CHAPTER 5 MERELY AS A MEANS

It is wrong, Kant claims, to treat any rational being merely as a means. We treat someone in this way when we both use this person and regard her as a mere tool, 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 (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 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 condemn such acts. On this doubly revised claim, it is wrong 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. Our only way to save these people's lives is to cause someone else, without her consent, to fall onto the track, thereby killing this person but stopping the train. It may seem that, if we acted in this way, we 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, it would not be condemned by either the Consent Principle or the Mere Means Principle.

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. When we discuss moral questions, we should try not to use ordinary words or phrases in special senses.

18 Harming as a Means

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, when it would *not* be wrong to impose certain lesser harms on people as a means of achieving these good 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 6 RESPECT AND VALUE

19 Respect for Persons

We ought to respect all persons, but that does not tell us how we ought to act. It is wrong, some writers claim, to treat people in any way that is incompatible with respect for them. But this claim would seldom help us to decide, in difficult cases, whether some act would be wrong.

20 Two Kinds of Value

Some things have a kind of value that is to be *promoted*. Possible acts and other events are in this way good when there are facts about them that give us reasons to make them actual. People have a kind of value that is to be *respected*. This value is not a kind of goodness. Human life may have such value. But we are not morally required to respect the value of anyone's life in ways that conflict with this person's well-being and autonomy.

21 Kantian Dignity

Kant uses 'dignity' to mean supreme value or worth. It is often 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 so. 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 also fails to distinguish between being supremely good and having the kind of moral status that is compatible with being very bad. But we can add this distinction to Kant's view.

CHAPTER 7 THE GREATEST GOOD

22 The Right and the Good

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

23 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 following his various formulas, Kant claims, that everyone could best help to produce this world. This part of Kant's view overlaps with one version of Act Consequentialism.

24 Free Will and Desert

According to one of Kant's arguments, if our acts were merely events in the spatio-temporal world, we could never have acted differently, and morality would be an illusion. Since morality is not an illusion, our acts are not merely such events. 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 the spatio-temporal world.

According to another of Kant's arguments, if our acts were merely such events, we could never be responsible for these acts in some way that could make us deserve to suffer because of what we did. Since we *can* be responsible for our acts in this desert-involving sense, our acts are not merely such events. This argument also fails. We ought, I believe, to accept Kant's claim that, if our acts are merely such events, we cannot 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 the spatio-temporal world. So we cannot deserve to suffer.

CHAPTER 8 UNIVERSAL LAWS

25 The Impossibility Formula

By our *maxims* Kant means, roughly, our policies and underlying aims. According to Kant's *stated* version of 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 self-interested killing, injuring, coercing, lying, and stealing. This formula rightly condemns the making of lying promises. But it condemns such acts for a bad reason, and it mistakenly condemns some other permissible or morally required acts. Though there are some ways in which we could revise this formula, none succeeds.

26 The Law of Nature and Moral Belief Formulas

It is wrong, Kant claims, to act on some maxim unless we could rationally *will* it to be true that this maxim is a universal law. Kant appeals to 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 the 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 beliefs about rationality and reasons. We might appeal to what Kant himself believed. But we are trying to find out whether a Kantian moral theory can help us to decide which acts are wrong, and help to explain why these acts are wrong. So, in asking what Kant's formulas imply, 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.

We should not, however, appeal to our beliefs about which acts are wrong, since Kant's formulas would then achieve nothing. When Kant applies his formulas, he rightly makes no appeal to such beliefs.

27 The Agent's Maxim

Whether some act is wrong, Kant's formulas assume, depends on the agent's maxim. Suppose that our Egoist has only one maxim: '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 this man could not rationally will that his maxim be a universal law, Kant's formulas imply that, whenever acts on his maxim, he acts wrongly. This man acts wrongly even when, for purely self-interested reasons, he brushes his teeth, keeps his promises, 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 when that is the only way to stop some would-be murderer from finding his intended victim. Since Kant could not have rationally willed that his maxim be a universal law, Kant's formula implies that, whenever Kant acted on his maxim by telling anyone the truth, he acted wrongly. That is clearly false.

Kant's appeal to the agent's maxim raises other problems. After considering such problems, some people have come 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, in the sense that covers policies. On a revised version of the Law of Nature Formula:

> We act wrongly unless what we are intentionally doing is something that we could rationally will everyone to do.

On a revised version of the Moral Belief Formula:

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

As we shall see, these formulas need to be revised in some other ways.

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 9 WHAT IF EVERYONE DID THAT?

28 Each-We Dilemmas

It will be simpler to go on discussing Kant's formulas, turning to our revisions 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 everyone. Similar claims apply when we have certain morally 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 this formula works best.

Kant's formula is especially valuable when the bad effects of any single act are spread over so many people that the effects on each person are trivial or imperceptible. One example are the acts with which we are over-heating the Earth's atmosphere. By requiring us to do only what we could rationally will everyone to do, Kant's formula gets us to see how much harm we together do, and provides a strong argument for believing that such acts are wrong. In such cases, common sense morality is *directly collectively self-defeating*.

29 The Permissible Acts Objection

Whether it is wrong to act on some maxim may depend on how many people act upon it. There are some maxims on which it would be 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 such maxims. So Kant's Law of Nature Formula condemns our acting on these maxims even when such acts are clearly permissible. This objection can be partly met by pointing out that most people's maxims are *implicitly conditional*. But, for a full solution, we must revise Kant's formula.

30 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, never to use violence, and to act in ways that ignore what other people will in fact do. This objection can be answered. Kant's formula does not require us to act in these ways. But there is a different problem. Once a few people have failed to do what we could rationally will everyone to do, Kant's formula may permit the rest of us to do whatever we like. Similar objections apply to some *rule consequentialist* and *contractualist* theories. To answer this objection, we should revise Kant's formula in a different 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 by everyone, or by *any other number* of people.

CHAPTER 10 IMPARTIALITY

31 The Golden Rule

Kant's objections to the Golden Rule can be answered.

32 The Rarity and High Stakes Objections

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

33 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 want other people to do such things to us. But, when we apply Kant's Law of Nature Formula, 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 advantages, 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 others 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 could not plausibly defend this claim. So Kant's formulas wrongly permit such acts. Similar claims apply to some of the ways in which those who are rich or powerful treat those who are poor or weak.

34 A Kantian Solution

To avoid this and our other objections, we should again revise Kant's formulas. According to

> *the Doubly Revised Moral Belief Formula*: 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 which 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 11 CONTRACTUALISM

35 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.

36 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.

37 Kantian Contractualism

To reach a better version of contractualism, we should appeal 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.

38 The Deontic Beliefs Restriction

According to Scanlon's similar formula, everyone ought to follow the principles that no one could *reasonably reject*. Since Scanlon appeals to 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.

When we apply any contractualist formula, we cannot appeal to our intuitive beliefs about which acts are wrong. To defend this feature of their view, some contractualists claim that we ought to ignore such intuitive beliefs, since they involve mere prejudice or cultural conditioning. We should reject that claim. And, when we are trying to decide which acts are wrong, we *must* appeal to these intuitive beliefs.

Contractualists should claim instead that 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 12 CONSEQUENTIALISM

39 What Would Make Things Go Best

Consequentialists appeal to claims about what would make things go best in the *impartial reason-involving* 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. Consequentialism can take many forms.

40 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 act on these consequentialist maxims.

41 to 45 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 *UA-optimific*.

Kantians could argue:

> KC: 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-involving sense.
>
> These are the principles whose universal acceptance everyone

would have the strongest impartial reasons to choose.

These impartial reasons would not be decisively outweighed by any conflicting self-interested reasons.

Nor would these reasons be decisively outweighed by any other 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 the optimific principles whose universal acceptance everyone would have sufficient reasons to choose, and could rationally choose.

Therefore

Everyone ought to follow these principles.

KC is the Kantian Contractualist Formula. This argument is valid, and its other premises are true. So this 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 might have no importance. But none of this argument's premises assume 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 appeal to 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 13 CONCLUSIONS

46 Kantian Consequentialism

According to *Act Consequentialism*, or *AC*, everyone ought always to do whatever would make things go best. AC 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* everyone could rationally will. This version of the Kantian Formula implies a version of Rule Consequentialism that is 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.

47 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 this

> *Triple Theory*: An act is wrong just when such acts are disallowed by some principle that is 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 explain what these other properties have in common.

For the Triple Theory to succeed, if 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

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

This claim is in itself very plausible. If we reject (A) that is because this claim's implications conflict too often, or too strongly, with some of our intuitive beliefs about which acts are wrong. Rule Consequentialism conflicts much less often or strongly with these intuitive beliefs. But, if Rule Consequentialists appeal to (A), 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 condemn, 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 cannot be wrong to do what we know 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 universally 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 wrong because they are disallowed by one of the only 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.

CHAPTER 1 REASONS

(The 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 Practical 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.

We can have reasons to believe something, to do something, to have some desire or aim, and to have many other attitudes, such as fear, regret, and hope. Reasons are provided 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. Our reasons to have some belief we can call *theoretical*. Our reasons to have some desire or aim, and to do what might achieve this aim, we can call *practical*.

If we were asked what it means to claim that we have some reason, it would be hard to give a helpful answer. 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 groups of fundamental concepts, such as those of time and space, or possibility reality and necessity, the concept of a reason is *indefinable* in the sense that it cannot be helpfully explained in other terms. [1]

We can have reasons 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. She should not assume that, because I don't know that walnuts would kill me, I have no reason to avoid them. [2]

Reasons may conflict, and they can differ in their 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 to avoid

them. When we have reasons to act in some way that are, when taken together, stronger than any set of reasons we may have to act in some other way, these reasons are *decisive*, and acting in this way is what we have *most reason* to do.[3] When such reasons are *much* stronger than any conflicting reasons, we can call them *strongly* decisive. Many facts give us decisive reasons only in some cases, but there may be some facts that always give us such reasons. On one view, for example, whenever some act would be morally wrong, that gives us a decisive reason not to do it. When we are aware of facts that give us decisive reasons, we can *respond* to these reasons by deciding to do, and then doing or trying to do, what we have these reasons to do. [4]

There is often nothing that we have most reason to do, or decisive reasons to do, because we have *sufficient* reasons to act in any of two or more ways. We have *sufficient* reasons to do something when these reasons are not weaker than any set of reasons we may have to do anything else. When we have reasons *not* to do something, these reasons count *against* acting in this way.

We can use the concept of a reason to explain or identify certain other concepts. These concepts are *normative* in the sense that they imply claims about reasons. Some possible act is

> what we *ought* to do, in what we can call the *wide reason-implying* sense, when this act is what we have most reason to do, or decisive reasons to do. [5]

Even if we never use the phrase 'most reason', most of us often use 'ought' in this sense. There are similar senses of 'should' and 'must', which differ only by implying reasons of different strengths. For example, I might say that you *should* see some movie, that you *ought* to give up smoking, and that you *mustn't* touch some live electric wire.

Though reasons are provided by facts, what it would be *rational* for us to do depends on our beliefs. Suppose that we have some set of beliefs, and that what we believe would, if it were true, give us reasons to act in some way. To save words, I shall call these *beliefs whose truth* would give us these reasons. In most cases, some possible act of ours would be

> *rational* when we have beliefs whose truth would give us sufficient reasons to act in this way, [6]
>
> *rationally required*, or what we *ought rationally* to do, when these

reasons would be decisive,

less than fully rational when we have beliefs whose truth would give us decisive reasons *not* to act in this way,

and

irrational when these reasons would be both clear and strongly decisive.

Similarly claims apply to our actual acts. In most cases, we act

rationally when we act in some way because we have beliefs whose truth would give us sufficient or decisive reasons to act in this way,

and

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

If we have inconsistent beliefs, some act of ours may be rational relative to some of our beliefs, but irrational relative to others. Rather than calling certain acts 'rational' or 'irrational', we may use words with similar meanings, such as 'sensible', 'reasonable', 'smart', 'foolish', 'stupid', and 'crazy'.

When we know all of the relevant, reason-giving facts, what we ought rationally to do is the same as what we ought in the reason-implying sense to do. But, when we are ignorant or have false beliefs, these *oughts* may conflict. Suppose that, while walking in some desert, you have angered some 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 in the reason-implying sense to do. You have no reason to run away, and a strongly decisive reason *not* to run away. As you would be told by any well-informed and friendly adviser, 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. If we say that people have such reasons, we would have to claim that, when

we give people advice, we ought to ignore their false-belief-provided reasons. It is better to say that false beliefs can give people what *merely appear* to be reasons. Suppose that we have some set of beliefs whose truth would give us some decisive reason to act in some way. If these beliefs are true, we would have this reason. If these beliefs are false, we would merely appear to have such a reason. But we wouldn't know that our beliefs were false. So, in both cases, we ought rationally to act in this way. We would then be responding rationally to this reason or *apparent* reason.

These claims are about *normative* or *justifying* reasons. When we believe that we have such a reason, and we act *for this reason*, this becomes our *motivating* reason, or the reason *why* we acted as we did. If I avoid walnuts, for example, my motivating reason may be that 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 snake, your motivating reason would be provided by your false belief that this act would save your life. [7] 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'. We can here ignore motivating reasons.

When we ask what we ought to do, we are most often using 'ought' in the wide reason-implying sense. But we sometimes use 'ought' in one of several *moral* senses, which I shall discuss in Chapter 3. These senses differ in at least two ways from the wide reason-implying sense. First, there are many things that we ought to do only in this reason-implying sense. If I hate commuting, for example, I may have most reason to live close to where I work. If I need to catch some train, I may have most reason to leave some meeting now. 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 can still ask whether this act is what we ought in the reason-implying sense to do. As most people use the words 'ought morally' and 'wrong', it makes sense to claim that we can have sufficient or even decisive reasons to act wrongly. On some widely accepted views, as we shall see, we may sometimes have *no* reason to do what we ought morally to do.

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

We can now turn to the concepts *good* and *bad*. When we call something

> *good* in the *reason-involving* sense, we mean that there are facts about this thing, or its properties, that would, in some situations, give us or others reasons to respond to this thing in some positive way, such as wanting, choosing, producing, using, 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'.

When something is in this sense good, Thomas Scanlon claims, this thing's goodness could not itself give us any reason. Such goodness is what Scanlon calls the *higher-order* property of having *other* properties that might give us certain reasons. The higher-order fact that we had these reasons would not itself, Scanlon claims, give us a reason. [8]

This view needs, I think, one small revision. Suppose that some reliable adviser truly tells me that there are certain facts that give me decisive reasons to go home. This adviser does not tell me what these reason-giving facts are, since she has promised to keep them secret. On Scanlon's view, the higher-order fact that I have these reasons to go home does not itself give me any reason to go home. If that were true, I could rationally decide to stay where I am. I could claim that, though I know that I have decisive reasons to go home, I am not aware of any fact that gives me a reason to go home. But that claim would be false. I am aware of the fact that there are some facts unknown to me that give me decisive reasons to go home. This higher-order fact about these reasons clearly gives me a reason to go home. Rather than denying that this fact gives me a reason, our claim should instead be that this reason is *derivative*, since its normative force derives entirely from the facts, unknown to me, that give me my *non-derivative* or *primary* reasons to go home. This derivative reason has no independent strength or weight.

Similar claims apply to the kind of goodness which is the property of having other, reason-giving properties. If some medicine is the best, this fact might be truly claimed to give us a reason to take this

medicine. But this reason would also 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 strength or weight, they are not worth mentioning in such a claim.

Of our reasons for acting, many are provided by facts about our own or other people's well-being. On *hedonistic* theories, our well-being consists in our having pleasure and happiness, and our avoiding pain and suffering. On *substantive good* theories, our well-being may also consist in some other states or activities, such as loving and being loved, moral goodness, knowledge, and some kinds of achievement. On *desire-based* theories, our well-being consists in the fulfilment either of our actual desires, or of the desires that we would have under certain conditions. On any plausible theory, hedonism is at least a large part of the truth, so my examples will often involve hedonistic well-being.

Facts about our own well-being can give us reasons that are *self-regarding*, or *self-interested*. [9] The different theories that I have just described make partly conflicting claims about which facts give us such reasons. These facts are about possible events, in the wide sense of 'event' that covers states of affairs and acts. When we claim that some possible event would be

> *good for someone*, in the *reason-involving* sense, we mean that there are facts about this event that give this person self-interested reasons to want this event to occur.

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, or on our character. Pain and suffering may not be in these senses bad for us. But it is bad to be in pain. Pain and suffering are bad for us in the sense that these are conscious states that we always have self-interested reasons to want not to be in.

Facts about the well-being of other people can give us *other-regarding* or *altruistic* reasons. 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 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*, since 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 we would have these reasons from an impartial point of view. When we think about possible events that would involve or affect people who are all strangers to us, our actual point of view is impartial. 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 strangers to me. I would have strong self-interested reasons to want these rescuers to save me rather than these many strangers. But, if I were in the impartial position of some outside observer, because I was not one of the people whose lives were in danger, I would have most reason to want the rescuers to save many people rather than saving only one. [10]

We can now can explain another kind of goodness. Of two possible events, one would be

> *better* in *the impartial reason-involving* sense if 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 the rescuers saved more people. This kind of goodness is often called *impersonal,* in the sense that it is not goodness *for* particular people. But many events are impersonally good because they are good for one or more people.

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

2 Reason-Giving Facts

There are two main kinds of view about reasons for acting. According to one group of theories, such reasons are all provided by certain facts about our present desires, or present aims. Some of these theories appeal to what we actually want to achieve. Others appeal to the desires or aims that we would now have, if we had gone through some

process of informed deliberation, knowing the relevant facts and thinking clearly. We can call these our *hypothetical, fully informed* desires or aims. On these *desire-based* or *aim-based* theories---which are sometimes called 'internalist'---what we have most reason to do is whatever would best fulfil or achieve these actual or hypothetical present desires or aims.

According to another group of theories, which we can call *value-based*, reasons for acting are all provided by the facts that make certain possible outcomes worth producing or preventing, or make certain things worth doing for their own sake. Many of these outcomes or acts are good or bad for particular people, or impersonally good or bad. But, as I have claimed, value-based reasons are provided, not by the goodness or badness of these outcomes or acts, but by the facts that make them good or bad.

Many people now accept desire-based or aim-based theories. In economics and the other social sciences, practical reasons and rationality are often *defined* in a desire-based or aim-based way. We ought, I believe, to reject all such theories, and to accept some value-based theory, according to which there are no desire-based or aim-based reasons. [11]

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

There are several possible explanations. First, in many cases, these two kinds of theory partly agree. According to all plausible value-based theories, we have reasons to try to promote our future well-being. Since most of us want to promote our future well-being, desire-based theories also imply that most of us have reasons to act in this way. Similarly, on both kinds of theory, we often have reasons to try to fulfil our other present desires, since what we want would often be claimed by value-based theories to be worth achieving.

Second, we have some reasons for acting that we wouldn't have if we didn't have certain desires. But, though these reasons *depend* on these desires, they are not *provided* by these desires, or by the facts that certain acts would fulfil these desires. These reasons are provided by certain other facts, most of which causally depend on our having these desires. When we have some desire, for example, that may make it true that this desire's fulfilment would give us pleasure or happiness, or that its non-fulfilment would be distressing, or distracting. These facts would give us reasons to try to fulfil this desire.

Suppose next that we must choose between several possible aims, which would all be equally worth achieving. If we want to achieve one of these aims, perhaps because we find it specially appealing, that may give us reasons to believe that we would find this aim's achievement especially rewarding. Our wanting to achieve this aim may also make it easier for us to make the efforts that would be needed to achieve this aim, and the thought of this aim's achievement may give us pleasure in advance. These facts would all give us reasons to adopt and try to achieve this aim. In these and other similar other ways it would not be our desires themselves but these desire-dependent facts that gave us our reasons to try to fulfil these desires. When people claim that our reasons are provided by our desires, it is often such other facts that they really have in mind.

Third, of those who accept desire-based theories about *reasons*, some also accept desire-based theories about *well-being*. On such theories, the fulfilment of some of our present desires would be in itself good for us, even if that fulfilment would give us no pleasure, perhaps because we would not even know that this desire had been fulfilled. If the fulfilment of such desires would be in itself good for us, we would have *value-based* reasons to fulfil these desires. This would be another way in which desire-based theories about reasons would partly agree with some value-based theories.

Fourth, there is a superficial sense in which our desires or aims can be truly claimed to give us 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, which I cannot do unless I leave now. But this would be another *derivative* reason, since this reason 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, I would have no reason to leave now. When I claimed above that no reasons are provided by our desires or aims, I was referring to our non-derivative, primary reasons.

Fifth, as we shall see, some desire-based theorists fail to distinguish between *normativity* and *motivating* force. Our desires give us reasons, these people believe, because it is our desires that move us to act.

Though these kinds of theory often agree that we have reasons to try to fulfil our present desires, they may disagree about how strong these reasons are. On most desire-based theories, the strength of these reasons depends on the strength of these desires. On most value-based theories, the strength of these reasons depends instead on how good, or worth achieving, the fulfilment of these desires would be. We often have stronger desires for what would be less worth achieving, as when

we prefer to have some enjoyable experience in the nearer future, though we know that, if we waited, our enjoyment would be greater. So these kinds of theory often disagree about what we have *most* reason to do, and what we ought rationally to do. There are also many desires that, on plausible value-based theories, we have no reasons to try to fulfil. Some examples are desires for revenge and for some kinds of wealth, power, and fame.

There is a deeper disagreement. On desire-based and aim-based theories, practical reasons are mainly reasons for acting, which are provided by facts about what would fulfil or achieve our present desires or aims. On value-based theories, practical reasons are not merely, or mainly, reasons for acting. As I have said, we also have reasons to *have* the desires or aims that our acts are intended to fulfil or achieve. Our aims are worth achieving when there are facts that give us reasons both to want to achieve these aims and to try to achieve them. Since our reasons for acting are given by the facts that also give us reasons to have these desires or aims, we would have these reasons even if we *didn't* have these desires or aims.

3 Value-based Theories

To save words, I shall now discuss only our desires, though many of my claims would also apply to our aims. Our desires are *instrumental* when we want something as a *means* to something else. Our desires are *telic* when we want something as an *end*, or for its own sake.

We often have long chains of instrumental desires, but such chains all end 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, not for its own sake, but only so that I can finish writing some great novel, and I might want to finish this novel, not for its own sake, but only to achieve posthumous fame. This desire might also be instrumental, since I might want such fame only to refute my critics, or to increase the income of my heirs. But, if I want posthumous fame not even partly as a means but only for its own sake, this telic desire would end this particular chain.

Many people have believed that, at the end 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 get pain from 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, we must have an independent desire to win.

Besides having telic desires for such other things, we may not even want pleasure as an end, or for its own sake. Suppose that we know some relentlessly ambitious politician, whom we find basking in the sun, sipping champagne. When we ask this man what he is doing, he replies 'Enjoying myself'. Given our knowledge of this man's character, this reply is baffling. This man never does anything merely for enjoyment. He then explains that his doctor warned that, unless he allows himself 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.

We can now sketch some of the ways in which, as value-based theories claim, we can have reasons to have particular desires. All desires have *objects*, which are *what* we want. These objects are all events, in the wide sense that covers states of affairs and acts. Though we can be said to want some persisting things, such as some book or bottle of wine, what we want is really to own, use, consume, or have some other relation to these things. Rather than saying that we want some event to occur, I shall say, for short, that we want this event.

Of our reasons to have some desire, some are provided by certain facts about this desire's object, or the possible event that we want. Such reasons we can call *object-given*. These reasons are *telic* when they are provided by facts that make this event good as an end, or worth achieving for its own sake. Such reasons are *intrinsic* when they are provided by this event's intrinsic properties or features. [12] We have *instrumental* reasons to want some event when and because this event would help to produce, or be a means of achieving, some good end.

According to some widely accepted views, 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 there are facts that would make our having some desire good, or good for us. These reasons might also be either telic or instrumental, since our having some desire might be good either as an end or as a means.

On these views, we can have at least four kinds of reason to have some desire, which can be shown 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 all these reasons to have the same desire. If you are suffering, for example, I might have all these reasons to want your suffering to end. What I want would be in itself good, and it may have the good effect of allowing you to enjoy life again. My wanting your suffering to end may be in itself good, and it may have good effects, such as your being comforted by my sympathy.

Similar claims apply to our reasons to have beliefs. Such reasons are *epistemic* when they are provided by facts which are related to the *truth* of some actual or possible belief, by being evidence, or by logically implying this belief, or in some other way. If the clouds are dark grey, for example, that may give us an epistemic reason to believe that it will soon rain. Since these reasons are related to the truth of *what* we believe, we can call them *object-given*. According to some writers, we can also have *state-given* reasons to have some belief: reasons that are provided by facts that would make our *having* this belief in some way good, either as an end or as a means. We are often claimed, for example, to have such reasons to believe in God. These reasons would not be *truth-related*, but *goodness-related*, or value-based. Such reasons to have beliefs are often called *pragmatic*.

The same facts can give us object-given reasons both to want something to happen and to try to make it happen, by acting in some way. Though these reasons are closely related, there is a striking difference between the ways in which we can respond to them. We can respond to reasons for acting in a direct, voluntary way, by doing,

or trying to do, what we have most reason to do. We can also respond directly to our object-given reasons to have certain desires, by coming to have them, and then continuing to have them. But, at least in most cases, these responses are *not* voluntary. These responses are, in this respect, like our responses to epistemic reasons. Though our responses to these two kinds of reason are seldom under our direct voluntary control, they 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 in these non-voluntary ways.

To illustrate some of these claims, we can suppose that it would be bad for us if we had some belief or desire. Consider

> *the Despot's Threat:* Some whimsical despot declares that I shall be tortured for ten minutes unless, at noon tomorrow, 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. [13]

This despot's threat might be claimed to give 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 possibly respond directly to such reasons.

One problem here is that I have *object*-given reasons that count strongly *against* believing that 2 + 2 = 1, and *against* wanting to be tortured. Suppose that, because I fail to respond to this despot's threat, he has me tortured for these ten minutes. 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 = 4. It would be as absurd to claim that I was an idiot in not wanting to be tortured. I might want to be tortured if I knew that being tortured would be my only way to achieve some great good. That might be true, for example, if I have some life-threatening illness and my pain would trigger some healing process in my body. But this example is not of that kind. My 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 cannot want to be tortured for its own sake.

Suppose next that, in a different version of this case, this despot threatens that I shall be tortured unless, at noon tomorrow, I have some belief that it would be easier for me to have, such as the belief

that some sealed box is empty. As before, this threat might be claimed to give me a pragmatic, state-given reason to have this belief. And this reason would be unopposed, since I have no epistemic reason *not* to believe that this box is empty. But, as before, I could not directly respond to this state-given reason. Since I am rational, I could not believe that this box is empty simply because I know that it would be better for me if I had this belief.

When it would be better for us if we had some belief or desire, it is not, I believe, helpful to claim that we have *state*-given reasons to *be* in these states. Our claim should instead be that we have *object*-given reasons to *want* to be in these states, and to *cause* ourselves to be in them, if we can. These are reasons to which we *could* directly respond. I would respond to this despot's threat by wanting to believe that this box is empty, and wanting to have a desire to be tortured for its own sake. And I might be able to cause myself to have this belief and desire by using some technique like self-hypnosis. That would then be what I had most reason to do. Since it is only these object-given reasons to which we can directly respond, these reasons are much more important. My claims below will all be about such reasons.

We might make a stronger claim. There are, I suggest, no state-given reasons. According to what we can call

> *the response requirement*, for some fact to give us a reason, it must be true that, at least in some cases, we or others could directly respond to reasons of this kind.

Suppose that it would be better for me if I were six inches taller, or if I were healthier, or if I knew how to get home. These facts could not give me reasons to *be* six inches taller, to *be* healthier, or to *know* how to get home. I could not have reasons to *be* in these states, because neither I nor others could possibly respond directly to such reasons. [14] I could at most have reasons to want or wish to be in these states, and to cause myself to be in them, if I can. While it is obvious that I could not have reasons to *be* six inches taller, or to *be* healthier, it may seem that, as well as having reasons to *want* to have some belief or desire, and to *cause* myself to have them if I can, I could also have state-given reasons to *be* in these states. But, if this claim seems plausible, that may be only because we can have *object-given* reasons to have beliefs or desires: reasons to which we *can* respond in direct though non-voluntary ways, by coming to have and continuing to have these beliefs and desires.

(More to be added here.)

We can now return to these object-given reasons to have desires, and to the rationality of these desires. Of our reasons to have desires, what are fundamental are intrinsic telic reasons. These are reasons to want some event as an end, or for its own sake, which are provided by this event's intrinsic features. Different value-based theories partly disagree about which events we have such reasons to want. Such theories can appeal, for example, to different views about well-being. These theories can also disagree about *whose* well-being we have reasons to care about, and try to promote. According to *Rational Egoism*, for example, we have reasons to promote only our own well-being. On some theories, the goodness of some ends does not depend, or depend only, on their contributions to our own or other people's well-being. And some of these ends are acts, which are in themselves good, or worth doing. As before, it will be enough here to consider reasons that are provided by facts about our hedonistic well-being.

When we are ignorant, or have false beliefs, our desires may be rational though we have no reasons to have these desires. We can start with the simpler cases in which we know all of the relevant, reason-giving facts.

In such cases, some desire is rational when we want some event that we have sufficient object-given reasons to want. Some desire is *contrary to reason* if we want some event that we have reasons *not* to want, and no reasons, or only weaker reasons, to want. When some desire is clearly and strongly contrary to reason, because we want some event that we have strongly decisive reasons *not* to want, this desire is irrational. Desires that are more weakly contrary to reason are merely less than fully rational.

We have some desire-like states which are not responses to reasons. One large and important group are the likings or dislikings of some present sensations that make our having these sensations pleasant, painful, or unpleasant. It is sometimes claimed that these sensations are in themselves good or bad, since their nature gives us reasons to like them or dislike them. But we do not, I believe, have any such reasons. Nor could these *hedonic* 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 produced by eating milk chocolate, having cold showers, and taking strenuous exercise. Some of these likings or dislikings are odd. Many people hate the

sound of squeaking chalk. I hate the feeling of touching velvet, the sound of buzzing house flies, and the effect of overhead lights. Whether we like, dislike, or are indifferent to these various sensations, we are not responding or failing to respond to reasons.

When we are in great pain, what is bad is not our sensation but our conscious state of having a sensation that we intensely dislike. If we didn't dislike this sensation, our conscious state would not be bad. The nature or quality of some 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, they would not be bad, and we would have no reasons to dislike them.

When we intensely dislike some sensation, we also strongly want not to be having it. And we have many other desires about our own and other people's pleasures and pains. These *meta-hedonic* desires are quite different from our hedonic likings or dislikings. This distinction is clearest when we think about our future conscious states. Though we may strongly want to avoid some future period of pain, we cannot now dislike this future pain.

Unlike our hedonic likings or dislikings, these meta-hedonic desires or preferences can be rational or irrational. These desires provide some of the clearest examples. If one of two future ordeals would be much more painful, for example, this fact gives us a strongly decisive reason to prefer the other. Unless we have 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 what they know would be the more painful of two ordeals, that may always be because this ordeal would have some other feature, such as being deserved, or enabling us to show how tough we are.

Some other cases 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 prefer agony on any future Tuesday to slight pain on any other day. That some ordeal would be much more painful is a strong reason *not* to prefer it; that it 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 someone with a *bias towards the next year*. This imagined man cares equally about his future throughout the next year, but he cares only half as much about the rest of his future. 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 future pains would be just over rather than just under a year from now, that is no reason to care about them only half as much.

No one has these attitudes to time. But many of us have an attitude that is partly similar: caring less about our further 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, at the foreseen cost of making this ordeal much worse. Rather than having slight pain later today, you prefer agony tomorrow. This preference would also be, though more weakly, irrational. [15] Many people often act on such irrational preferences, thereby making their lives go worse.

These claims may seem too obvious to be worth making. Who could possibly deny that we have intrinsic telic reasons to care about our future well-being, such as reasons to want to be happy, and to prefer slight future pain to future agony?

4 Desire-based Theories

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 desire-based theories about reasons. On such theories, we cannot have reasons to want anything as an end, or for its own sake.

According to these theories, all reasons are provided by facts about what would fulfil our present desires. Some theories appeal to our actual present desires. Others appeal to the desires that we would

now have if we had carefully considered all the relevant facts. We can here consider cases in which we *have* carefully considered all these facts, so that these two kinds of desire-based theory coincide.

On these theories, we can have instrumental reasons to want something as a means to something else that we want. We would have such reasons to want be happy, for example, if and because our happiness would help to fulfil some other desire. That might be true if our future happiness would enable us to work more effectively, or would make those who love us happy, as we want them to be. But we cannot have any desire-based reasons to want future happiness, not as a means, but as an end.

Here is why we cannot have such reasons. According to desire-based theories, such reasons would have to be provided by facts about what would fulfil our present desires. If, after informed deliberation, we want future happiness as an end, this fact could give us instrumental reasons to have certain *other* desires, since it would give us reasons to want whatever would make us happy. But the fact that we had this desire could not be truly claimed to give us a reason to have it. Desires cannot be self-supporting. Our wanting happiness as an end could not give us a reason to want happiness as an end. [16]

Suppose next that, after such deliberation, we *don't* want future happiness as an end, nor would this happiness have effects that we want. On desire-based theories, we would then have no reason to want to be happy. There is no other possible fact about our desires that could give us such a reason. Similar claims apply to our future suffering. Suppose that, after informed deliberation, we have no desire to avoid some future period of agony, nor do we have any other desire whose fulfilment this agony would prevent. On desire-based theories, we would then have no reason to want to avoid this agony.

It might be objected that, when we are later in agony, we shall have a very strong desire not to be in this state. This fact may seem to give us a desire-based reason to want to avoid such agony. But this objection forgets what is claimed by desire-based or 'internalist' theories about reasons. On these theories, reasons are provided only by facts about the fulfilment of our *present* desires. This is one of the central claims of such theories. On these theories, reasons involve *motives*, and only our present desires could motivate us.

What may be confusing here is that a *value-based* theory about *reasons*

could be combined with a *desire-based* 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, since they are provided by facts about our future well-being, or what would be 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 want these future *desires* to be fulfilled.

Unlike this desire-based theory about *well-being*, desire-based theories about *reasons* make no appeal to facts about what would fulfil our future desires. These theories could not appeal to such facts, since our future desires could not now motivate us. If we know the relevant facts, and we have no present desires whose fulfilment would be prevented by our having some period of future agony, these theories unavoidably imply that we have no reason to want to avoid such agony. [17]

Desire-based theorists might reply that

(A) everyone wants to avoid all future agony.

If (A) were true, it might seem not to matter that, according to desire-based theories, we have no reasons to want to avoid such agony. If everyone had this desire, these theories would imply that everyone had reasons to try to fulfil this desire, by avoiding agony.

These claims are not a good defence of these theories. First, if we have reasons to have certain desires, acceptable theories ought to imply that we have such reasons.

Second, (A) is false. Some people do not care about the prospect of future pain, if this pain would be far enough in the future. Of the people who have believed that their sins would be punished in the fires of Hell, many have 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 everyone did want to avoid all future agony, we can imagine people who didn't have this desire. Any good theory about reasons must have acceptable implications when applied to imaginary cases, if it is clear enough what such cases would involve.

Some desire-based theorists might give a different reply. These

people appeal to the desires that we would have after informed and *rational* deliberation. So they might claim that

> (B) if we were fully rational, we would want to avoid all future agony.

These people might then claim that, since (B) is true, everyone would always have desire-based reasons to try to avoid such agony.

(B) is ambiguous. Understood in one way, (B) is a normative claim, which would be made by any plausible value-based theory. These theories make claims about what we can call *substantive* rationality. On these theories, we have strong reasons to have certain aims or ends, and to be substantively rational we must want to achieve these ends. One such end is avoiding future agony. If we did not want to avoid such agony, we would not be fully rational, because we would be failing to respond to our strong reasons to have this desire.

Desire-based theorists cannot make such claims. As I have just argued, these theories imply that we have no reasons to want, for its own sake, to avoid future agony. When some desire-based theorists appeal to the desires that we would have after informed and *rational* deliberation, they are referring to *procedural* rationality. According to these writers, when we are deciding what to do, we ought to think carefully about the possible outcomes of our acts, adopt aims that are easier to achieve, use our imagination, and follow certain other rules. But we are not rationally required to have any particular desires, or aims. That is what makes these theories *desire-based*. We can be procedurally rational whatever we care about, or want to achieve.

If desire-based theorists appealed to (B), they would have to mean that

> (C) if we deliberated in ways that were fully procedurally rational, we would in fact want to avoid all future agony.

When applied to some actual people, this claim, I believe, is false. And we can easily imagine people who, though deliberating in these ways, did not want to avoid some period of future agony. If such people had no other present desires whose fulfilment would be prevented by such agony, all desire-based theories imply that these people would have no reason to want to avoid this agony.

We can next return to the reason-involving senses in which events can

be good or bad. Future events would be in themselves good for us when their intrinsic features give us self-interested reasons to want them to occur. According to desire-based theories, as I have just argued, we have no such reasons, so no future event could be good or bad for us in this reason-involving sense. Nor could future events be in themselves good or bad in the impartial reason-involving sense. According to desire-based theories, there are no future events that, from an impartial point of view, everyone or anyone would have reasons to want as an end.

Some desire-based theorists about reasons use 'good for someone' in a different 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. [18]

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. [19]

When 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.

As Rawls notes, many other writers propose or accept such definitions.

If this is how we define what is best for someone, or what is someone's good, it would be a merely psychological fact that some life would be best for someone in this sense. Rawls's theory of the good may seem normative, since he appeals to what it would be *rational* for people to choose. But, though it is a normative question which kinds of deliberation are procedurally rational, it is a psychological question what, after such deliberation, someone would in fact choose. [20] The most rational plan 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. [21]

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. [22] There is nothing, Rawls assumes, that we have any 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'. [23] 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 facts, this man would in fact choose this plan of life. Consider, for example,

> *Blue's Ideal*: After informed and procedurally rational deliberation, *Blue's* strongest desire is that the rest of his life consists only of unrelieved suffering. Blue therefore chooses the 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, since no one would choose a life of unrelieved suffering. But this claim is irrelevant. 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 this merely imaginary person. As I have said, any acceptable theory must be able to be applied successfully to such imaginary cases, if it is clear enough what these cases would involve.

It might next be objected that my example is *not* clear, since my description of *Blue's Ideal* makes no sense. For it to be true that we are suffering, we must have a strong desire *not* to be in this state, so it may seem impossible that anyone could want to be in a state of unrelieved suffering. But this objection overlooks the difference between our attitudes to *present* and *future* suffering. As I have also said, some people care little or not at all about the prospect of future suffering, if this suffering would be far enough in the future. And we can imagine people who don't care about any future suffering, or who want their future to be filled with suffering. Perhaps no actual person would be so irrational. But it is clear enough what such cases would involve. And it is by considering such cases that we can best see the implications of desire-based theories about reasons, and of present-choice-based theories about well-being, or about someone's good.

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 informed and rational 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. [24] And Rawls, I assume, 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. When we call some life

> 'best for someone' in the reason-involving sense, we mean that this is the life that this person would have the strongest self-interested reasons to want, and to choose.

Rawls cannot use 'best for someone' in this sense, since he accepts a desire-based theory about reasons, and such theories imply that we have no self-interested reasons. But Rawls might use 'best for someone' in some sense that is not reason-involving. And he might then claim that, in this other sense, it would be bad for Blue to have a life of unrelieved suffering.

Such a claim, however, would achieve little. If we accept some desire-based theory about reasons, we cannot avoid implausible conclusions by appealing to claims about what is good or bad for people. On these desire-based theories, what we have most reason to do is whatever would best fulfil our present informed desires. What Blue now most wants, after informed deliberation, is a life of unrelieved suffering. So these theories unavoidably imply that, even if such a life would be in some other sense very bad for Blue, this is the life that Blue now has most reason to give himself, if he can. If Blue could ensure that he will have a life of unrelieved suffering, by getting himself enslaved to some cruel owner, or committing some crime for which the punishment is endless hard labour, this would be what, on desire-based theories about reasons, he has most reason to do, and what, if he knew

the facts, he would be rationally required to do.

We have been discussing some extreme imaginary cases. But similar claims apply to actual cases. According to desire-based theories, just as we cannot have reasons to want, for its own sake, to avoid suffering, we cannot have such reasons to want ourselves or others to live happy and worthwhile lives, or to achieve any other good ends. Similarly, on aim-based theories, we cannot have reasons to have any ultimate aim. On these widely accepted views, we can now conclude, *nothing matters*.

Some desire-based theorists would admit that, on their view, nothing matters in an impersonal sense. It is enough, these writers claim, that things matter to particular people. But this reply shows how deeply these views differ. On value-based theories, things matter in the normative sense that we have *reasons* to care about these things. When desire-based theorists claim that things matter to particular people, they mean only that these people *do* care about these things.

These bleak views are seldom defended. Most desire-based or aim-based theorists take it for granted that we cannot have reasons to care about anything for its own sake.

Some desire-based theorists appeal to the claim that 'ought' implies 'can'. These people argue:

> For us to have reasons to do something, it must be true that we *could* do it.
>
> We couldn't do something if, even after informed deliberation, we would not be motivated to do this thing.

Therefore

> For us to have reasons to do something, it must be true that after informed deliberation, we would be motivated to do this thing.

We ought to reject this argument's second premise. Suppose I claim, 'You ought to have helped that blind man cross the street', and you reply, 'I couldn't have done that'. If I ask 'Why not?', it would not be enough for you to say, 'Because I didn't want to'. We *could* do something, in the relevant sense, if nothing stopped us doing this thing except our desires, or other motives.

Some other people argue:

> For some fact to give us a reason, it must be possible that we act *for this reason*.
>
> Whenever we act for some reason, we are motivated to act in this way, so this reason is a desire-based or *internal* reason.

Therefore

> All reasons for acting are desire-based or internal. [25]

As before, we ought to reject this argument's second premise. When we are motivated to act for some reason, this fact cannot imply that this reason must be desire-based rather than value-based. It would be absurd to claim that, for some reason to be value-based, it must be impossible for us to be motivated to act for this reason.

There is another line of thought that leads many people to accept desire-based theories. On what I call value-based theories, the fact that we have some reason is an *irreducibly normative* truth. Of those who accept desire-based 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 described by the natural and social sciences, such as physical and psychological properties and facts. Irreducibly normative truths are incompatible, these people assume, with a scientific world-view.

These naturalists give *reductive* accounts of desire-based or aim-based reasons for acting. According to some *analytical* naturalists, 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 informed deliberation, this person would be motivated to act in this way, or we mean something 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 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*. It may then seem that we should identify reasons with certain facts about our desires, or about how we might be motivated to act, since we can thereby explain the normativity of these

reasons. Value-based reasons cannot be regarded in such ways, since we have these reasons even if we are *not* motivated to act upon them. [26]

Of the writers who give such reductive accounts, some 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.

Naturalism is, I believe, mistaken. Though I shall not try to defend this belief here, and my claims will not depend on this belief, it is worth making one remark. If there could not be normative reasons for acting, there could not be normative reasons for having beliefs. Such 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. [27]

If naturalism is false, we ought, I believe, to accept some value-based view about practical reasons. If we can have reasons for believing, and reasons for acting, we can also have reasons for having the desires or aims that our acts are intended to fulfil or achieve. We can have reasons, for example, to prefer slight future pain to future agony.

CHAPTER 2 RATIONALITY

5 Rational Desires

We can now return to questions about the rationality of our desires. When we are ignorant, or have false beliefs, it may be rational for us to want what we have no reason to want. So we can start by asking how the rationality of our desires depends on our beliefs.

We should again distinguish here between some desire itself, and our state of 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. And there are some merely possible desires, which no one has. Similar claims apply to beliefs. The words 'desire' and 'belief' are ambiguous, since they can refer either to some desire or belief itself, or to our having this desire or belief. Though I shall sometimes say which of these I mean, I hope that, in other passages, the context will make my meaning clear.

Our having some desire *causally* depends on our having some belief when we have this desire because we have this belief. 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 might want to sleep, for example, because I believe that, unless I get some sleep, I shall perform badly in some interview tomorrow. Since my having this desire would be a rational response to what I believe, my having this desire would be both caused by, and justified by, my belief.

My having this desire would be rational because this desire *itself* would

be rational. So we can now ask how the rationality of our desires *normatively* depends on our beliefs. The rationality of our *having* some desire may partly depend on the relation between this mental state and our other mental states. But no such claim applies, I believe, to our desires themselves. That is clearest when we consider some merely possible desire. If no one has some desire, this desire's rationality cannot depend on its relation to other mental states. But such a desire can be *intrinsically* or in itself rational or irrational. One example is a desire to live a life of unrelieved suffering. This desire is in itself irrational, since suffering is a state that everyone has strong reasons to want to avoid.

I believe that, at least in most cases:

> (1) Whether some desire is rational depends only on facts about this desire's object, or the possible event that we want. And what is relevant is this desire's *intentional object*, or the possible event that we want with the features that we believe that this event would have.
>
> (2) This desire is rational if these features give us sufficient reasons to want such an event.
>
> (3) it is irrelevant whether our beliefs about this event are true, or rational.

These claims express what we can call the *intentional-object* view.

Some possible *act* would be rational, I have claimed, when we have beliefs whose truth would give us sufficient reasons to act in this way. We can similarly claim that

> (4) some possible desire would be rational when we have beliefs whose truth would give us sufficient reasons to have this desire.

But this claim may be misleading. At least in most cases, as I have just said, the rationality of some desire does not depend on this desire's relation to any *other* mental state. Suppose for example that I want to be happy. I couldn't have this desire unless I had some true beliefs about what happiness involves. But these beliefs are *part* of my desire, since happiness is what I want. Similar claims apply to our instrumental desires. Suppose that I want to take some dose of morphine because I believe that this act will relieve my pain. This desire depends on a particular belief about the effects of this act. But

this belief is also part of my desire, since what I want is to take morphine and thereby relieve my pain. In both cases, the rationality of my desire depends only on facts about its intentional object, or what I want.

There is a partial parallel here with the rationality of some beliefs. The rationality of most of our beliefs depends on their relation to our other beliefs, and to other mental states, such as our perceptual experiences. But some of our beliefs are rational or irrational simply in virtue of their content, or *what* we believe. Some belief is intrinsically irrational, for example, if its content is an obvious contradiction. And some belief is intrinsically rational if its content makes it obviously true. Two examples are the beliefs that 2 + 2 = 4, and that we have reasons to want to avoid agony. The rationality of each of these beliefs depends only on the content of this belief.

Many people accept views that conflict with these claims. Our desires are irrational, some people claim, when they causally depend on false beliefs. Hume, for example, writes that though desires cannot be 'contrary to reason', they are, in a loose sense, 'unreasonable' when they are 'founded on false suppositions'. That is not true. Just as false beliefs can be reasonable or rational, so can desires that depend on false beliefs.

According to many other writers, 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 to the age of 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?

Our answer should be No. Given my belief, 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 believe that smoking would damage my health.

When some people make the claims that I am here rejecting, they might be discussing the rationality, not of our desires themselves, but only of our *having* these desires. These people might claim that, though my desire to smoke is rational, my *having* this desire is *not* rational. But that claim, I believe, is also false. If I want to smoke because I believe that smoking will protect my health, my having this desire is rational. I have this desire because I have a belief whose truth would give me a strong reason to have it. My having this desire is a rational response to this apparent reason.

When we want something as an end, similar claims apply. 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. Given my belief, however, my desire to go to this party is rational. And, if I wanted go to this party because I had the rational belief that I would *not* enjoy it, my desire would *not* be rational.

In these examples, my desires are rational only when they depend on *irrational* beliefs. If this claim seems paradoxical, that is because we are failing to distinguish between two kinds of rationality and reason. The rationality of our *beliefs* normatively depends on whether, in having these beliefs, we are responding to *epistemic* or truth-related reasons or apparent reasons to have these beliefs. The rationality of our *desires* normatively depends, not on the rationality of our beliefs, but on whether, in having these desires, we are responding to *practical* reasons or apparent reasons to have these desires. We might respond well to either set of reasons or apparent reasons, while responding badly to the other. We might be practically rational but epistemically irrational, or practically irrational but epistemically rational.

Of those who claim that the rationality of our desires depends on the rationality of our beliefs, most assume that we have no reasons to have our desires. Our desires can be rational or irrational, these people claim, only in the indirect and derivative sense that these desires causally depend on rational or irrational beliefs. But we do have reasons to have our desires. As value-based theories claim, we have reasons to want some events as ends, or for their own sake; and, as desire-based theories also claim, we have reasons to want the means of achieving some of our ends. Since we can have reasons to have our desires, it is on how well we respond to *these* reasons or apparent reasons that the rationality of our desires depends.

Similar remarks apply to our acts. Given my irrational beliefs that smoking will protect my health and that I shall enjoy some party, it would be rational for me to smoke and to go to this party. Our claim should be only that, if my beliefs are false, I may have no reasons to act in these ways.

As well as being aware of the facts that give us certain reasons, we may have beliefs about *whether* these facts give us reasons. Our desires often partly depend on such beliefs. As Scanlon claims, many of our desires can be more fully described as states of being motivated by some belief about reasons. [28] We want many things because we believe them to be good in some reason-involving sense.

We can have rational beliefs and desires, and act rationally, without having such beliefs about reasons. Some children respond rationally to their awareness of the facts that give them certain reasons, though they do not yet have the concept of a reason. Similar claims apply to some animals, such as dogs and cats, though these animals will never have the concept of a reason. And some rational adults seem to lack this concept, or to forget that they have it. That seems true of Hume, for example, when he declares that no desires or preferences could be unreasonable, or contrary to reason.

If we have beliefs about which facts give us reasons, our desires and acts may in part be rational responses to what we believe. But we sometimes fail to respond rationally to what we believe. We may want some things that we believe that we have no reasons to want, and strong reasons not to want. That is true of some exhausted parents when they want to hit their howling babies. And it is true of me whenever 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, do not *match* our normative beliefs.

Most of our desires are rational, I have claimed, when they depend on certain beliefs about what we want, and what we believe would, if true, give us reasons to have these desires. It is irrelevant whether these beliefs are true, or rational. But, when our desires partly depend on certain *normative* beliefs, such as beliefs about which facts give us reasons, it *may* be relevant whether *these* beliefs are true, or rational. For some fact to give us a reason, it is not enough that we believe that it gives us a reason. And, for our desires to be rational, it is not enough

that we believe that we have reasons to have them. Rationality is not so easily achieved. For us to be fully practical rational, our beliefs about our reasons must be rational.

It might be objected that, if 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.

That claim applies, however, only to most cases. When we are considering beliefs about practical reasons, these kinds of rationality and reason overlap. As Scanlon claims, our desires often consist in our being motivated by some belief about what we have reason to want. 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, our having such beliefs makes us in one way practically irrational. Much of our practical reasoning consists in theoretical reasoning about practical questions, by trying to reach true beliefs about what we have most reason to want and to do. To be fully practically rational, it is not enough to respond rationally to our awareness of the facts that give us reasons, or to our apparent reasons. If we have beliefs about what we have most reason to want and to do, these must be rational beliefs, and we must respond to these beliefs by having these desires, and by doing, or trying to do, what we believe that we have most reason to do.

To illustrate some of these claims, let us compare three imagined people:

> *Scarlet* prefers one hour of agony tomorrow to five minutes of slight pain on any other day of next week.
>
> *Crimson* prefers one hour of agony tomorrow to five minutes of slight pain later today.
>
> *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 they all believe that, other things being 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 we have met before. On Scarlet's view, we have reasons to care about our future well-being, including all of our future pleasures or pains, *except* when these pleasures or pains would come on any future Tuesday. Since tomorrow is a Tuesday, Scarlet believes that he has decisive reasons to prefer agony tomorrow to slight pain on any other day of next 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 one hour of agony tomorrow to five minutes 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, does *not* have this preference, so he chooses the slightly longer pain tomorrow.

When Scanlon discusses people like Scarlet and Crimson, he claims that these people are not irrational, but are merely making substantive mistakes about which facts give them reasons. We should call someone irrational, Scanlon suggests, only when this person 'fails to respond to what he or she acknowledges to be relevant reasons'. [29]

We are irrational in the ordinary sense when our beliefs, desires, or acts make us open to strong rational criticisms. When we are open only to weaker criticisms, we are merely less than fully rational. If Scanlon is using 'irrational' in this ordinary sense, his claims about these cases are not, I believe, justified. When Scarlet doesn't care about agony next Tuesday, he is not failing to respond to what he acknowledges to be some relevant reason. Scarlet believes that he has no reason to care about this agony. Since Scarlet's preference matches his beliefs about reasons, he avoids one kind of irrationality. But, in failing to care about this future agony, he is failing to respond to a very clear and strong reason. And he has a very irrational belief. If some pain of ours will be on a future Tuesday, that does not give us the slightest reason to care about it less. Scarlet's substantive mistake is so gross that he is not merely irrational but in one way insane.

Crimson's preference is less irrational, since this preference does not draw an arbitrary line, and it is not so implausible to believe that we

have reasons to care more about our nearer future. But Crimson's version of this view is much too extreme. It is irrational to believe that we have decisive reasons to prefer one hour of agony tomorrow to five minutes of slight pain later today. Since Crimson's preference also matches his beliefs about reasons, he avoids one kind of irrationality. But, in preferring this agony to this slight pain, Crimson is also failing to respond to a clear and strong decisive reason, and his preference matches his beliefs only because he has an irrational belief.

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 these 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 whether we 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 reasons, and we admit our faults, we may be less irrational than those who have irrational beliefs and much greater unadmitted faults. [30]

There is one way of arguing that our desires or preferences are irrational only when they fail to match our own beliefs about reasons. On some views, there are no truths about rationality or reasons. According to some *non-cognitivists*, for example, when we say that people have reasons to have some belief or desire, or to act in some way, we are not claiming something that could be true or false. We are merely expressing some positive attitude towards such beliefs, desires, or acts. If there were no truths about which facts give us reasons, so that we could not make substantive mistakes, that might seem to support the view that we are irrational only when we fail to want or do what we ourselves believe that we have most reason to want or do.

Some 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 did what he believed to be his duty, by committing mass murder, what he did was 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 now turn briefly to different versions of these imagined cases. Scarlet and Crimson, we can now suppose, both accept some desire-based theory about reasons. Though these people both prefer agony tomorrow to slight pain on some other day, they do not believe that they have 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 informed desires. Since Scarlet and Crimson now prefer the future agony to the 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 one hour of agony to five minutes of slight pain, they are failing to respond to a clear and strongly decisive reason. But their beliefs are less irrational. It is crazy to believe that, though we have reasons to prefer slight pain to most future agony, we have no such reasons if our agony would be on a future Tuesday. It is not crazy to believe that all reasons are given by desires, and that we have no reasons to want anything for its own sake. And some people accept such desire-based theories because they were taught to accept them, and their teachers didn't even mention any value-based theory. Though desire-based theories are, I believe, false, it may not be irrational for these people to accept such theories.

Many of these people, moreover, have rational 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 at the foreseen cost of making this ordeal more painful. Such people respond rationally to the facts that give them reasons to care about their future. Their mistake is only in their failing to believe that they have these reasons. Some desire-based theorists 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.

We can next consider some other views about the rationality of desires. According to one common 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 is in itself rational depends, I have claimed, on facts about this desire's intentional object, or what we want. Whether it is rational for us to *have* some desire may partly depend on other facts. It may be relevant, for example, how we came to have some desire. If I am in prison, and 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 could. 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. [31]

According to some writers, the rationality of our desires 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 in themselves rational even if we were hypnotized or indoctrinated into having them. If we often act against our interests, for example, because we have little concern about our further future, we might be hypnotized into having such concern. Or we might be indoctrinated into loving our enemies, and wanting to do at least one good deed in every day. Such love and such desires are, I believe, fully rational. Suppose next that, after autonomous deliberation, we want to avoid eating even at the cost of starving ourselves to death, or we have some other desire for something that is wholly undesirable. The autonomous origin of these desires would not make either them, or us, rational. On the contrary, we would be *less* irrational if, rather than forming these desires through autonomous deliberation, we were made to have them by some form of outside interference, like hypnosis.

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

According to a different kind of view, our desires or preferences are irrational when they are inconsistent. Two beliefs are inconsistent if they could not both be true. This definition cannot apply 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 would 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. So our desires might be claimed to be 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 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, they 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 sets of preferences, there is more scope for inconsistency. We might believe that A is better than B, which is better than C, which is better than A; and we might have preferences that match these beliefs. Such beliefs and preferences may, I believe, be inconsistent, and in that way irrational. But such cases would also be fairly rare. [32]

Some claims to be added here.

The rationality of our desires seldom depends, I have claimed, either on their origin, or on their consistency with 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 other beliefs, or both. There are few beliefs whose rationality depends only on their content: or what is believed. That is true of beliefs about some necessary truths or falsehoods, such as some mathematical or logical beliefs. Some belief is intrinsically irrational, as I have said, 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 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 their 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. Whether we can rationally believe that this match obtains depends on our other beliefs, our perceptual experiences, and the other evidence available to us.

No such claims apply to our telic intrinsic desires. As I have said, 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, the rationality of this desire depends only on our beliefs about this desire's object, or *what* we want. These desires are in themselves rational, as value-based theories claim, when their intentional objects would be in themselves relevantly good, or worth achieving. This is the central, fundamental truth which is either denied or ignored by most of the theories that we have just been considering. [33]

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, or contingent, but necessary. Undeserved suffering, for example, could not have failed to be in itself bad. For these 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.

6 Sidgwick's Dualism

Value-based theories about reasons can differ in several ways. One difference is in the range of things that these theories claim to be good or bad as ends. On some theories, all such ends involve our own or other people's well-being. On other theories, as I have said, some things are good in ways that do not depend, or depend only, on their contribution to anyone's well-being. Nor is it only outcomes that are worth achieving, since some things are 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.

Value-based 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-involving sense, if we are doing what, from an impartial point of view, everyone would have most reason to want us to do. On one version of this view, what would be impartially best is what 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. [34] 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.[35]

Of these three views, Sidgwick's, I believe, is the closest to the truth. According to Rational Egoists, it could not be rational to act in any way that we believe would be worse for ourselves than some other possible act. That is not true. Such an act might be rational, for example, when and because we believe that this act would make things go impartially much better. I could rationally injure myself, for example, if that were the only way in which some stranger's life could be saved. According to Rational Impartialists, it could not be rational act in any way that we believe would be impartially worse than some other possible act. That is not true. Such an act might be rational, for example, when and because we believe that this act would be much better for ourselves. I could rationally save my own life, for example, 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 theories, all reasons are precisely comparable, since there are precise truths about the relative strengths of all of our desires. According to value-based theories, when we must choose between two things that are exactly similar, such as two cherries or two copies of some book, we may have precisely equal reasons to make either choice. But, on plausible value-based theories, most reasons are only roughly 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. Even in principle, there is no scale on which we could compare the reasons that are provided by some pain's intensity and its duration. Nor are there such truths 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. But such reasons *are* comparable, 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 views are hard to defend. Suppose that we are choosing between some architectural plans for some new building. If economic and aesthetic reasons were wholly incomparable, this would imply that

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

and that

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

We can imagine how one of these preferences might be rational, since we might have reasons to give absolute priority either to our building's beauty, or to its cost. But it would be most implausible to claim that *both* these preferences would be rational, as would be true if these reasons were wholly incomparable. Similar claims apply to our reasons to keep our promises and to help strangers. It would be most implausible to claim that even the weakest reasons of either of these kinds could not be outweighed by even the strongest reasons of the other kind. As these examples suggest, 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. [36] 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 called 'the fundamental fact for ethics.' [37]

Sidgwick's Dualism also rests on what Thomas Nagel calls our *duality of standpoints*. [38] We live our lives from our own, personal point of view. But we can also think about the world, and our relations to other people, as if we had the impartial point of view of some outside observer. When we ask what we have most reason to do, we reach different answers, Sidgwick claims, from these two points of view.[39] 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, however,

impartial reasons are supreme, since we always have most reason to do whatever would be impartially best. [40]

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

Sidgwick's reasoning seems to be this:

> When 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.
>
> From our personal point of view, self-interested reasons are supreme. 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.
>
> 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 could rationally do either what would be impartially best, or what would be best for ourselves. [42]

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 years of agony. These are unacceptable conclusions. If someone acted in such a way, our first reactions would be horror and indignation. But, as well as being very wrong, such acts would not be fully rational. No one could have sufficient reasons to give such absolute priority to their own well-being. [43]

Such acts would not be rational, we might claim, 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 can reject that claim. We might also reject Sidgwick's claim that we could always rationally do what would make things go best. As an *Act Consequentialist*, Sidgwick believes that such acts are always right. Most of us reject this view. It is often 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 do them.

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. Sidgwick overstates the rational importance of personal identity. As Sidgwick claims, we have reasons to be specially concerned about our own future well-being. But we have other, similar reasons. Our reasons to care about our future are 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 give us reasons to be specially concerned about their well-being. We can have reasons to benefit these people which are much stronger than some of our reasons to benefit ourselves. So we can reject Sidgwick's claim that, 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, I have claimed, 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? [44]

This egoistic attitude would be, as Nagel remarks, 'very peculiar.' Unless we are psychopaths, 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. [45] And we have such impartial reasons even when our actual point of view is not impartial. We can have reasons to benefit strangers that conflict with, and are stronger than, some of our self-interested reasons. As I have said, we would not have sufficient reasons to give ourselves some minor benefits rather than saving 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. We can agree that, for some purposes, it is worth asking what we would have most reason to want, or prefer, if we were in the impartial position of some outside observer. That may help us to avoid some kinds of bias. And it is worth asking how it would be best for things to go in the impartial reason-involving sense. But, when we ask 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. [46]

When we assess the strength of all our reasons from the same, actual point of view, our partial and impartial reasons *are* comparable. [47] Some reasons of either kind could be stronger than, or outweigh, some reasons of the other kind. But Sidgwick's view is partly right, I believe, since these reasons are only *very roughly* comparable. According to what we can call

> *wide value-based views*: 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.

The word 'often' allows for various exceptions. Different wide value-based views make conflicting claims about when we would *not* have sufficient reasons to act in such ways. We ought, I believe, to accept some view of this

kind.

To illustrate such a view, suppose that, in *Case One*, I could either save myself from some serious injury, or save some stranger's life. 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 some stranger's life. And I would have such reasons, I believe, whether my injury would be as little as losing one finger, or as great as losing both legs. I am also inclined to believe that, in *Case Two*, I could rationally save either my own life, or the lives of several strangers, 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 would be neither weaker nor stronger than my reason to save my own life. If that is 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 they are provided by facts whose description need not make any reference 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 this suffering if we can, regardless of the sufferer's relation to us. 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. [48]

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 *our own* well-being, and the well-being of 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 benefit ourselves are always stronger than our reasons to give some equal benefit to some stranger; but this difference in strength 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 required amount of 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 just as much weight as, or somewhat greater weight than, the weight 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 to give this morphine to the stranger in the next bed. And I would have such reasons even if this stranger's pain was no worse than mine, or even if my pain was worse.

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.

When the stakes are as high as this, we might be claimed to be rationally required to give some priority, or 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. Though I would not be making my decision from an impartial point of view, I would know that 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. [49]

This belief may seem hard to defend. 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 may seem to break this tie, or tip the scale, giving me decisive reasons to save myself. [50]

This objection can, I think, be answered. Though we have reasons to be specially concerned about our own well-being, that does not imply that we are always rationally required to benefit ourselves rather than giving some equal benefit to some stranger. As I have said, I could rationally give my dose of morphine to the stranger in the next bed. On one view of this kind, which 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 rationally either give greater weight to our own well-being, or give equal weight to everyone's well-being. [51]

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. But they may also require us *not* to give such 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 *my* life. In permitting us to give great priority to our own well-being, but requiring us *not* to give such great priority to the well-being of strangers, these versions of Pure Dualism recognize and endorse our reasons to be specially concerned about our own well-being.

Here is another way to make this point. On some wide value-based views, when we are choosing between morally permissible acts, we are always rationally required to give to our own well-being greater weight than we give to any stranger's well-being; but this requirement is very imprecise, since the amount of extra weight might be anywhere between *slightly more* and *many times as much*. On Pure Dualism, the required amount of extra weight might be anywhere between *none* and *many times as much*. These views are very similar.

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. There are no other relevant differences between these people.

It would be implausible to claim that I could rationally save this stranger rather than saving my child, or my friend. But Pure Dualists

could reject this claim. In such cases, they could claim, I would not be choosing between morally permissible acts, since I ought morally to give priority to my child or friend. If I saved this stranger rather than my child or friend, this act would *not* be generous and fine. And I would have other decisive reasons to save my child or friend.

Similar claims might apply to *First Shipwreck*. I might have young children who depend on me, or have other special obligations to certain other people. That might make it wrong for me to save this stranger rather than myself, since I would then fail to care for my children, and I would not fulfil these other obligations. These facts would give me further reasons to save myself, which would tip the scale, and be decisive. This stranger might have similar obligations that his death would leave unfulfilled, but those obligations would not be mine. So Pure Dualists could claim that, 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 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. But we need not here decide whether that is true, or whether this act, though morally admirable, would be less than fully rational.

CHAPTER 3 MORALITY

7 The Profoundest Problem

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 our duty, we shall get the happiness that we deserve. That is claimed by most of the world's great religions.

Sidgwick doubted whether 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'. [52]

Sidgwick's problem was in part that Moral Rationalism and Rational Egoism both seemed to him intuitively very plausible, but that, if duty and self-interest sometimes conflict, these views cannot both be true. If we had to choose between two acts, of which one was our duty but the other would be better for ourselves, these views imply that we would have most reason to act in each of these ways. That is inconceivable, or logically impossible. Just as we could not keep most of our money in each of two different wallets, we could not have most reason to act in each of two different ways. So, if duty and self-interest sometimes conflict, we would have to reject or revise either Moral Rationalism or Rational Egoism.

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'. We would be knaves if we didn't do our duty, but fools if we did. Other writers reject Rational Egoism. According to these writers, we could never have sufficient reasons to act wrongly, not even if some wrong act 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 view, 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 whatever 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 all such cases, we would not have stronger reasons to act in either of these ways. Either act would be rational. [53]

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. [54]

These remarks are 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. But it would be bad if, in cases of conflict, we and others would have sufficient reasons to act wrongly. The *moralist's problem*, we might say, is whether we can avoid that conclusion. And it would be disappointing if, in such cases, reason gave us no guidance. We may hope that, in at least some of these cases, we could make a rational decision, since 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 could 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. On some views, some such reasons might be as strong as our reasons to do our duty. We might have sufficient reasons to act wrongly if, for example, that was our only way to save from great pain or death, not ourselves, but our close relatives, or other people whom we love. Such an act might be fully rational.

We might also be claimed to have 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 that view, since we believe that some acts would be wrong even if they would make things go best. It might be wrong, for example, to kill one person even when that is our only way to save the lives of many other people. Despite this act's wrongness, the fact that we would be saving these many people's lives might give us a sufficient reason to act in this way. This might 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, or to avoid acting wrongly. This view we can call *Weak Moral Rationalism*. On some other views, we may sometimes fail to have such reasons. Rawls for example claims that, if our informed desires would be best fulfilled by acting unjustly, we would not have sufficient reasons to do what justice requires. [55] According to such desire-based theories, we might have no reason to do our duty, and have decisive reasons to act wrongly. It might then be *irrational* for us to do our duty.

To cover these 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 different answers, since we might sometimes have sufficient or even decisive reasons to act wrongly. Our problem is to decide whether we do sometimes have such reasons, and, if that is true, what conclusions we should draw.

In considering these questions, it will help to distinguish between two senses of 'normative', and two conceptions of normativity. On the *reason-involving* conception, normativity involves normative 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 'disinterested' to mean 'uninterested'. Such requirements or rules are sometimes called 'norms'.

These conceptions of normativity are very different. On the rule-involving conception, we can create new normative truths merely by 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 no rules about the correct spellings of English words. Later writers of English have created such rules. In contrast, on the reason-involving conception, there is normativity only when there are normative reasons. We cannot create such reasons merely by introducing some rule, or requirement.

There is a deeper difference. When there are rules or requirements, we often have reasons to follow them. But these reasons are provided, not by the mere existence of these rules, but by certain other facts, such as some facts that depend on people's acceptance of these rules. We have reasons to use words with their correct meaning, because that will help us to be understood. We have reasons to drive on the correct side of the road, because we shall then 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 superficial ways. What we can create are only the particular forms that, in different communities, more fundamental, universal, and uncreated requirements take.

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 requirements of the code of honour are all that matter in this code's terms. To be able to make significant claims about the relative importance of moral requirements and requirements of these other kinds, we need some non-moral, neutral criterion.

Reasons provide such a criterion. We can compare the strengths of our reasons to follow these requirements. The men who fought duels, for example, 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-involving 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 have some belief, or to act in some way, we may be rationally required to have this belief, or to act in this way. Such requirements might have little importance in the reason-involving 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-involving sense, we are not assuming that rationality matters. [56]

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 might 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 certain duels, and it would 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-involving* sense.

This objection is in part correct. 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. 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, as some of us believe, have supreme importance in the reason-involving 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 reasons to care about morality, and to avoid acting wrongly. In discussing morality, we shall be discussing these reasons. And these are among the reasons that most need discussing, because they raise some of the most difficult questions.

Second, to judge the strength of these reasons, we must answer certain questions about which acts are wrong. Here is one example. According to Act Consequentialists, we are morally required to do what would make things go best, whatever the cost to ourselves. If I could save either my own life, or the lives of two strangers who were relevantly like me, it would be wrong for me to save my own life. If such acts would be wrong, it would be more plausible to claim that we can 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 this way much less demanding, it is less plausible to claim that we can have sufficient or decisive reasons to act wrongly.

8 Moral Concepts

Before discussing which acts are wrong, we should briefly consider the concept *wrong*, or what is meant by the word 'wrong' and by other words with the same meaning. There are several versions of this concept, some of which refer to different kinds of wrongness.

Like the concept of *a reason*, and of the wide reason-implying *ought*, one version of the concept *wrong* is indefinable, in the sense that it cannot be helpfully explained in other terms. We can use this concept to define some other moral concepts. We can say that some act is

> *right,* or *morally permitted,* when this act would not be wrong,

and that some act is

> *our duty*, *morally required*, or what we *ought morally* to do, when it would be wrong for us *not* to act in this way.

We might instead define this version of the concept *wrong* by appealing to

a similar version of one of these other concepts. Some act would be wrong, we might say, when we have a duty not to act in this way. But, though we can explain how these concepts are related, they 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*'. [57]

These moral concepts 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 resentment or indignation',
>
> 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. When people call some act wrong, they might mean that this act is blameworthy because such acts are unjustifiable to others. Or they might mean that this act mustn't-be-done because such acts are forbidden by God.

Some writers use

> 'our duty' to mean 'what we have decisive reasons to do',

and use

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

But these *decisive-reason* senses are not worth using. They add nothing to our conceptual scheme, since we already have the concepts of what we have decisive reasons to do, and of what we ought rationally to do. These senses of 'duty' and 'wrong' are also misleading, since they are very different from other more familiar senses. We often believe that we have decisive reasons to act in some way, though we do not believe that this act is our duty. Consider next the claim that

> (A) we always have decisive reasons to do our duty.

It is of great importance whether (A) is true. But, if we used 'duty' in the decisive-reason sense, (A) would mean

> (B) we always have decisive reasons to do what we have decisive reasons to do.

This claim would be trivial. And, if Rational Egoists used this sense of 'duty', they would claim that

> (C) we always have a duty to do what would be best for ourselves.

But that is a misleading way to state this view. Rational Egoism is not a moral view, but an external rival to morality. On this view, we always have decisive reasons to do what would be best for ourselves, whether or not such acts would be our duty, or would be wrong. [58]

In the *impartial reason-involving* sense,

> 'ought' means '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. These senses of 'ought' and 'wrong' *are* worth using, since they add something to our conceptual scheme. As I have said, there are similar senses of 'good' and 'best'. According to Act Consequentialism, or

> AC: We ought to do what would make things go best.

If this claim used both 'ought' and 'best' in these reason-involving senses, it would mean

> (D) What we have the strongest impartial reasons to do is whatever would make things go in the way in which we all have the strongest impartial reasons to want things to go.

We can call this view *Impartial-Reason Act Consequentialism*. To express this sense of 'ought', we can use the phrase *ought-impartially*.

This sense of 'ought' differs significantly from more familiar moral senses. Sidgwick, for example, writes

> the good of any one individual is of no more importance, from the point of view. . . of the Universe, than the good of any other. .

> . And. . . as a rational being I am bound to aim at good generally. . . not merely at a particular part of it. . . I ought not to prefer my own lesser good to the greater good of another.[59]

Sidgwick does not mean that, if he promotes his own lesser good, he would be blameworthy. Nor does he mean that such an act would give him reasons to feel remorse or guilt, and give others reasons for resentment or indignation. He seems to mean that, from an impartial point of view, he has most reason to do what would do most good. This is what he *ought-impartially* to do.

This version of consequentialism might be better regarded, not as a moral view, but as being, like Rational Egoism, an external rival to morality. Given this view's impartiality, it is closer to morality. That makes it, in some ways, a more serious rival, since this view may be accepted by some people who reject Rational Egoism, because they believe their own well-being to be what Sidgwick calls a 'narrow', 'limited', and 'ignoble end'. [60]

(D) may seem to be a trivial claim, which is true by definition. 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-involving senses. On desire-based theories, and Rational Egoism, there are no such acts or outcomes, since there is nothing that we all have impartial reasons to want or prefer. And, even if (D) is true by definition, Impartial-Reason Consequentialists can make other, substantive claims. For example, they might claim that

> (E) what we ought-impartially to do is whatever would produce the most happiness. [61]

These consequentialists might believe that we all have strong reasons to do what we ought in this sense to do. This belief might often lead them to try to maximize happiness. And they might not have, or act upon, 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.

There is another way in which words like 'wrong', 'ought', and 'right' can have different senses. To give some rough definitions, some act of ours would be

> *right* in the *knowledge-supposing* sense when this act would be right if we knew all of the relevant facts,

> *right* in the *evidence-relative* sense when this act would be right if the facts were as the available evidence gives us sufficient reasons to believe them to be,

and

> *right* in the *belief-relative* sense when this act would be right if our beliefs were true. [62]

Suppose that, as your

> *Doctor*, I must choose between two ways of treating you. Given the available evidence, treatment A would be almost certain to save your life, and treatment B would be almost certain to kill you. But I have the unjustified belief that treatment A would kill you and that treatment B would save your life.

This story could continue in several ways. Suppose that, in

> *Case One*, because I want to save your life, I give you treatment B, thereby killing you.

In giving you a treatment that kills you, as it was almost certain to do, I am acting wrongly in the knowledge-supposing and evidence-relative senses. But, in the belief-relative sense, my act is right. Suppose instead that, in

> *Case Two,* because I hate you, and want to kill you, I give you treatment A, thereby saving your life.

In giving you a treatment that saves your life, as it was almost certain to do, I am acting rightly in the knowledge-supposing and evidence-relative senses. But, in the belief-relative sense, my act is wrong.

According to some writers, it is enough to use 'right' and 'wrong' in their evidence-relative senses. On this view, when I try to kill you in *Case Two*, I am not acting wrongly. If some believer in voodoo 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. Nor is it wrong for me, in *Case Two*, to give you a treatment that is almost certain to save your life.

It is *not* enough to make such claims. It is worth claiming that, in the evidence-relative sense, my act is right. I am doing what, given the evidence available, any good doctor ought to do. But we should also

claim that, in the belief-relative sense, my act is wrong. Since I am trying to kill you, my act makes me seriously to blame, giving me reasons for remorse and guilt, and giving others reasons for indignation. When our acts are wrong in the belief-relative sense, that makes them wrong in the blameworthiness and reactive-attitude senses.

We should also claim that, when I try to save your life in *Case One*, my act is in one sense right. 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. But I may have read these journals, and my fault may instead be only that I have failed to believe what the available evidence gave me reasons to believe. Failing to respond to epistemic reasons is not morally wrong.

Suppose next that, though treatment A nearly always succeeds, and treatment B nearly always fails, you are one of the rare exceptions. And suppose that, in

> *Case Three,* I give you treatment B, believing truly though unjustifiably that I shall thereby save your life,

and that, in

> *Case Four,* I give you treatment A, believing truly though unjustifiably that I shall thereby kill you.

In the evidence-relative sense, I act wrongly in *Case Three* when I intentionally save your life, and I do *not* act wrongly in *Case Four* when I intentionally kill you. As before, it is not enough to make these claims. It is true that, in *Case Four*, I have given you the treatment that, on the evidence available, was most likely to save your life. But I have also murdered you, and that makes my act in other senses wrong.

In these and many other cases, what we ought to do depends on the goodness of our act's effects. In such cases, some act would be

> *right* in the *knowledge-supposing* sense just when this act would in fact make things go best.

It may seem that

> (F) if we don't know all the relevant facts, we ought to *try* to do what would be right in the knowledge-supposing sense, by trying to do what would make things go best.

But this claim is often false. Consider

> *Mine Shafts:* A hundred miners are trapped underground, with flood waters rising. We know that all these miners are in one of two shafts, but we don't know which. There are three flood-gates which we could close. 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 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 knowledge-supposing sense, because our act would save all of the hundred miners. If we closed Gate 3, we would have *no* chance of doing what would be in this sense right. But, given our beliefs and the evidence available, it would be clearly wrong for us to try to act rightly in the knowledge-supposing sense. We ought to close Gate 3.

In such cases, we can roughly claim, some act would be

> *right* in the evidence- or belief-relative senses when it is true that, given the evidence or our beliefs, this act would make things go in the way that would be *expectably-best*.

In saying that some act would make things go 'expectably-best', we do *not* mean that we expect this act to make things go best. If we closed Gate 3, we would be certain not to make things go best, as we might have done by closing one of the other gates. When what matters is the number of lives that are saved, some possible act would make things go expectably-best if this act would save the greatest *expected* number of lives. These expected numbers are the actual numbers that each act might save, divided by the chances that these acts would save these numbers. If we closed either Gate 1 or Gate 2, the expected number of lives saved would be 100 divided by 2, or 50. If we closed Gate 3, this expected number would be 90, since we would be certain

to save 90 lives. In the same way, in other cases, the expected goodness of some act's effects would be the goodness of these effects divided by the chance that this act would have these effects.

When I claim that, in this example, we *ought* to close Gate 3, I am using 'ought' in the evidence-relative and belief-relative senses. These are the senses of 'ought', 'right', and 'wrong' that are most important both when we are trying to decide what we ought to do, and when we ask whether some past act makes us blameworthy, giving us or others reasons for remorse or indignation. When we ask these questions, it is irrelevant which acts are right in the knowledge-supposing sense. In that sense, according to Act Consequentialists, Hitler's mother's doctor ought to have killed Hitler just after he was born. This doctor could not possibly have known that fact, and if he had killed the newborn Hitler, as in this sense he *ought* to have done, he would have acted wrongly in any of the ordinary senses of 'wrong'.

The knowledge-supposing senses of 'ought', 'right', and 'wrong', though having little practical importance, have some theoretical importance. That is true whether or not the rightness of some act depends on the goodness of its effects. In both kinds of case, it is often worth asking which acts would be right if we knew all of the relevant facts. Though we also need to ask how we can best respond to risks, and to uncertainty, these questions are best discussed separately. So I shall often suppose that, in my imagined cases, everyone would know all of the relevant facts. In such cases, what is right in the knowledge-supposing sense is the same as what is right in the other two senses. In other cases, I shall often use 'best' to mean 'best or expectably-best'.

According to some writers, there is only a single moral sense of 'ought', 'right', 'wrong' and other such words. 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 one definable sense that can be plausibly claimed to be the only sense in which anyone uses the words 'morally wrong'.

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 definable senses might then be claimed to express, not the belief that certain acts are wrong, but certain other beliefs about such acts. The divine command sense might express the belief that acts are wrong when and because they are forbidden by God. The decisive-reason sense

might express the belief that we have decisive reasons to avoid acting wrongly. And the reactive-attitude sense might be claimed to appeal implicitly to the indefinable sense, because the attitudes of guilt, remorse, and indignation all involve the belief that some act is wrong.

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, since we wouldn't be discussing the same questions.

This argument is not, I believe, strong. Different people may use these words in different definable senses that partly overlap. That may be enough to make disagreement possible. Suppose 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 we use 'wrong' in such overlapping senses, that may *increase* our disagreements. In this example, 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. And 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' and 'ought' in such different, partly overlapping senses, they have reasons to move to other, thinner senses, which they can all use. That would make their disagreements clearer. In the example just given, 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' in different senses may *not* be disagreeing. On Sidgwick's view, as I have said, it would be wrong for me to save my life rather than the lives of two relevantly similar strangers. If Sidgwick were using 'wrong' 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 strangers, I would have no reason to feel remorse or guilt, and the strangers 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

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

and others claim that

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

If (G) uses 'ought' in the knowledge-supposing sense, and (H) uses 'ought' in the evidence-relative or belief-relative sense, these claims do not conflict, and we could accept them both.

We need not decide whether the various senses that I have described should be called different senses of 'wrong', which may refer to different kinds of wrongness. It is enough to distinguish these senses, and the concepts they express. We can then decide which of these concepts are most worth using. And it may be worth using several of these concepts. For example, it may be worth asking 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 deep and difficult questions about whether acts can have the properties to which these concepts seem to refer. It might be true, for example, that no acts have the indefinable property of being things that mustn't-be-done, but that some acts are blameworthy and unjustifiable to others. There also deep and difficult questions about how we are able to understand irreducibly normative concepts, and able to recognize irreducibly normative truths. In this book I shall say little about these *meta-ethical* questions. These questions will be easier to answer when we have made more progress in our moral thinking. As Rawls remarks, our moral theories 'are primitive, and have grave defects'. [63]

Rather than proposing a new moral theory, I shall try to develop and combine existing theories of three kinds. 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'. [64] I shall ask whether he succeeds.

CHAPTER 4 POSSIBLE CONSENT

9 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. [65]

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.[66]

Christine 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. [67]

Onora O'Neill similarly writes:

> if we coerce or deceive others, their dissent, and so their genuine consent, is in principle ruled out. [68]

Korsgaard concludes:

> According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrong-doing to others. [69]

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 objection, Korsgaard and O'Neill 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. [70]

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 told you the truth, you could reasonably complain with your dying breath that I ought to have saved your life by deceiving you. I could not plausibly 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 giving your consent, you could then deliberately forget our conversation, so that I could tell my life-saving lie. Since you would consent to my act 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 and O'Neill 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 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 they are acts whose nature makes consent impossible.

10 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. [71]

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'. [72] 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 consenting 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, *act-affecting* sense. For people to be able to give or refuse consent in this sense, they must have what Korsgaard calls 'power over the proceedings'. These people must know that we shall treat them in some way only *if* they consent. So (B) could be restated as

> *the Choice-Giving Principle*: It is wrong not to give other people the power to choose how we treat them. [73]

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 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. [74] We could not possibly give to more than one of these people the power to choose how we act. [75] 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 Kant's claims about consent are not restricted to these special cases.

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 possibly 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 do not mean that giving consent is not one of the choices that is open to them. They 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, if they knew the relevant facts, they could not *rationally* consent.

We can call this the *Principle of Possible Rational Consent*, or---as I shall say

for short---the *Consent Principle*. [76]

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. [77] Consider next Kant's remark that someone whom he is treating wrongly

> cannot possibly agree to my way of treating him, *and so himself contain the end of this action*. [78]

If Kant were claiming that we ought always to let other people choose how we treat them, he would have no reason to add the remark that, for our treatment of others to be justified, these people must be able to agree to our treatment of them, and so 'contain', or share, the end or aim with which we act. When we let other people choose how we 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 with some aim that other people could *rationally* share, so that they could rationally consent to our way of treating them. [79]

Kant's claims about consent give us an inspiring ideal of how, as rational beings, we ought all to be related to each other. We might be able, I believe, to 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, if they knew the facts, they could rationally consent. And, if that is possible, Kant may be right to claim that this is how everyone ought always to act.

11 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 rationally will that the agent's maxim be a universal law. [80]

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 rationally will to be universal laws.

By our *maxims* Kant means, roughly, our policies and underlying aims.

(F) is not, however, true. Suppose that, as your doctor, I ask you whether you consent to my giving you some medical treatment. To know whether you have sufficient reasons to consent, you might need to know whether I am a well-qualified and conscientious doctor, and what effects this and the other possible treatments would be likely to have. But you wouldn't need to know on which maxim or policy I would be acting, or whether you could rationally 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'. [81] Rawls takes this remark to imply that Kant's other principles 'cannot add to the content' of Kant's Formula of Universal Law. Rawls therefore proposes that we should try to interpret Kant's other principles in ways that make them contain no other ideas. [82]

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. [83] The truth is that, in the cascading fireworks of a mere thirty pages, Kant gives us more new and fruitful ideas than all the philosophers of several centuries. 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. [84]

Kant's Consent Principle is one example. It is surprising that this principle has been so little discussed.

According to this principle, more fully stated:

> 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.

For this principle to be acceptable, it must both be in itself plausible, and have plausible implications. This principle must not conflict too strongly with our intuitive moral beliefs, by requiring too many acts that seem to us to be clearly wrong, or condemning too many acts that seem to be clearly required. If this principle both implies and supports many of our 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 reasons and rationality. Since we are asking what this principle can achieve, we should appeal, not to Kant's assumptions, but to what seems to us to be the truest or best view.

To avoid complications, we should suppose that, in most of my examples, everyone would know the relevant facts. When people know these facts, they could rationally consent to some act if they would have sufficient reasons to consent. We have *sufficient* reasons to act in some way when these reasons are no weaker than any reasons that we may have to act in any other possible way.

If we assume either some desire-based 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. [85] These people are both strangers to me, and there are no other morally relevant differences between them.

It is clear that I ought to save Blue's life. But we can plausibly suppose that, if I saved Grey's leg, that would be much better for Grey and would much better fulfil Grey's present desires. On that assumption, according to both desire-based theories and Rational Egoism, Grey could not 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. [86] 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 why this principle has been so little discussed.

We ought I believe to accept some *wide value-based* theory. On such theories, when one possible choice would be impartially best, but some other choice would be 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 such a 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 treat 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. [87] 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 these people, I could use a lifeboat to save either White or the five. These people are all strangers to me, and they do not differ in any other morally relevant way.

In this case, as most of us would believe, I ought to save the five rather than White. If White could choose how I shall act, she would have sufficient reasons, I believe, to make either choice. She 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. That is not how we should interpret 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 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 others in the different outcomes that would be produced by this and the other possible acts. It makes a difference how great these benefits or burdens would be, to how many people they would come, and how badly off we and the other people are. And it may make a difference whether we or the others are responsible for the situation that we are in, as might be true, for example, when we have been reckless. There are also some 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. In many such cases, for example, we could not rationally consent to being deceived or coerced. 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 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. This fact 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. [88]

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 personal 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. Like the five, White would have sufficient reasons to consent to my failing to save her life so that, in the time available, I could save the five. If White would have such reasons, as I believe, that strongly supports the view that in other kinds of case, in which the stakes are lower, there would be at least one possible act to which everyone could rationally consent.

If there would always be such an act, we could argue:

> Whenever someone could not rationally consent to some act, there must be some facts that give this person decisive reasons to refuse consent to it. These facts provide moral objections to this act.
>
> These objections must be stronger than the objections to any possible act, or acts, to which everyone *could* rationally consent.
>
> Whenever there are 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.

12 A Superfluous Principle?

According to some writers, nothing is achieved by appealing to the possibility of rational consent. 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 underlying 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 acts of some 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, have to make some 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 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 acts. According to *Act Utilitarianism*, or

> *AU*: We ought always to do whatever would, on the whole, benefit people most, by producing 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 gives the true account both of what we ought morally to do,

and of why we ought to act in these ways, 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 are 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 would 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 gives us a strong reason to act in this way, and may be enough by itself to explain why the other possible acts would be wrong. It is also worth asking whether we could achieve Kant's ideal.

13 Actual Consent

It is often morally important whether people *actually* consent to being treated in some way, or whether, if they had the opportunity, they *would in fact* consent. We should not ignore these questions, by asking only whether people *could* rationally consent to being treated in some way. Some rapist might claim that his victim could have rationally consented to having sexual intercourse with him. That claim could not justify rape. Even if this man's victim could have rationally consented to his sexual act, she did not in fact consent. To explain the wrongness of such acts, we must appeal to some other principle.

According to

> *the Veto Principle*: It is wrong to treat people in any way to which they either do, or would, 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 do or would refuse consent. But someone might 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 or would 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. [89]

Should we accept (J)? Suppose we are trying to decide whether it would be wrong to treat someone in some way to which, while we are acting, she refuses her consent. We might ask whether this person could rationally consent to the act to which she in fact refuses consent. But that question may be confusing, since this person 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

> CP2: It is wrong to treat people in any way to which they could not have rationally given, in advance, their irreversible consent.

People's consent to some act is *irreversible* when they know that, if they later withdrew their consent, that would make no difference to how they would be treated.

This principle rightly condemns almost all cases of rape. People could seldom rationally give irreversible consent in advance to sexual acts to which, at the time of the acts, they would refuse consent. That would seldom be rational because the nature of most sexual acts is greatly affected by whether, at the time, the people involved consent. There are, however, many kinds of act to which people 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 it to be rational for us to give such consent, we might have to believe 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. That would give Grey a reason to give irreversible consent, by saying 'Go ahead and save Blue's life, even if I later change my mind'. We can next suppose 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.

When we apply this version of the Consent Principle, 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 irreversible consent, thereby restricting their future freedom. But, given the aims of this imagined thought-experiment, we can *suppose* that these people would have had such reasons. Our question can be whether, on that assumption, these people would have had sufficient reasons to give irreversible consent to their being later treated in some way without their actual consent.

Since people could rationally give such irreversible consent to being

later treated in such ways, we can reject the argument given above. The Consent Principle does not imply the Veto Principle, and avoids the objections to that principle.

14 Deontic Beliefs

The Consent Principle claims to describe only one of the ways in which our acts may be wrong. As I have said, many acts are wrong even though everyone could rationally consent to them.

Many such acts are wrong because some people do not in fact consent to them. 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 these ways, we mean in part that such acts would be wrong. 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. On this view, all such acts are wrong even when the people who are killed could have rationally given, in advance, their irreversible consent.

Another large 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 drive your car, live in your apartment, and eat what is in your kitchen, even though you could have rationally consented in advance to my doing all these things.

There might also be some ways of treating people that would be wrong even if these people have actually and rationally given their consent.

Many people have that view, for example, about voluntary euthanasia, and assisting suicide. And some kinds of act are wrong for reasons other than the ways in which they treat people, so that the question of consent does not arise. One example is cruelty to animals.

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. The wrongness of such acts might give some people sufficient reasons to give consent to some alternative. When we apply the Consent Principle to any choice between certain possible acts, we must therefore ask whether there are people who would have such reasons. So we must appeal to our beliefs about the wrongness of any of these acts. These beliefs I shall call *deontic*, and reasons 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 belief that certain acts are wrong, our moral reasoning would be circular, or question-begging. Such reasoning could not support our belief that these acts are wrong, or help to explain why these acts are wrong.

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

> (1) some act is wrong because someone could not rationally consent to it,

and that

> (2) 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.

There is, however, a different way 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 saving someone else's life.

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.

This principle may seem to fail in a lesser way, by mistakenly permitting this wrong act. But the Consent Principle does not claim that we act wrongly if *and only if* people could not rationally consent to our way of treating them. 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.

It may also seem an objection that, in arguing that the Consent Principle does not require me to save Blue's life in this way, we must appeal to our belief that this act is wrong. But this objection has no force. As I have just implied, the Consent Principle does not claim to be the only principle we need. And we should expect that, in some of the cases in which our acts would be very bad for certain people, these people could rationally consent to these acts only if and because any possible alternative would be wrong. It is not surprising that if, in *Means*, Blue would have a sufficient reason to consent to my letting her die, this reason would have to be provided by the wrongness of the only act with which I could save her life.

Similar claims apply to other cases. We are considering acts that are wrong, not even in part because some people could not rationally consent to them, but for other reasons. We can argue:

> The Consent Principle requires some act only when one or more people would not have sufficient reasons to consent to our failing to act in this way.
>
> If 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. [90]

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. [91]

When we apply the Consent Principle to acts that may be wrong for other reasons, we must appeal to our beliefs about whether these acts would be wrong. But that is no objection to this principle, and it implies only that this principle does not support these particular moral beliefs. Similar claims apply to other more familiar principles. In most cases, for example, if we promise to act in some way that would be wrong for other reasons, we would have no obligation to keep this promise. So, when we apply some principle about our obligations to keep promises, we must appeal to our beliefs about whether the promised acts would be, for such other reasons, wrong.

15 Extreme Demands

Suppose next that, in

> *Self*, it is I who is 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 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 be No. From Blue's point of view, *Self* may be relevantly like *Earthquake*. Blue may not have sufficient reasons to consent to my saving someone else's leg rather than Blue's life, whether this leg is Grey's or mine.

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 often either benefit ourselves or give some greater benefit to others. When the benefits to other people would be *much* greater, these people may not have sufficient reasons to consent to our failing to benefit them. Suppose that, in

> *Aid Agency,* I could either spend $200 on some evening's entertainment, or give this money to some efficient aid agency, such as *Oxfam*, which would use this money to save some poor person in a distant land from death, blindness, or some other great harm.

When applied to these two alternatives, the Consent Principle clearly implies that I ought to give this money to this aid agency. This is the only act to which this poor person would have sufficient reasons to consent. [92] 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 is no objection to the Consent Principle that, 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 theory makes similar 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. [93]

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. [94]

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 requires more than that.

If this principle is too demanding, it could be revised. We might claim

> CP3: 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. [95] 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.

Kant's Consent Principle is, I conclude, fairly successful. This principle may be too demanding, and there may be some other ways in which it should be revised. [96] 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. [97] 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 5 MERELY AS A MEANS

16 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. I might use you as a dictionary, by asking you how some word is spelt. Or I might 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. [98]

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,

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 way would

best achieve our aims.

Frances Kamm rejects this second definition. She 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 the slightest weight to the well-being of his slaves, perhaps by letting them rest in the hottest part of the day, that would have been enough to make it true that this man did not treat his slaves merely as a means. But slave-owners surely failed to meet Kant's requirement. [99]

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. [100]

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 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 if 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 my slave-owner never flogs his slaves because he believes that such acts would be wrong. But what would make such acts wrong, he believes, is not that he would be inflicting pain on his slaves, but that he would be giving himself sadistic pleasure. If that is why this man never flogs 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. [101] 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 flogging 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 1930's, 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. Even if these people acted wrongly, they did not treat my mother 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. [102]

For condition (2) to be met, it is not enough that we would be prepared to bear great burdens 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 this 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 which are the facts that would make such claims true.

On Kant's view, it is wrong to treat any rational being merely as a means. On a similar but wider view, it is wrong to treat any conscious or sentient 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 seems also to claim 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 man 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 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. [103]

We could similarly claim that, when this Egoist *keeps* some promise to

someone whose help he needs, he wants to make use of this other person merely as a means. Suppose next that this man saves some child from drowning, at some 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. [104]

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 claims imply that Green acts wrongly. Perhaps we can accept that conclusion. But we should not claim that, when my Egoist keeps his promises, or risks his life to save some drowning child, he acts wrongly. Our claim should be only that this man's acts do not have what Kant calls *moral worth*. [105]

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. [106]

In moving to this revised principle, we would be giving up the view that, if we treat someone merely as a means, that is 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 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 treat them only in ways to which they could rationally consent. [107]

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 into 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 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 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 very glad 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 would 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 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. There are also people who would 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. [108] 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 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 kill herself by redirecting the train, she could also rationally consent to my doing that. So the Consent Principle does not condemn this act.

Similar claims also 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 White would have no reason to prefer to die as a side-effect rather than as a means of saving the five. If anything, White would have more reason to prefer to die 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 doing that.

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, 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.

Given these facts, I might argue:

> According to the Consent Principle, we ought to treat people only in ways to which they could 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

> If 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 may 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. If this act is wrong, its wrongness must be explained in some other way.

17 *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 murder 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. When we defend ourselves from some attack, we are not *using* the attackers. We might 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. [109]

Turn next to the cases in which, when we harm people as a means, we *do* also treat these *people* as a means. That may be true, for example, when we use someone's body as an instrument, or tool. On what I am calling 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

> *Accident*: Some malfunctioning machine threatens to kill both you and your child. You cannot save your child's life except by using

> *Black's* body, without her consent, in a way that would crush one of her toes. If you caused Black to lose another toe, you would also 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. Your act harms Black, without her consent, as a means of achieving your aim. On the standard view, you are thereby 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 do not treat someone merely as a means if we let 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, and makes your act less effective in achieving your aims. That would not be true, in *Bridge*, if I killed White as a means of saving the five. 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,

> (D) 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 will be likely to 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 *Accident*, 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 malfunctioning machine 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 (D) 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 through some 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 this building, (D) 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 (D), 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. This act may be wrong. And, if it is, what makes it wrong may be that I would be killing White as a means of saving the five. But I would not be treating White *merely* as a means.

I have rejected the standard view about what is involved in treating people merely as a means. Some writers make other claims. 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. . . [110]

Korsgaard similarly writes:

> Coercion and deception are the two ways of using others as mere means. [111]

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 a 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 do not treat someone as a mere means if we sacrifice our life for this person's sake. [112]

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. . . [113]

This claim implies that whenever people tell us to do something---such as to write a student essay, 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'. [114] But, when poor people feed themselves with the scraps that others throw away, they do not treat these other people as a mere means.

Suppose next that, in

> *Bad Samaritan*, while driving across some desert, I see you lying injured by the road, needing help. I ignore you, and drive on.

According to some writers, Kant would claim that I am here treating you merely as a means. That claim would be false. In ignoring you, I am not using you in any way, so I cannot be merely using you.

These writers might reply that, when Kant uses the phrase 'merely as a means'---or, more accurately, its German equivalent---Kant does not use this phrase in its ordinary sense. Kant often uses words in special senses. When I drive past you, ignoring your need for help, it might be true that, in Kant's special intended sense, I am treating you merely as a means. O'Neill and Korsgaard might similarly claim that all deception and coercion does, in Kant's special sense, treat people merely as a means.

We can sometimes rationally use words in something other than their ordinary senses. For example, it is 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 write 'painful or unpleasant', and the distinction that we are ignoring seldom matters. 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. [115] 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. [116] That would make it trivial to claim that these chosen principles are true. [117]

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 sometimes uses this phrase in a special sense, [118] he also uses it, I believe, in the ordinary sense. It is not misleading to say that, on Kant's view, we must never treat people merely as a means. And this is the version of Kant's claim 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 mostly 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 say, 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.

18 The Harmful Means Principle

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 *treat* people in this way.

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. [119]

To meet such objections, as I have said, several writers revise Kant's claim. 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.

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 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. 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*. 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) our act is the least harmful way to achieve this aim,

and,

(2) given the goodness of this aim, the harm we cause 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. [120] 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 a million other people. But we may believe it to be wrong to kill one person as a means of saving only only five, or only fifty other people.

On what I have called the standard view, if we harm someone, without this person's consent, as a means of achieving some aim, we thereby treat this person merely as a means. As I have argued, that may not be true. When I break Brown's leg to stop him from murdering me, I am *harming* Brown as a means of defending myself. But I am not treating *Brown* as a means, so I cannot be treating Brown merely as a means.

Turn next to cases in which, when we impose harm on someone as a means, we *do* also treat this person as a means. When we ask whether such acts 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 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 *Accident*, 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 bruise someone's leg, causing her 'a mild, short-lasting pain'. [121] So we can suppose that, in

> *Second Accident*, my gangster cannot save his child's life except by bruising Black's leg, without her consent, causing Black 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 be both harming Black and treating Black merely as a means.

According to Kant's Formula of Humanity, which includes the Mere Means Principle, it is wrong to treat 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 *Accident*, by causing Black to lose one toe, my gangster would be morally permitted to save his child in the same way. [122]

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, when 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 6 RESPECT AND VALUE

19 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. [123]

This requirement to respect all persons is one of Kant's greatest contributions to our moral thinking. It does not, however, tell us how we ought to act.

Allen Wood suggests that

> (A) we must always treat people in ways that express respect for them. [124]

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. [125] But, on this suggestion, (A) would not help us to decide which acts are right, since we could not decide whether some act could be taken to express respect except by deciding whether this act was 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. [126] 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. But, 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. [127]

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. [128] Though (C) cannot directly solve all moral problems, this principle provides, Wood claims, 'the correct basis for deciding moral questions'. [129] To support this claim, Wood points out that in his longest book about morality, Kant often makes remarks that seem to appeal to (C). [130]

Kant's remarks do not, I believe, show (C) to be a useful principle. Some of these 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'. [131] Kant makes other claims that Wood rightly rejects. It would be wrong, Kant claims, for each of us to give ourselves sexual pleasure, or to hasten our deaths to avoid suffering, because such acts debase or defile humanity. [132] 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'. [133] 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'. [134]

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.

20 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 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'. [135] Wood however 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.' [136] 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'. [137]

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-involving* 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. [138]

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 covers states of affairs, processes, outcomes, and acts. 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 pleasure, 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

> *the Actualist View*: Possible acts and other events are good as ends when they have intrinsic properties or features which 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 are good as a means when our making them happen, or be actual, would be an effective way of achieving some end. [139]

Similar claims apply to events that are bad as ends, or bad as a means to some end. Possible events may be good as ends either for particular people or in the impartial reason-involving sense, or both. As well as having reasons to try to produce 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 at some time actual or real, being a part of the way things go.

Since the Actualist View 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. The Actualist View 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. [140]

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 are in themselves good, such as good people, books, jokes, or 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, tell good jokes, 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.'[141] 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, or some religious painting by refusing to use these things as a dishcloth, firewood, or 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, rather than throwing them into 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 others 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 to be

promoted, when we can respond to the value of such things by treating them in respectful ways, these acts may be good as ends, having the kind of value that is to be promoted. [142]

We can turn next to 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. [143]

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-involving senses. Such happy and well-lived lives are good as ends. We have reasons to protect the living of such lives, since that can help people to achieve these good ends.

On some views, human life has a different kind of value. Suppose that *Blue* has begun to die a slow, painful and undignified death, and she has nothing important left to do. Blue may have strong reasons to kill herself, and other people may have strong reasons to help her, if she needs help. Of those who appeal to the value of human life, some believe that such acts would be wrong. These people might agree that it would be both better for Blue, and impersonally better, if Blue died an earlier, natural death. That would be, in a different sense, a better end. But Blue ought not to kill herself, these people believe, and other people ought not to help her, since these 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 be both best for this person and what she chooses.

Scanlon rejects this view. We have reasons not to end some life, he writes, only 'as long as the person whose life it is has reason to go on living or wants to live'. [144] Scanlon here denies that a person's life has the kind of value that we ought to respect in ways that might 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. [145]

It is not human life but the people who live these lives who can best 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.

21 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-effected* 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* ends, of which his main examples are rational beings, or people. Kant's third kind of end are what he calls *ends-in-themselves*. Such things have what Kant calls *dignity*, which he defines as absolute, unconditional and incomparable value or worth. [146] 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, who are ends-in-themselves, and whose value is of the kind that is to be respected rather than promoted. Though Kant sometimes makes such remarks, this interpretation of Kant's view is, I believe, mistaken. There are several kinds of thing which Kant claims to have supreme value, and some of these things are ends-to-be-effected, which Kant claims 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. A good will can be taken to be either a mental state or disposition, or an activity which consists in good willing. [147] Regarded in either way, having a good will is something that, on Kant's view, we ought to try to achieve. In Kant's words, 'the true vocation of reason must be to produce a will that is good.'[148]

Another end-to-be-effected 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. [149]

A third such end is what Kant calls the *Greatest Good*. [150] 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. [151] Kant claims that 'we ought to try to promote' this end, and that 'reason. . . commands us to contribute everything possible to its production.'[152]

There may be a fourth such end. Kant calls rational beings 'something whose existence in itself has absolute worth'. [153] And he writes that, if there were no such beings, the Universe would be 'a mere waste, in vain, without a final purpose'. [154] These remarks suggest that, on Kant's view, the continued existence of rational beings is another end-to-be-effected with supreme value. [155]

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-effected. And their value is of a different kind. On Kant's view, as Wood and Herman claim, 'even the worst human beings have dignity',[156] and a person whose will is good 'is of no greater value' than someone with an ordinary or a bad will. [157] This part of Kant's view is, I believe, a profound truth. But the value of the worst human beings cannot be claimed to be 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 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, while the value of rational beings is not a kind of goodness, their use and development of their rationality 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'. [158] Wood even calls Kant's claim about rationality's value 'the most fundamental proposition in Kant's entire ethical theory'.[159]

Like having a good will, rationality is in part an end-to-be-effected, 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 'infinitely far above' a lower kind of value, which he calls *price*.[160] 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.[161] 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, 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 people from any amount of pain. That conclusion would be insane.[162]

To avoid this objection, we might claim 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 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.[163] So it would be wrong for us to damage our ability to play chess or solve crossword puzzles, even if these were the only ways of 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 morality and for 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. [164]

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 morality, 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', [165] 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. [166] 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'. [167] On Kant's view, 'past moral philosophy . . . mistakes the nature of the good'. [168]

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, by the Realm of Ends, and by Kant's Greatest Good, the possible state of affairs in which everyone would be virtuous and happy. On Kant's view, these are all ends-to-be-effected, 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-involving 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. [169] Kant here ignores our non-moral reasons to want to be happy. He also ignores our non-moral reasons to try to achieve various other good aims. 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 or

badness 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. [170] 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. That claim would be obviously false. Kant means that all rational beings have a kind of value that is to be respected rather than promoted, since their rationality makes them beings who ought to be treated only in certain ways. This value is a kind of *status*, or what Herman calls 'moral standing.' [171] 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 needs to 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 most 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 just 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 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 the 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.

We can now look more closely at some of Kant's claims about the relations between what is good and what is right, or what we ought morally to do.

CHAPTER 7 THE GREATEST GOOD

22 The Right and the Good

The Greatest Good, Kant claims, would be a world in which everyone was wholly virtuous, or morally good, and had all of the happiness that their virtue would make them deserve. [172] Kant also writes:

> Everyone ought to strive to promote the Greatest Good. [173]
>
> the moral law commands me to make the greatest possible good in a world the final object of all my conduct. [174]

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.[175]

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. [176] But this seems unlikely to be Kant's view. Everyone's happiness would 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 happiness, but some people did not have all of the happiness that they deserved. So we should 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.

We would be wholly virtuous if we had good wills and always did what we ought to do. Kant assumes 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. [177] We can 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. We can call moral theories *telic* if they claim that everyone ought morally to promote, or try to promote, one or more common ends or aims. According to one such theory, *Hedonistic Act Utilitarianism* or

> *HAU:* Everyone ought to produce, or try to produce, the greatest possible amount of happiness. [178]

This theory is wholly telic, since it makes no other claims about what we ought morally to do. Many other moral theories are partly telic. So are the overlapping sets of similar, untheoretical moral beliefs that most people hold, which I am calling 'common sense morality'. On such views, we ought all to try to achieve certain common aims, such as that of saving people from being killed by famines or plagues. But we ought also to have certain personal aims. For example, rather than having the common aims that promises be kept and children cared for, each of us ought to try to keep our own promises, and to care for our own children. A third group of theories are not telic at all. According to one such view, our only duties are to obey the Ten Commandments. These commandments do not give everyone any common aims. [179]

Some telic theories are *value-based*, in the sense that they appeal to claims about the reason-involving goodness of what they tell us to try to achieve. According to one such version of HAU, happiness is the only thing that is good as an end, and we ought to try to maximize

happiness because that is how we can make things go best in the impartial reason-involving sense. Other telic theories are not value-based. That is true, for example, of Bentham's version of HAU, since Bentham doesn't use the concept *good* in this or any other relevant sense. Wholly telic theories are often called 'consequentialist'. I shall use this word more narrowly, to cover only theories that are both wholly telic and value-based.

As well as making claims about what is good and what we ought morally to do, moral theories may make claims about certain concepts, or the meanings of certain words. It will be enough here to distinguish three views about how the concept *good* is related to the moral version of the concept *ought*. [180] According to some theories, the concept *good* is fundamental, and can be used to define the concept *ought morally*. On one such definition, when we say that we ought morally to do something, we mean that this is the act that would do the most good. [181] According to some other theories, it is the concept *ought morally* that is fundamental, and can be used to define the concept *good*. I shall soon discuss some such definitions. 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 morally* 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.

Kant claims that we must define *good* in terms of *ought*. In his words,

> the concepts of *good* and *evil* must not be determined before the moral law. . . but only after it. . . and by means of it. [182]

Surprisingly, Kant also writes:

> All imperatives are expressed by an 'ought'. . . and say that. . . some act would be good. [183]

Kant may here seem to be doing just what he claims that we must not do, by defining *ought* in terms of *good*. Similarly, Kant describes some acts as 'practically necessary, that is, good.' [184] But these remarks do not use 'good' in any of its ordinary senses. [185] In these and other passages, Kant does not distinguish between some act's being good and its 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*. Such senses of 'good' we can call *ought-based*.

Kant also claims

> K1: Good wills are supremely good. [186]

Kant is referring here to the kind of wills that people have when they do their duty because it is their duty. We can call such wills *dutiful*. When Kant claims that such wills are good, he may be using 'good' in another ought-based sense. In calling some outcome

> 'supremely good' or 'best' in what we can call the *required-aim* sense, we mean that we ought to try to produce this outcome.

If Kant is using 'supremely good' in this sense, K1 means

> K2: We ought to try to have dutiful wills.

Though Kant believes K2, this may not be all that K1 means. When Kant claims that dutiful wills are supremely good, he may be using 'good' in some sense that is not fully definable in terms of 'ought'. As some other passages suggest, Kant may mean that we ought to try to have dutiful wills *because* such wills are supremely good. In this respect Kant's moral theory may be, as Herman claims, an ethics of value. But Kant would not here be doing what he claims that we must not do, by deriving the content of the moral law from our 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*. [187] As we have seen, however, Kant himself describes an ideal world which he calls the Greatest Good, and he claims that everyone ought 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?

In considering this question, it will help to compare Kant's claim with the kind of theory that most clearly does what Kant condemns. According to one version of *Act Consequentialism*, or

> AC: Everyone ought always to try to produce the greatest amount of good.

As we have seen, Kant claims

> K3: Everyone ought always to strive to promote the Greatest Good.

Given the similarity between these claims, Kant's claim may seem to be another statement of Act Consequentialism.

This is not, I believe, the best way to interpret K3. Kant, I suggest, uses the phrase 'the Greatest Good' in the required-aim sense, to mean 'what we ought to strive to promote'. When Kant claims K3, he has already told us what he believes to *be* the Greatest Good. Kant's view could be more fully stated as

> K4: Everyone ought always to strive to promote a world of universal virtue and deserved happiness,

and this claim doesn't even use the word 'good'. In calling this world 'the Greatest Good' in the required-aim sense, Kant would not be *supporting* his belief that everyone ought to strive to promote this world. He would be merely saying in a different way that this world is the end or aim that everyone ought always to strive to promote. On this interpretation, Kant's claim would not be a version of Act Consequentialism, since it would not be value-based.

There could, however, be Act Consequentialists who both accepted K4 and believed that Kant's ideal world would be best in some sense of 'best' that is not ought-based, such as the impartial reason-involving sense. So, even if Kant's Formula of the Greatest Good is not consequentialist, this formula partly coincides with one version of Act Consequentialism.

There is another way in which Kant is not making the error of the Ancient Greeks, by deriving his beliefs about what we ought to do from his beliefs about what is good. Kant's Greatest Good consists in part of everyone's doing what they ought to do. It might be objected that Act Consequentialists could not make this claim; but, as I argue later, that is not true. [188]

23 Promoting the Good

Kant's Formula of the Greatest Good might be claimed to be the only principle we need, since we should always 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. [189]

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). [190]

In these and other passages, Kant assumes

> K5: 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 must assume

> K6: 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.

It may seem unlikely that Kant assumed K6. 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 it is by following such rules that everyone could best promote deserved happiness.

That is not, I believe, unlikely. We should not assume that earlier writers 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. For example, they would often deceive themselves in their own favour, as when they believe that their need for the property they 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 institutions, such as that 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, since 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. Sidgwick assumes that, though we ought generally to act on the rules of common sense morality, there would be some exceptions, since it would sometimes be sufficiently likely that we could best promote happiness by breaking such a rule. But these cases, he believes, are fairly rare. Moore makes the stronger claim that, in trying to do the most good, we ought *always* to try to follow these common sense rules. [191]

Sidgwick and Moore are both, in one sense, Act Consequentialists, since they both believe that acts are in one sense right only if they do the most good. But, when Act Consequentialists claim that we ought not to try directly to do the most good, since we can do more good by trying to follow certain other policies or rules, the resulting view is not straightforwardly Act Consequentialist.

Most of us use 'right' and 'wrong' only in what I have called the *evidence-relative* and *belief-relative* senses. Some Act Consequentialists use 'right' and 'wrong' only in the knowledge-supposing sense, since they call acts right only when these acts would in fact do the most good. But these people ought also to use these words in these other senses. And, if Act Consequentialists also used 'right' and 'wrong' in these other senses, their claims would be closer to the claims of non-consequentialists. For example, Moore might mean that, in the evidence-relative sense, it is always right to act on the rules of common

sense morality, since that is how we would be most likely to make things go best, or more precisely expectably-best.

There is another way in which some people have come to believe that we can best promote happiness by always acting on these common sense rules. When Sidgwick claims that we could sometimes produce more happiness by breaking one of these rules, he is assuming one view about how we ought to assess the effects of our acts. According to what we can call this

> *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 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 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. [192] 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 world 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 makes similar claims. 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'. [193] 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 consider only the effects of this particular act. Even when some act would, if 'considered in itself', have bad effects, it may be one of a set of acts, done by us at different times or done by different people, that would together produce the best effects. This will then be the act that we ought to regard as doing the most good.

If Act Utilitarians rejected the Marginalist View and accepted the Whole Scheme View, they might accept Hume's claim that we ought to follow certain strict rules, such as 'Never steal'. On this view, it is by following such rules that we could benefit people most.

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. [194] It is often assumed that, in claiming that we must never lie, Kant states a view that could not possibly be Act Utilitarian. 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. [195]

Kant here condemns all lies by appealing to the harm that these acts bring about. As before, these claims might be made by Act Utilitarians who accepted the Whole Scheme View.

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. . .[196]

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. [197]

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 the greatest happiness. This claim states one version of *Hedonistic Rule Utilitarianism*.

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

> K5: 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 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 do not yet know what is implied by some of Kant's other formulas. We have not yet considered Kant's other main formula, his Formula of Universal Law.

We could not fully answer Q1 until we have answered Q2. But we can now 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 his other formulas. That is not so. These claims add 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. [198] As Rawls points out, these claims about desert cannot be plausibly derived from, or supported by, Kant's other formulas. [199] 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. [200]

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

24 Free Will and Desert

[This section is still too rough to be worth sending you. It will have little connection with my other claims.]

CHAPTER 8 UNIVERSAL LAWS

25 The Impossibility Formula

The rightness of our acts, Kant claims, depends on our *maxims*, by which he usually means the *policies* on which we act. Some of Kant's examples are: ''Increase my wealth by every safe means', [201] 'Let no insult pass unavenged', [202] 'Make lying promises when that would benefit me', and 'Give no help to those who are in need'. [203]

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. [204]

This claim needs to be explained. In one passage, Kant refers to a maxim's being 'a universal permissive law'. [205] 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), nor would (A) be a useful claim. [206]

Some writers suggest that Kant means

> (B) It is wrong to act on any maxim that we could not all accept.

On this suggestion, Kant's formula would be unreliable. If (B) condemned acting on maxims 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 worlds in which everyone accepts bad maxims, such as 'Deceive, coerce, or injure others when that would benefit me'. Such worlds might be claimed to be psychologically impossible, since there are some good people who would be unable to accept these bad maxims. But, in the sense in which that is true, there are also some bad people who would be psychologically unable to accept some good maxims. So, if (B)

appealed to such psychological impossibility, this formula would mistakenly condemn acting on these good maxims. As these remarks suggest, (B) is also implausible. We have no reason to believe that whether maxims are good or bad 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 also be unreliable, condemning many permissible or morally required 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 that does not make caring for our children wrong. We might revise (C), so that (C) applied only to maxims on which it would be impossible for everyone to act even if everyone had both the opportunity and the abilities that such acts would need. But no maxims would fail this test. And (C) is also implausible. We have no reason to believe that it is only bad maxims on which it would be impossible for everyone to act.

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. [207]

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: 'Adopt an orphan', 'Become a doctor or a lawyer', '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. And, besides condemning many permissible or good acts, (D) is 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. Nor is it wrong to make attempts some of which will fail.

We have been asking what Kant means when he claims it to be wrong

to act on maxims that could not be universal laws. (A) to (D) are the most straightforward ways to interpret Kant's claim. But, as well as being either unhelpful or both unreliable and implausible, (A) to (D) are not claims to which, when he applies his formula, Kant 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 upon 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. [208]

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. [209] 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. [210]

Korsgaard similarly writes:

> lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive. . . [211]

For (F) to condemn some actual people's lies, however, it is not enough for (F) to condemn acting on the maxim 'Always lie'. On Kant's view, the wrongness of an act depends on the agent's actual maxim, not on other maxims that the agent might have had. If someone acts on the maxim 'Lie when that is the only way to prevent some murder', we could not justifiably condemn this act by condemning the maxim

'Always lie'.

No one acts on the maxim 'Always lie'. Many liars act on the maxim 'Lie when that would benefit me'. Kant's formula condemns acting on this maxim 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 our lying, since a lie would not be believed. 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. [212]

But, as before, no one acts on the maxim 'Always steal'. Many thieves act on the maxim 'Steal when that would benefit me'. If this maxim were universally accepted and acted upon, that would not produce a world in which theft could never achieve its aim. There would still be property, which would not always be successfully protected. Self-interested theft would sometimes succeed.

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'. [213] 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 person acts on the maxim 'Make lying promises

when that would benefit me'. [214] 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. [215]

In assessing this claim, as Rawls points out, 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.[216] Kant would be right to predict that, after such a period, no one would be able to benefit themselves by making any lying promise. If everyone accepted the lying promiser's maxim, and believed that such promises were permissible, the practice of morally motivated, trust-involving promises would have ceased to exist. [217]

Now that we have found one kind of act that Kant's actual Impossibility 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) seems more plausible, however, than it really is. That can be shown with some imaginary examples. 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 might help me to 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. If these claims had been true, (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. [218] But such lies would be clearly justified. And, when applied to this example, (G) has no plausibility. It would be no objection to this way of saving people's lives that, if everyone believed such acts to be permissible, that would make them impossible.

This imagined case is intentionally similar to Kant's case of a lying promise. Lying promisers achieve their aims 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 make it impossible for anyone to act successfully on this lying promiser's maxim. Similar claims apply to my imagined case. My imagined helper of Jews achieves her aim because there are many people who can be trusted not to lie to the police, given their belief that such lies are wrong. I have similarly 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. The most important difference between these cases is in what these people's acts are intended to achieve; and this difference is ignored by (G).

For another example, we can suppose that most German soldiers of this period could be relied upon to obey orders, because they believed that disobedience is always wrong. That might have allowed some soldier to act successfully on the maxim 'Disobey orders when that would help any Jew to escape.' We can next suppose that, if all these soldiers had been known to believe that such disobedience would not be wrong, their officers would not have given any order whose being disobeyed would allow Jews to escape. On these assumptions, (G) would have mistakenly condemned my imagined soldier's life-saving act.

This example can illustrate another point. These German soldiers were required to swear oaths of unconditional obedience to Hitler. Kant condemns lying promises on the ground that, if everyone believed that they were permitted to make promises that they did not intend to keep, such promises would be a 'vain pretense', or sham. We could similarly claim that, if all German soldiers had believed that they ought to disobey Hitler if they were commanded to act wrongly, such oaths of

unconditional obedience would be a vain pretense or sham. If Kant's claim showed it to be wrong to make lying promises, our claim would show it to be wrong to ignore oaths of unconditional obedience when we are commanded to act wrongly. But, as Kant says, everyone ought to disobey such commands. Kant's formula ignores the difference between good practices, such as trust-involving promises, and bad practices, such as oaths of unconditional obedience. So, as well as condemning lying promises, this formula mistakenly condemns these refusals to obey immoral commands.

As these cases show, (G) is unacceptable. [219] This formula condemns some acts that are clearly right. And though (G) may condemn some wrong acts, it would condemn these acts for a bad reason.

Kant's actual Impossibility 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 some 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 before long make it impossible for anyone to act successfully on these maxims, since no one would offer bribes, and there would cease to be any poor people. Given these facts, these objectors claim, Kant's formula mistakenly implies that it is wrong to refuse bribes, or 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, Korsgaard claims, because these people's acts would thereby *achieve* their aim. [220]

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 by themselves and others for their generosity. If all rich people acted on this maxim, that might abolish poverty, making it impossible for any of these people later to act on

their maxim in a way that would achieve their aim. If that were true, (H) would condemn these people's acts. But when these people give generously to the poor, though their acts have no moral worth, they are not acting wrongly.

Consider next those men who accepted codes of honour, like the code that led the 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 defeated Pushkin's 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 clearly 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 it 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.

Turn now to the maxims 'Never 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'. If we all acted on these maxims, that might prevent any of us from achieving our aims. Cakes would never get eaten, conversations would never get started, and journeys would never end. That does not show that acting on these maxims is wrong. For a more serious example, consider the maxim, 'Have no children, so as to have more time and energy to work for the future of humanity.' If we all acted on this maxim, that would make it impossible for anyone successfully to act upon it, since humanity would have no future. So (H) condemns this maxim, in a way that is clearly mistaken.

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. [221]

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.' [222] But this claim is false. Of those who have been deceived or coerced, most can deceive or coerce others. O'Neill also claims that, when we deceive or coerce people, we 'undercut their agency', thereby preventing them 'for at least some time' from acting successfully in the same way as us. [223] But this claim is too strong. 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.

O'Neill might reply that, to show that (I) condemns deception and coercion, it is enough to defend the weaker claim that some deceivers and coercers prevent *some* of their victims from deceiving or coercing others. This claim is true. Similarly, if we acted on what O'Neill calls 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 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 permissible acts. There are many good or permissible maxims of which it is true that, if some of us successfully acted on them, that would prevent some other people from doing the same. To start with a trivial example, (I) would condemn playing competitive games with the overriding aim of winning. [224] Perhaps we could accept that conclusion. But there is nothing wrong with acting on the maxim 'Become a lawyer', even when, by taking the last available place in some Law School, we make it impossible for someone else to become a lawyer. 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 act successfully on these maxims. 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 formula. What this formula forbids, she suggests, are acts whose success 'depends upon their being exceptional.' This test, she adds, 'reveals unfairness'. [225] 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 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, [226] pay the debts on our credit cards before interest is charged, [227] buy only second-hand books, or give surprise parties. [228]

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 acts are wrong if their maxims could not even *be* universal laws.

26 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 for us to act on any maxim that we could not rationally *will* to be a universal law. [229]

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 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 fill this gap, by implying that some kinds of act are always wrong. As Herman remarks, it would not be enough if Kant's view implied that, though it would be wrong to have a *policy* of killing others for our own convenience, such acts are sometimes permitted. [230]

When we apply Kant's formula, we must imagine or suppose that we have the power to choose or will that certain things be true. Kant sometimes claims that, when we apply his formula, we should ask whether we could rationally will it to be true that our maxim be a universal law of nature, in the sense that everyone would accept this maxim, and act upon it when they can. [231] 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. Thus the maxim 'Give generously to the poor' applies only to the rich.

In 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. [232]

When Kant applies this formula, he assumes that, if we were permitted to act on some maxim, we would be more likely to act upon it. This effect would be produced, not by our *being* permitted to act on this maxim, but by our *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. [233]

Unlike the Permissibility Formula, this formula can be used on its own. And this formula is in other ways better. [234] 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'. [235] 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 our main aim, I shall assume, is to find out whether a Kantian moral theory can help us to decide which acts are wrong, and help to explain why these acts are wrong. So, in asking what Kant's formulas imply, 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.

Combining these claims 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 we and others act wrongly. Kant knows that, if he appealed to such beliefs, his formulas would achieve nothing, since they could not help us to reach true beliefs about which acts are wrong, nor could they 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 a world in which everyone *believes* some kind of act to be wrong, we should not appeal to our beliefs about whether such acts *are* wrong. As before, Kant 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 greater burdens on others. On some views, such acts are not rational, since everyone is rationally required to give equal weight to everyone's well-being, or at least 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 promote everyone's well-being, they could not even rationally will 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 claim that this man is rationally required to give such help. As Rawls and Herman suggest, when we apply Kant's formulas to people who benefit themselves at greater costs to others, we should suppose that these people's maxims and acts are both rational. [236]

27 The Agent's Maxim

Whether some act is wrong, Kant's formulas assume, depends on the agent's maxim. Kant sometimes uses the word 'maxim' to refer only to the policy on which someone acts. In some other passages, Kant takes a maxim to consist both of someone's policy and of this person's underlying aim. To adapt one of Kant's examples, we can suppose that two merchants both act on the policy 'Never cheat my customers'. One merchant never cheats because he believes this to be his duty, while the other's motive is to preserve his reputation and his profits. These merchants, we might say, act on the same *policy maxim* but on different *policy-and-motive maxims*.

Kant's appeal to the agent's maxim raises various problems. 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 it to be true that everyone acts on this maxim whenever they can, since I would know it to be most unlikely that anyone else could ever act in precisely this way. So Kant's Law of Nature Formula mistakenly permits my act. [237] Similar claims apply to other cases. Whenever wrong-doers act on highly specific maxims, they could rationally will it to be true that everyone acts on their maxims, since they would know that other such acts would be either impossible or very rare. Kant's formula mistakenly permits these wrong acts. We can call this the *Rarity Objection*.

This objection can be partly answered. It is a factual matter what the maxim is on which someone is acting. And real people seldom act on such highly specific maxims. When describing someone's maxim, as

O'Neill claims, we should not include any details whose absence would have made no difference to this person's decision to do whatever she is doing. [238] In a realistic version of my example, I would have stolen from my victim even if her wallet had contained only $62, or 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 might *not* be a maxim on which I could rationally will everyone to act, and Kant's formula would then condemn my act.

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 formula ought to be able to condemn these imagined people's acts. [239] And, as we shall see, this objection applies to some actual cases.

Kant's appeal to the agent's maxim raises other problems. Return first to my imagined Egoist, who has only one policy and underlying aim, '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 morally permitted. Such a world would be much worse for him. Egoists have strong self-interested reasons to want other people to accept and follow, not their egoistic maxim, but various moral principles. Since my 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 tells the truth, keeps his promises, and helps other people. These are unacceptable conclusions. When my Egoist saves some child from drowning because he hopes to get some reward, his act has no moral worth, but he is not acting wrongly.

It might be claimed that, when my Egoist saves this child, *what* he is doing is not wrong, but *his doing* of it is. Kant suggests this 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. [240] Kant similarly claims that to fulfil a duty of gratitude, we must feel grateful. [241]

This distinction cannot answer this objection to Kant's formulas. We may have some duties of virtue, which we can fulfil only by acting with the right motive. My Egoist could not fulfil such duties, 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. Our problem is that, when my Egoist pays his debts, Kant's formulas mistakenly imply that this man *is* acting wrongly, since he is acting on a maxim that he could not rationally will to be universal.

Return now to the drowning child. Suppose that, given the strength of the river's current and the nearness of some 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, she would be nobly acting beyond the call of duty. My Egoist may think 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 my Egoist brushes his teeth, he is 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 this man's act is not wrong, his doing of it is. There is no sense in which this man's brushing of his teeth 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'. [242] And, if we claimed that such maxims are too general, we would be ignoring some people's actual maxims. Kant discusses the maxim 'Make a lying promise when that would benefit me'. There are other, similar maxims, such 'Steal, cheat, and break the law when that would benefit me'. But, since these maxims all have the same underlying self-benefiting aim, they are unnecessary clutter, and could all be replaced by the single maxim 'Do whatever would be best for me'. It is undeniable, I believe, that there are people who are like my imagined Egoist. These people act intentionally in certain ways, and they have a single underlying policy and aim. It would 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 can be Kant himself, during the period in which he accepted the maxim 'Never lie' because he believed that lying is always wrong. 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. So Kant's formula implies that, whenever Kant acted on his maxim by telling anyone the truth, Kant 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, even when these are the only ways 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, in the sense that it is always wrong to act upon them. One example is a sadist's maxim 'Torture others for my own amusement'. When applied to some such maxims, Kant's formulas succeed. But there are many maxims that are, in this sense, only partly bad. When people act on these maxims, they sometimes act wrongly, but at other times their acts are permissible or even morally required. That is true of the Egoist's maxim 'Do whatever would be best for me' and Kant's maxim 'Never lie'. Kant overlooks such maxims. In proposing his Law of Nature Formula, Kant assumes that, if we could not rationally will it to be true that everyone always acts on some maxim, we ought to conclude that no one should ever act upon it. As our examples show, that assumption is a mistake. Since such maxims are only partly bad, we can call this the *Flawed Maxims Objection*.

There are, we shall see, other objections to Kant's Formula of Universal Law, in both its law of nature and moral belief versions. After considering some of these objections, some writers conclude that Kant's formula cannot be used to decide which acts are wrong. Wood claims that, when used as such a criterion, Kant's formula is 'pretty worthless'.[243] Herman claims that, despite a 'sad history of attempts. . . no one has been able to make it work'.[244] O'Neill suggests that, when applied to people who act on certain kinds of maxim, Kant's formula cannot tell us whether these people's acts are right, and may 'give either unacceptable guidance or none at all'.[245] Hill doubts whether, when

used on its own, Kant's formula can provide 'even a loose and partial action guide'. [246]

Given their belief that Kant's formula cannot provide a criterion of wrongness, some of these writers 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 doing some kinds of act for certain reasons. [247] 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. [248]

Kant, I believe, had more ambitious aims. For example, 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? [249]
>
> 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. [250]

Kant also claims that his formula 'determines quite precisely what is to be done. . . with respect to all duty in general'. [251]

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.

To judge whether some act is, or would be, wrong, we need to know what the agent is, or would be, *intentionally doing*. We must know this person's immediate aims, or what she is directly trying to achieve. We must also know what effects the agent believes that her acts might have. What people intentionally do is not the same as what they intend. If some terrorist shoots down the airplane in which his country's President is travelling, he may be intending only to kill the President, but what he is intentionally doing is shooting down this airplane knowing that he will thereby kill many other people.

Some of Kant's maxims merely describe what someone would be intentionally doing. That is true of the maxim 'Kill myself to avoid suffering'. This maxim could be acted on only once. [252] In most

cases, however, Kant uses 'maxim' to refer either to someone's policy, or to some policy and underlying aim. As I have just argued, in judging whether some act was or would be wrong, we don't need to know on which *policy* the agent was or would be acting. When Kant told people the truth, it was irrelevant that he was acting on the policy 'Never lie'. And, when my Egoist saves the drowning child's life, it is irrelevant that he is acting on the policy of doing whatever would be best for himself. These facts at most give us reasons to believe that, in some *other* cases, Kant or this Egoist might or would act wrongly.

Since an act's wrongness does not depend on the agent's policy, Kant's formulas need to be revised. According to some writers, Kant's Law of Nature Formula could become

> LN2: We act wrongly unless what we are intentionally doing is something that we could have done while acting on some maxim on which we could rationally will everyone to act. [253]

This formula avoids the Flawed Maxims Objection. When my Egoist saves the child's life, 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 $63 from a white-dress-wearing strawberry-eating reader of Spinoza's *Ethics*. What I am intentionally doing is something that I could have done while acting on a maxim, or policy, of acting in precisely this way, whenever I can. I could rationally will it to be true that everyone acts on this maxim whenever they can, since it is most unlikely that anyone else could ever do that. So LN2 mistakenly permits my act. [254] Similar claims apply to countless other cases. If our formula applies to all of the maxims on which someone *might* have acted, there will nearly always be some such possible maxim that the agent could have rationally willed to be universal. So this formula would fail to condemn most wrong acts.

To avoid these objections, we should revise Kant's formulas in another way. These formulas should not use the concept of a maxim, in the sense that covers policies. Kant's Law of Nature Formula could become

> LN3: We act wrongly unless what we are intentionally doing is something that we could rationally will everyone to do.

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 permitted.

The phrase 'such acts' here refers to what we are intentionally doing. As we shall see, these formulas need to be revised in some other ways.

Like LN2, these formulas avoid the Flawed Maxims Objection. When my Egoist saves some child's life, and Kant tells most people the truth, they could rationally will it to be true both that everyone does what they are intentionally doing, and that everyone believes such acts to be permitted. So these formulas do not mistakenly condemn these acts. And, since these formulas do not appeal to merely possible maxims, they avoid the Rarity Objection. In describing what someone is intentionally doing, we should not include morally irrelevant details. Suppose that, being a whimsical kleptomaniac, I really *am* acting on the maxim of stealing only from a strawberry-eating woman who is reading Spinoza's *Ethics*. I wouldn't steal from anyone who was eating any other fruit, or reading any other book. Though my actual maxim is highly specific, what I am intentionally doing is, in part, stealing money from someone no richer than me, to benefit myself. I may not be able rationally to will it to be true either that everyone acts in this way, whenever they can, or that everyone believes such acts to be morally permitted. These revised formulas would then condemn my act.

It is sometimes unclear what someone should be counted as intentionally doing. It may be unclear, for example, how much should be included in the agent's aim, or among some act's foreseen or foreseeable effects, or what should be regarded as separate acts or as parts of a single complex act. But, when we are trying to decide whether some act or kind of act was or would be wrong, these are the questions that we need to answer.

It might now be claimed that, if we revise Kant's formulas so that they do not refer to maxims in the sense that covers policies, we are no longer discussing Kant's view. That is true, but no objection. We are asking whether Kant's ideas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise Kant's formulas in a way that improves them, we are developing a Kantian moral theory. And the policy-covering 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 suggests, is not 'a detachable or dispensable part of Kant's theory', since this feature of Kant's view enables us to claim that, when someone tries to universalize some bad maxim, there is a contradiction in this person's will. We can thereby argue that wrong-doing involves 'failures to have coherent intentions'. [255] But, as Kant points out, wrong-doers do not in fact try to universalize their maxims, so 'there is really no contradiction' in these people's wills. [256]

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. [257] But, as we have seen and O'Neill elsewhere claims, [258] that is not so. If all we know is that my Egoist has acted on the maxim 'Do whatever would be best for me', we cannot possibly decide whether this man has acted wrongly. We don't know whether he has broken or kept some promise, killed someone, or saved someone's life. And, if all we know is that Kant has acted on his maxim '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.

Return next 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. His motive is always to benefit himself.

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 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. [259] Similar claims apply to many other maxims, such as 'Never steal' and 'Never break the law'. 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 9 WHAT IF EVERYONE DID THAT?

28 Each-We Dilemmas

Though I have claimed that we ought to revise Kant's formulas, it will be clearer to go on discussing Kant's own formulas, returning to our revisions when that is needed. It is worth showing that there are other ways in which these formulas need to be revised. And some of my claims will also apply to our revised formulas.

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. This question is incomplete, since the answer may depend on what the alternative would be. One alternative might be the actual world. But Kant's formula would then permit us to act on many bad maxims. If many people were already acting on these maxims, it might make little difference if these maxims 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*. [260]

To apply this version of Kant's formula, we also need to know what people would do if no one acted on some maxim. 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 even worse maxim. On the best version of Kant's formula, we could not rationally will it to be true that everyone acts on some maxim if there is some other *better* alternative maxim on which we could rationally will everyone to act.

Kant's Law of Nature Formula works best when it is applied to maxims or acts of which three things are true:

> (1) it would be possible for many people to act on this maxim, or in this way,

> (2) whatever the number of people who act in this way, the effects of each such act would be similar,
>
> (3) 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 *better* in some kind of way, *we* would be doing what would be, in this way, *worse*.

We can call such cases *each-we dilemmas*.

It will be enough to consider cases in which each person's act would have effects on people's well-being. One large class of each-we dilemmas are often misleadingly called *prisoner's dilemmas*. [261] There are many groups of people of whom it is true that

> each person could either (A) benefit herself or (B) give some greater benefit to others,
>
> these greater benefits would be roughly equally distributed,

and

> what each person does would have no significant effects on what others do.

If each person does (A) rather than (B), what she is doing is certain to be better for her, whatever the other people do. But if all rather than none of us act in this way, what *we* are doing 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. [262]

Though such cases 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. Some of these cases we can call *Samaritan's dilemmas*. Each of us can sometimes help some stranger who is in need, at some small but real cost or burden to ourselves. [263] One example would be helping

someone on the street who has had some accident. If all of us always gave such help to strangers who are in need, that might be better for all of us than if none of us ever gave such help. But, if we live in some large community, such as London, Los Angeles, or Beijing, 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, in a large community, whether this person received such help from others would very seldom depend on whether she gave such help. The strangers whom she failed to help would hardly ever be the same people as the strangers by whom she could later be helped. So this person's failure to help others would hardly ever lead others, bearing a grudge, to deny this person help. As I have said, however, if no one helps strangers when that would involve a real cost, though each of us is doing what is better for herself, we are doing what is worse for all of us.

Some each-we dilemmas 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 involve clean air, national defence, and law and order. [264] 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 the avoidance of certain bad outcomes, and the contributions that are needed are 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 decline, so that, before long, they 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. For example, there might be some fishermen who are so skilful that, even if there was overfishing, they would still catch as many fish. When that is true, however, the other fishermen may still face an each-we dilemma. In my description of these cases 'everyone' means 'all the members of some group'. It would be no objection to my claims if there are some people who, though acting in the same ways, are not members of this group.

Many each-we dilemmas do not involve, or involve only, benefits to ourselves. 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 her 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 benefit those people to whom we are related in certain ways. These are people such as our children, parents, pupils, patients, clients, customers, colleagues, or those whom we represent. We can call these our *M-related people*. If we all believe that 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 may 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, by taxation that is either unavoidable, or enforced by penalties for non-payment. In many other cases, however, political solutions cannot be achieved, or are too costly. In some of these cases, we have achieved solutions that are *psychological*, in the sense that, without a change in our situation, all or most of us choose to give the greater benefits to others. Such solutions often depend on our having and acting upon certain moral beliefs. We may contribute to some public goods, despite the costs to ourselves, because we believe that we ought to contribute.

Of these *moral* solutions to each-we dilemmas, two are especially relevant here. Suppose that we were all Act Utilitarians, who believed that we ought always to do whatever would produce the greatest sum of benefits. If we all acted on this moral belief, we would all contribute to such public goods, giving the greater benefits to others in ways that would make the outcome better for all of us. These solutions are rare, however, since there are few people who are both Act Utilitarians 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 a maxim of failing to contribute. So, if we were all conscientious Kantians who always acted on Kant's Law of Nature Formula, we would all contribute to these public goods.

When we have achieved some moral solution to some contributor's dilemma, common sense morality requires everyone to go on contributing. In such cases, there are often some *free riders*: people who benefit from these public goods, without making any contribution. Each free rider benefits herself in a way that imposes a greater total cost on others. Common sense morality condemns such acts as unfair. And these are 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 contributes 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 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 could say, 'What if *no one* did that?' 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 things continue as they are, so that no one gets these goods.

Kant's formula is especially valuable here for at least two reasons. First, this formula conflicts with, and may lead us to revise, some widely held and at least partly mistaken moral beliefs. In unsolved each-we dilemmas, most of us believe that we are either permitted or required to give the lesser benefits to ourselves or 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 all of our M-related people. [265]

As well as conflicting with some widely held beliefs, Kant's formula challenges these beliefs in an especially forceful way. Though Act Utilitarians would also claim that everyone ought to give the greater benefits to others, the Kantian argument for this conclusion is more plausible, and 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, we are acting in ways that are *directly collectively self-defeating*. If we were egoists, that would be no objection to our view, since Rational Egoism is a theory about *individual* rationality and reasons. But moral principles or theories answer questions about what *all* of us ought to do. So such principles or theories clearly fail when they are directly self-defeating at the collective level.

Kant comes close to giving such an argument. When Kant discusses the limits on our duty to benefit others, he writes,

> a maxim of promoting the happiness of others with a sacrifice of one's own happiness. . . would conflict with itself if it were made into a universal law. [266]

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. Given some further assumptions, 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. Given 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. [267]

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 people that would be trivial, and may even be imperceptible.

These claims apply to most of the cases that I listed above. When we know that our acts would impose only such trivial or imperceptible burdens on each of many people, our ordinary concern for others

would not be aroused. Even if we were Act Utilitarians, 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 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 that we together inflict such damage on ourselves or our 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 to ourselves without receiving the greater benefits from others. This requirement may sometimes be too demanding. [268] 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 act in partly similar ways. [269]

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

29 The Permissible Acts 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 it is wrong to act on some maxim may depend, however, on how many people act upon it. When applied to such maxims 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 those acts whose rightness depends at least in part on their predictable effects. There are many maxims of which it is true that

> (1) 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.

If such effects are sufficiently important, it may be true that

> (2) 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,

> (3) most of us could not rationally will it to be true that everyone acts on these maxims.

When applied to these maxims, Kant's formula implies that most of these acts are wrong. So this formula mistakenly condemns such acts even when they are either permissible or 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. [270] Consider next the maxims: 'Consume food without producing any', 'Become a dentist', and 'Live in Iceland, to absorb the spirit of the Sagas'. [271] It is not wrong, in the world as it is, to act on these maxims. But, since most of us could not rationally will it to be true that everyone acts on these maxims, Kant's formula seems to imply that most of these acts are wrong. We can call this the *Permissible Acts Objection*.

Thomas Pogge suggests that, to answer this objection to Kant's view, we should turn from Kant's Law of Nature Formula to his Moral Belief Formula. [272] Though most of us could not rationally will a world in which everyone *acts* on such maxims, we could rationally will a world in which everyone believes such acts to be 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. 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*, since they are maxims of acting in some way only when our acts would have certain effects. Such maxims would not apply when our acts would not have the effects that we intend, 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 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.' As Pogge claims, 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 here makes 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. We might here 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 Permissible Acts 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. [273] 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 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 a world in which everyone acted on these unconditional maxims, or believed 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. Perhaps we ought not to accept these unconditional maxims, but it would not be wrong to act on these maxims in the world as it is.

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 *flawed 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 covers policies, but to what the agent is intentionally doing. On our revised version of Kant's Law of Nature Formula,

> LN3: We act wrongly unless what we are intentionally doing is something that we could rationally will everyone to do.

If we acted on these unconditional maxims, what we would be intentionally doing 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. So LN3 does not mistakenly imply that these acts would be wrong.

30 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

(4) 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.

When such effects are sufficiently important, it may be true that

> (5) 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 is true, Kant's formula requires us never to use violence.

Pacifism has considerable intuitive appeal. And many people (one of them my father) have been pacifists on Kantian grounds. But, like Kant's belief that we must never lie, pacifism is too simple. Return to the time of the Second World War. If everyone outside Germany had been pacifists, that would have allowed Hitler to dominate the world, with effects that would have been likely to be even worse than this terrible war. If Kant's Law of Nature Formula implied that it was wrong to fight against Hitler's armies, that would count against this formula.

Suppose next that, in

> *Mistake*, several people's lives are in danger. You and I must choose between two ways of acting. The possible outcomes are these:

		I: do A	I: do B
You	do A	we save everyone	we save no one
	do B	we save no one	we save some people

We ought both to do A, since that is our only way to save everyone. But suppose that, because you misunderstand our situation, you do B. Despite knowing that you have made this mistake, I do A, with the result that we save no one. I know that, by doing A, I shall prevent us from saving some people whom we would have saved if I had done B. But, as a Kantian, I believe that I ought to do A, since that is the only thing that I could rationally will us both to do. [274]

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. The standard of conduct Kant sets for us, 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. [275]

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. [276]

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. [277] Consider this maxim:

> M1: Do whatever I could rationally will everyone to do.

According to the Ideal World Objection, compared with a world in which everyone acts on M1, we could not rationally will a world in which no one does. If this claim were true, Kant's formula would require us to act on M1 even when 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 don't act in this way, in which case do whatever, given the acts of these other people, 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 have failed to act 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 necessary 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. As Kant seems to have assumed, his formula 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 mistakes or wrong acts of other people.

The Ideal World Objection can, however, take another form. Kant's formula merely *permits* us to act on these better maxims. Consider another maxim:

> Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can.

Everyone could rationally will it to be true that everyone acts on this maxim, since that would *also* produce a world in which no one ever uses violence. But in the real world there will always be some people who use aggressive violence. Since this maxim passes Kant's test, Kant's formula mistakenly implies that the rest of us are morally permitted 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. [278] In the real world, however, there will always be some people who don't act 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 illustrate this problem in its clearest form, we can turn to

> M3: Do whatever everyone could rationally will everyone to do, unless some other people haven't acted in these ways, in which case do whatever I like.

Everyone could rationally will it to be true that everyone acts on this maxim, since that would produce a world in which everyone does only what everyone could rationally will everyone to do. So Kant's formula permits everyone to act on this maxim. 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 original Ideal World Objection, Kant's formula

sometimes requires us to act as if we were living 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, if we appeal 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 may permit the rest of us to do whatever we like.

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.

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 it would make things go best. [279] So RC does *not* require us to follow those ideal rules whose being followed by only some people would have disastrous effects. In acting on R1, as RC permits us to do, we could respond in the best ways to the mistakes and wrong acts of others. But this objection can take another form. 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 must at least permit us to follow this rule. We know that, in the real world, some people have not followed 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. [280] Similar objections also apply to several 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.

As before, the implied comparison is with no one's acting on this maxim. 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 no one, but also by *most* people rather than no one, by *many* people rather than no one, by a *few* people rather than no one, and by *any other number* of people rather than by no one. We must be able rationally to will that, *whatever* the number of people who don't act on this maxim, *everyone else* does.

If we widen Kant's formula in this way, it condemns the bad maxims that we have discussed. One such maxim 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 this maxim be acted on by any other number of people. If anyone uses aggressive violence, thereby failing to act on this maxim, everyone else would act on this maxim by killing as many people as they can; and that is not something that we could rationally will.

When we consider many maxims and acts, this revision of Kant's formula would make no difference. There are many acts whose moral status does not depend on the number of people who act in this way. In such cases we could rationally will that any number of people act on certain unconditional maxims. 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 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 those below which acts of certain kinds cease to have certain good effects, or 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 would make things go best.

Some of these rules could take such conditional forms.

This revision makes Rule Consequentialism in some ways closer to Act Consequentialism. That is most importantly true when we ask what proportion of their wealth or income the world's rich people ought to give to those who are poor. When applied to this question, most versions of Rule Consequentialism make claims that 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 they ought to give, the best rule would require the others to give a great deal. [281]

In revising Kant's Law of Nature Formula so that it avoids the Ideal World Objection, we appeal to conditional maxims, and to whether we could rationally will that any number of people act on these maxims. We thereby depart from the idea expressed in the question 'What if everyone did that?' As I have argued, this idea can be successfully applied only to certain kinds of case. In each-we dilemmas, for

example, if we are free-riders who fail to contribute to some public good, we can be rightly challenged with the question 'What if everyone did that?' But in many other cases it is enough to reply 'Not everyone will'. [282]

Kant's Moral Belief Formula appeals to a different idea, which might be successfully applied to all kinds of case. Though we should not 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 to reply 'Not everyone will'. 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. [283]

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

CHAPTER 10 IMPARTIALITY

31 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. [284]

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, one that was independently asserted, and claimed to be fundamental, in at least three of the world's earliest civilisations. [285] Though Kant calls his own formula 'the supreme principle of morality', he dismisses the Golden Rule as 'trivial' and unfit to be a universal law. [286] 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 helped by others, if they could thereby be excused from helping 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. These people, Kant claims, would agree not to be helped if they could thereby be excused from helping others. To put this claim in Kantian terms, these people would will it to be true that the maxim of not helping 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 people to act on maxims that they 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, if 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.' [287] 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 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 its most common statements, 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 most of us do not have such power over others, we can only want or wish that others act in some way. Kant's formula asks us to imagine or suppose that we have the power to will, or choose, how others will act. 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 would be *willing* to be treated. Some familiar statements of the Golden Rule, such as 'Do as you *would* be done by', seem already to 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, this is clearly what Kant has in mind. 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 rationally choose that they 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, we should not merely 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 should ask whether we would rationally choose it to be true that, in some imagined case in which we ourselves were starving, other people give us no food.

Consider next some white racist who, in the period of racial segregation 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 affect, by having these people's desires, attitudes, and other physical or psychological features. Thus, for a man to imagine 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 we would rationally choose that we ourselves be treated, if we were going to be in these other people's positions, and if we would also be relevantly like them.

The phrase 'would choose' can be misleading. When we apply the Golden Rule, our question should not be how, if we were in the position of some other people and were relevantly like them, 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. (That is clearer with a question like 'Would you want your organs to be used after you are dead?' This question asks us, not to predict our *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'. Though Kant does not describe this argument, he must be thinking of something like: 'Since you would not want to be punished, you ought not to punish me.' Kant seems here to assume that the Golden Rule is

> G4: We ought to treat each other person only in ways in which we would choose that we ourselves be treated if we were going to be in this person's position.

Kant would be right to reject *this* rule. Suppose that, in

> *Case One,* I could save either Blue's life, or White'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. White could similarly argue that I ought to save White's life. So G4 mistakenly implies that, whatever I do, I shall be acting wrongly, by failing to treat someone as I ought to do. [288]

When Jesus appealed to the Golden Rule, was he appealing to G4? Was he intending to imply that, whenever we could save either of two people from death, or lesser harms, whatever we did would be wrong? The answer is clearly No. The Golden Rule should be taken to mean, not

> G4: We ought to treat *each* other person as we would rationally choose that we be treated if we were going to be in *this* person's position,

but

> G3: 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.

So understood, 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 Two*, 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 split and become two people, one in Green's position and the other in Grey's. [289] 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. [290] 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. [291]

When we apply the Golden Rule to certain questions, such as questions about distributive justice, it can make a difference which of these proposals we adopt. But in most cases these proposals would have the same implications. However 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 G3 and G4. But if these people had compared these claims, and seen what they imply, they would have regarded G3 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 G3, judges could reject this argument. [292] These judges would 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 imply that most punishment is wrong.

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 go astray when we apply this rule to some of our duties to others. [293] Suppose that, in

> *Case Three*, I could either save Grey from going blind or save my own life.

If the Golden Rule tells me only how I ought to treat *others,* this rule would mistakenly imply that I ought to save Grey from blindness at the cost of my life. To meet this objection, this rule could become

> G5: 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 most cases, *we* are one of these people. On this version of the Golden Rule, when applied to *Case Three*, 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 before, I ought rationally to choose to save myself from death in one of these positions rather than

from blindness in the other.

This revision better states one of the assumptions of the Golden Rule, which is 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. [294]

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 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'. [295]

These principles mainly differ in the ways in which they make our moral thinking more impartial. Both principles tell us carry out *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. [296] 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 one way harder. When we apply Kant's Law of Nature Formula, we compare two *possible world*s, 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. 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. These four possible worlds may all be very different from the actual world. It would often be hard to predict what these worlds would be like. 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 people whom our acts might affect, but from the imagined point of view of some outside observer, who is not involved in the events 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 we would not be involved. 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 giving or 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 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 starving 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 facts that, in the real world, our acts would affect different people, and that one person's burdens cannot be compensated by benefits to other people. We shall then be ignoring facts that may give us reasons to accept principles of distributive justice. [297]

In these and 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 what has made this rule 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. Though this formula condemns many wrong acts, there are also many exceptions. As we shall see, however, these problems have a Kantian solution.

32 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 these people. I have called this *the Rarity Objection*. Consider, for example,

> *Unjust Punishment:* Unless *Brown* goes to the police and confesses, *Black* will be convicted and punished for some crime that Brown committed. Though Brown knows this fact, he does nothing.

Suppose that Brown acts on the maxim 'Let others be punished for my crimes'. To apply Kant's Law of Nature Formula, we ask whether Brown 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 Brown'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 Brown could rationally will it to be true that everyone acts on his maxim. If Brown lets Black be punished for Brown's crime, Brown would avoid many years in prison. If everyone else acted on Brown's maxim when it applied to them, that would increase the risk that Brown 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 Brown of avoiding these years in prison. Kant's formula therefore permits Brown to let Black be punished for Brown's crime, though this act is clearly wrong. Nor does Kant's Moral Belief Formula condemn this act, since Brown could rationally will it to be true that everyone believes such acts to be morally permitted.

For another example, consider

> *Murderous Theft*: While travelling across some desert, *Grey* and *Blue* have both been bitten by a cobra. Blue has prudently brought some drug that is an antidote to this snake's 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's age is 40, he can expect that his act would give him many more years of life worth living. Blue 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. As before, 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.

Suppose next that 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 that everyone acts on some maxim if there is some other, better maxim on which we could rationally will everyone

to act. One such maxim might be

> E2: Do whatever would be best for me, except when such acts would impose significant 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 E 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 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 may be some ways in which we could partly answer this objection. For example, we might appeal to Rawls's claim that, in asking whether we could rationally will a world in which everyone acts on some maxim, we should suppose that this maxim would already have been acted on for a long enough time for such acts to have had their full effects. We might then claim that, if Grey chose the world in which everyone always acted on the egoistic maxim, he would be running the risk that he would already be dead, having been earlier killed by some other egoist. This somewhat puzzling claim 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 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 three other principles, we take into account Blue's 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 give no weight to 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 be able to claim that Grey could not rationally will 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.

33 The Non-Reversibility Objection

There is another, similar, but more serious 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. [298] 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 some greater burden on someone else. 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*. 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 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. Around here, everyone *does* act in this way. Every hotel owner excludes blacks. And I could rationally will that everyone believes such acts to be right. That's what most of us do believe. And if the blacks believed that too, 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. [299] 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 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. If he is acting on the maxim 'Exclude blacks from my hotel', this man 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. [300] 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. And many wrong-doers' maxims are already nearly universal laws.

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 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. Those to whom some maxim applies may be some powerful and privileged group, who act 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 arguing that these people could *not* rationally will such a world, 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 other people'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, such as self-interested 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 continues to act like them, they could rationally will it to be true that everyone believes such acts to be morally permitted. They 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 rationally will that they keep their privileged position. Similarly, 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 would 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 whenever 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 continue to profit in these ways, or that everyone believes such exploitation to be justified. Since we cannot appeal to the unjustifiability of such exploitation, we could not plausibly defend these claims. For one last example, we can return to global inequality. On any plausible moral view, those who control much the greatest shares of the world's resources ought to transfer much of their wealth or income to the poorest people in the world. Most rich people transfer nothing. To argue that Kant's formulas condemn these people's acts, we would have to claim that these rich people could not rationally will it to be true either that they and others continue to give nothing to the poor, or that everyone believes that, in giving nothing, they are acting rightly. Since we cannot usefully appeal to the wrongness of these people's acts, or to altruistic rational requirements, we could not plausibly defend these claims.

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. [301]

What Kant's formula handles best is not, I have argued, selfishness 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 the rich and powerful are tempted by, and often commit, in their ordinary lives.

34 A Kantian Solution

Some may think that, in presenting these objections, I have misinterpreted Kant's Formula of Universal Law. 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 shall be relevantly like all these other people. [302] This suggestion makes Kant's formula more like the Golden Rule.

None of Kant's claims about his formula support Nagel's interpretation. [303] 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 he wishes for *himself*. [304]

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. [305] 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 Law of Nature Formula, Kant never suggests that we ought to imagine that we know nothing about ourselves or our circumstances. [306]

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 makes Kantian moral reasoning impartial. [307] 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. [308] 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. [309]

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. For reasons that I give in a note [not yet written], it will be enough to revise the moral belief version of Kant's formula. According to

> 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

> MB3: 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 suggested by several of Kant's claims about two of his other principles, the Formulas of Autonomy and of the Realm of Ends. [310] 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 MB3 would always coincide.

This assumption, I have claimed, is false. What could be rationally willed, for example, 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 MB3 sometimes conflict, and we must choose between them. If Kant had seen the need to make this choice, he would have rightly chosen MB3. [311]

Remember next that we ought to revise Kant's formula so that it appeals, not to the agent's maxim, but to what this person is intentionally doing. When we describe the way in which someone acts, that is usually what we are describing. So our revised formula can be

> MB4: 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 people believe some kind of act to be permitted, they accept some principle that permits such acts. So MB4 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?"' [312]

To avoid the New Ideal World Objection, this formula could use

'universal' to mean 'by everyone, and by any other number of people'. This formula also avoids all of our other objections to Kant's original formulas. Since this formula does not appeal to the agent's maxim, it avoids the Flawed Maxims Objection. Since this formula allows us to appeal to conditional principles, it avoids the Permissible Acts Objection. And this formula makes our moral reasoning impartial in a way that avoids the Rarity, High Stakes, and Non-Reversibility Objections.

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, or help to explain why these acts are wrong. When applied to such questions, Wood calls this formula pretty worthless, Herman claims that it cannot be got to work, and O'Neill claims that it often gives either unacceptable guidance or no guidance at all. [313] 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 it appeals only to what the agent could rationally will. But we can revise Kant's formula by dropping Kant's appeal to the agent's maxim, 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 will, or choose, to be the principles that everyone would accept, our view is of the kind that is called *contractualist*. Several writers, such as Rawls and Scanlon, propose 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 11 CONTRACTUALISM

35 The Rational Agreement Formula

Most contractualists ask us to imagine that we and others are all 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 universal acceptance it would be rational for everyone to agree. [314]

Some contractualists appeal instead to the principles to whose being universally *followed* it would be rational for everyone to agree. Many 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 likely to be best for ourselves. We can start with that assumption.

Though there are some principles whose universal acceptance would be likely to be best for everyone, there are other principles whose acceptance would be best only for certain people. What would be best for some men, for example, would not be best for some 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, in this imagined thought-experiment, everyone would know that everyone would accept only the principles that everyone chose. Given this fact, what each of us ought rationally to choose would depend on what other people were likely to choose. There would be no point in our choosing the principles whose acceptance would be best for ourselves, if these principles would not be chosen by some other people.

What we ought rationally to choose would also depend on the effects of our failing to reach agreement. Many 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. This amoral *no-agreement world* would be likely to be bad for everyone. So everyone would have 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. [315] We must all know that, if we failed to reach agreement in this last round, we would have lost our chance, and could not try again. In earlier rounds, it might be rational for us to vote tactically, trying to induce others to vote for principles that favoured ourselves. 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. It might then be rational for everyone to choose the principles that other people would be most likely to choose. This might be everyone's best hope of avoiding the burdens of the no-agreement world.

Such reasoning is, for some writers, the essence of contractualism. On such views, morality 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 claim, 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.

There is now a complication. The no-agreement world would be less bad for certain people, such as those who are rich in the sense of controlling more resources, and those who have greater abilities of various kinds. 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 would 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 could 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 David 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 new lands, they were morally permitted to kill the native inhabitants. [316] On such Hobbesian views, there would be no moral objection if those with greater bargaining power used threats to secure agreement on principles that favoured them.

Kantian contractualists, like Rawls, rightly reject these implications. As Rawls writes, 'to each according to his threat advantage is not a conception of justice'. [317] So we can now turn to Rawls's view.

36 Rawlsian Contractualism

Rawls's version of contractualism does not, I shall argue, succeed. But, if we removed the contractualism from Rawls's great book, the result would be a liberal egalitarian view about social justice that is both in itself very appealing and well supported by some of Rawls's non-contractualist arguments and claims. [318]

In considering Rawlsian Contractualism, we can start with Rawls's assumptions about rationality. Rawls accepts a desire-based theory, according to which we have most reason to do whatever would best achieve what we would most want after informed deliberation. Of those who accept this theory, many believe that it coincides with Rational Egoism, according to which we have most reason to do whatever would be best for ourselves. These people mistakenly assume that, after informed 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. [319]

Since Rawls's theory 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, if these people don't care about morality, we could not honestly recommend justice as a virtue to them, since they would not have sufficient reasons to do what justice requires. [320]

On desire-based 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 would not care even after 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. [321]

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. . . [322]

Rawls here assumes that we have no reasons to care about anything for its own sake. 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.

Many 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. We can here ignore these claims. What is relevant here is Rawls's suggested account of morality, which he calls *rightness as fairness*. [323]

As a contractualist, Rawls appeals to the principles that it would be rational for everyone to choose. 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. [324] On this assumption, Rawls's desire-based theory and Rational Egoism coincide. If we cared most about our own interests, or our future well-being, it would be rational for us, according to desire-based theories, to make the choices that we could expect to best promote these interests. So, though Rawls rejects Rational Egoism, his motivational assumption 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'. [325]

Rawls revises the Rational Agreement Formula by adding 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.

Rawls gives two main reasons for adding this veil of ignorance. First, if people 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'. [326] And 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, it would be rational for everyone to choose the same principles, so agreement would be guaranteed. It is 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 of who we are would ensure that, in choosing principles, we would give equal weight to everyone's well-being. [327]

One of Rawls's main aims, he writes, is to produce a systematic theory which provides an alternative to all forms of utilitarianism. [328] It is surprising that, to achieve this aim, Rawls proposes his version of

contractualism. If we appeal to a combination of self-interested rationality and impartiality, we should expect our moral reasoning to support some form of utilitarianism, or some similar view. [329] As Rawls himself says, utilitarianism is, roughly, self-interested rationality plus impartiality. [330]

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. [331] By choosing this *utilitarian average principle*, we would maximize our own expected level of well-being.

Rawls rejects what we can call this *Equal Chance Formula*. If we were behind the veil of ignorance, he 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 Rawlsian Contractualism to support non-utilitarian conclusions, Rawls 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. [332]

This remark treats 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. What would be wrong with that version of veil of ignorance contractualism? Rawls himself points out that, since there are different contractualist formulas, he must defend 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'. [333] 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. [334]

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. [335]

By supposing that we know as little as possible, Rawls suggests, we would make our reasoning as similar as possible to the reasoning of our noumenal selves in Kant's timeless noumenal world, and we would thereby best express our freedom and autonomy.

This defence of the No Knowledge Formula does not, I believe, succeed. If we start by supposing that, behind Rawls's veil of ignorance, we would have *no* information, and we ought then to add the *least* amount of information that would make a rational choice possible, we ought to appeal to a more extreme version of the No Knowledge Formula. We need not know, for example, 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 might have 'just enough' information to allow us to make a rational decision. [336] 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. [337] 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'. [338] 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. [339] Since Rawls cannot

justify his rejection of this version of Rawlsian Contractualism, Rawls's theory does not, as he intends, provide an argument against all forms of utilitarianism. [340]

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.

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, I believe, 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 choose how to use some scarce medical resources, treating children who 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 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 view. 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, its 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. [341]

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 best for those who would be worst off. There are many moral questions to which this argument would be hard to apply. Suppose that we are comparing different principles about when we could justifiably fail to keep our promises, or tell lies, or impose risks on 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. This 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 does not 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. On Rawls's version of contractualism, all such considerations are irrelevant. Though Rawlsian moral reasoning differs from utilitarian reasoning, it differs only by subtraction. When Rawls describes how people would choose moral principles 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. [342]

Rawls merely denies these people most of the knowledge that self-interested calculations need. Since Rawls's contractors choose principles for purely self-interested reasons, there is no way in which non-utilitarian considerations could possibly enter in.

When he first presents his theory, Rawls writes

> It is perfectly possible . . . that some form of the principle of utility would be adopted, and therefore that contract theory leads eventually to a deeper and more roundabout justification of utilitarianism. [343]

He also writes

> for the contract view, which is the traditional alternative to utilitarianism, such a conclusion would be a disaster. [344]

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. [345]

37 Kantian Contractualism

To reach a more plausible and successful version of contractualism, we should turn 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.

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 will, or 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 all and only the principles that *everyone* chose. According to the Kantian Formula:

> Everyone ought to follow the principles that everyone could rationally choose, if each person supposed that everyone would accept all and only the principles that *she herself* chose.

In applying this formula, we carry out *many* thought-experiments, one for each person. We imagine that each of us applies Kant's Moral Belief Formula, by asking which moral beliefs, or principles, she could rationally will everyone to accept. In making these separate choices, none of us would need to reach agreement with other people. The Kantian Contractualist Formula appeals to the principles that, in these separate thought-experiments, everyone would have sufficient reasons to choose.

Rawls, I believe, is right to reject the Rational Agreement Formula. But 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 different and better way. When we ask which principles everyone could rationally choose, we can suppose that everyone knows all of the relevant facts. Since there is no need to reach agreement, there is no

scope for bargaining, so no one would have greater bargaining power.

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 that would give a greater sum of benefits to others, even if these other people are all much better off. 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. [346]

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. 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. 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*.

The Rational Agreement Formula, as we have seen, fails to achieve this aim. Precisely *by* requiring such unanimous agreement, this formula gives greater power, not to those who *most* need morality's protection, but to those who *least* need such protection, because their greater control of resources or other advantages give them greater bargaining power.

Rawls's formula does little to achieve this protective aim. Though Rawls's veil of ignorance eliminates bargaining power, it 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, 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 possible benefits. [347]

Since the Kantian Formula requires unanimity without appealing either to a veil of ignorance or a need to reach agreement, this formula might better achieve 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. We might plausibly reject that claim.

The Kantian Formula has other advantages. Though Rawls's veil of ignorance ensures impartiality, it does that crudely, like frontal lobotomy. The disagreements between different people are not resolved, but suppressed. Since no one knows anything about themselves or their circumstances, unanimity is guaranteed. In the thought-experiments to which the Kantian Formula appeals, there is no veil of ignorance. Everyone would know how their interests conflict with the interests of others. Since unanimity is not guaranteed, it would be morally more significant if unanimity could be achieved. That would be true if 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 some desire-based theory, or Rational Egoism, the Kantian Formula would not succeed. If each person supposed that she had the power to choose which principles we would all 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 desire-based or aim-based theories. And though Rational Egoism is, in being value-based, a theory of the right kind, this theory is too narrow. On wide value-based theories of the kind that I believe we should accept, 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 can 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 for everyone 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. [348]

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 would have sufficient reasons to choose, and could

therefore rationally choose. If there was no principle that everyone could rationally choose, 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 would either permit or condemn 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. [349] The uniqueness condition would, I believe, be sufficiently often met. I shall partly defend this belief below.

To illustrate the Kantian Formula, suppose that

> some quantity of unowned goods can be shared between different people,
>
> no one has any special claim to these goods, such as a claim based on their having greater needs, or their being worse off than others,

and

> if these goods were equally distributed, that would produce the greatest sum of benefits, with everyone receiving equal benefits.

It is clear that, in such cases, everyone should get 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 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 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 themselves more than equal shares. We must also reject (A) if we accept a desire-based theory. There are many people whose desires 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. 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. Each of us would have both personal and impartial reasons *not* to make any such choice, since such choices would both be worse for us and produce 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 implies that, in such cases, everyone should get equal shares.

38 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. [350]

In a fuller statement:

> Some act is wrong when such acts are disallowed by some principle that no one could reasonably reject, or when someone could reasonably reject any principle permitting such acts.

Though 'reasonable' sometimes means the same as 'rational', Scanlon 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. [351]

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 objectswwwwwwwwww, 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. [352] Remember 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, 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. This view can be roughly stated as

> *the Harmful Means Principle*: It is wrong to impose a serious injury on someone as a means of benefiting others. [353]

According to another 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. [354]

Blue might say 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 beliefs. Grey might claim that

> (1) my act would be wrong, because no one could reasonably reject the Harmful Means Principle.

But she could not defend (1) with the claim that

> (2) no one could reasonably reject this principle because such acts are wrong.

Combining such claims, as I have said, would be like pulling on our boot laces to hold ourselves in mid air. Grey must argue in some other way that no one could reasonably reject the Harmful Means Principle. [355]

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 such a 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*. [356] 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 clearly wrong. As Scanlon writes,

> the implications of act utilitarianism are wildly at variance with firmly held moral convictions. [357]

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

Even without appealing to such convictions, however, Scanlonian Contractualists 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, I could use her transplanted organs to save the lives of five other

people.

According to

AU: We ought always to do whatever would benefit people most.

This principle requires me to save the five 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. [358] This relation is of great importance, since we often depend on our doctors, 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*. [359]

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, 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 produce no anxiety or mistrust.

We can reasonably reject some principle, Scanlon claims, only if we can propose some better alternative. If we reject AU, what alternative should we propose?

In considering alternatives, it will help to compare *Transplant* with two

other cases. Remember that, in

> *Tunnel*, I could redirect a driverless, runaway train, so that it kills White rather than the five,

and that in

> *Bridge*, I could save the five 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.

According to this principle, what is morally important 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 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 would 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. When someone needs some money, for example, we might have sufficient reasons to give this person what she needs, but not have sufficient reasons to want her to get this money from us by deception, theft or coercion. But, when some 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. Partly because we have no such reasons, Scanlon's Formula seems to count against the view that there is a 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, to which we might be led by the anxiety and mistrust argument. According to what we can call

> *the Emergency Principle*: Doctors must never kill their patients either as a means or as a side-effect of saving more lives. In *non-medical emergencies*, however, everyone is permitted to do whatever would save the most lives.

Such non-medical emergencies involve unintended threats to people's lives, such as some fire, flood, avalanche, or driverless run-away train.[360] This principle condemns my killing White in *Transplant*, since I am here White's doctor. But this principle permits me to kill 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 *Transplan*t, 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 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 such 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. Similar claims apply to Kant's Formula of Universal Law.

We might now challenge this Deontic Beliefs Restriction. When we try to answer moral questions by applying these contractualist or Kantian 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. There is no point in claiming both that

> (1) acts are wrong when any principle permitting them would fail some Kantian or contractualist test,

and that

> (2) principles would fail this test when the acts they permit are wrong.

But that is not a good enough reply. Even if these formulas would be useless unless they are combined with 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. This reply misunderstands the way in which good moral reasoning consists in part in appeals to moral intuitions. In trying to decide whether some 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 both some of these principles and some of these intuitive beliefs. 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.[361] 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. [362] 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. [363] If we are constructivist contractualists, and we decide that it would be rational to choose principles that permit slavery, we should conclude that slavery is not wrong. Though slavery may seem to us to be wrong, constructivists reject appeals to our moral intuitions, which some claim to involve mere prejudice, or cultural conditioning, or to be mere illusions.

I shall here assume that we should 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. [364]

If Kantians and contractualists accept that our moral reasoning should appeal to such intuitive beliefs, they must give a different defence of the Deontic Beliefs Restriction.

There is one straightforward and wholly satisfactory 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. 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 can 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. This is the kind of *making wrong* that is relevant here.

Scanlon claims that his contractualism gives an account of wrongness itself, or *what it is* for acts to be wrong. Contractualist formulas are better claimed, I believe, not to describe wrongness itself, but to

describe one of the properties or facts that can make acts wrong. According to

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

If we accept this formula, we should add

> If some act is disallowed by such a principle, that makes this act wrong.

We should not claim that this formula describes the only property or fact that makes acts wrong. As I have just said, there are other wrong-making properties or facts. 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 act one that is disallowed by one of the only principles whose acceptance everyone could rationally will. On this version of Kantian Contractualism, both of these facts can be truly claimed to make such acts wrong.

Scanlon's theory should, I believe, take the same form. According to

> Scanlon's Formula: An act is wrong just when such acts are disallowed by some principle that could not be reasonably rejected.

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. But his formula would then be a *concealed tautology*, whose open form would be the claim that acts are disallowed by such unrejectable principles when they are disallowed by such unrejectable principles. We could all accept that trivial 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 have the *different* property of being wrong in one or more non-contractualist senses. On 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 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 it conflicts too often and too strongly with our intuitive moral beliefs. [365]

CHAPTER 12 CONSEQUENTIALISM

39 What Would Make Things Go Best

Before we ask what is implied by Kantian Contractualism, it will help to say some more about consequentialist moral theories.

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 this belief. Hume, for example, does not use 'good' or 'bad' in reason-involving 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' to mean 'pleasure' and 'pain'. [366]

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. [367]

When Kant claims that pain cannot be bad, he means that pain cannot be morally bad. Like Hume, Kant seems to be unaware of, or to forget, the reason-involving sense in which it is non-morally bad to be in pain. [368]

So does Ross. If some outcome would be bad, Ross assumes, we have what he calls a *prima facie duty* to prevent this outcome, if we can. [369] Because Ross believes that we have no such duty to prevent our own pain, he 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. [370] Ross reaches this strange conclusion because he ignores the non-moral sense in which things can be 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.' [371] Such badness involves reasons that are *omnipersonal*, in the sense that they are reasons for everyone. We all have reasons to regret

anyone's pain, and to try to prevent or relieve this pain if we can. These reasons are also *impartial*, in the sense that they are reasons to regret anyone's pain whatever that person's relation to us. When we consider possible events that would involve and affect only strangers, our actual point of view is impartial, and we have only impartial reasons. But we also have impartial reasons even when our point of view is not impartial, as would be true, for example, if I could use my only dose of morphine to relieve either my own or someone else's pain.

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. But, as I have said, we can explain such a sense. Of some set of possible outcomes, one would be

> *impersonally best* in the *impartial reason-involving* sense when this is the outcome that, from an impartial point of view, everyone would have most reason to want.

These writers might reply that outcomes cannot be in this sense good or bad, since there are no outcomes that everyone has some reasons to want or to regret. And that must be claimed by those who accept either desire-based theories about reasons, or Rational Egoism. On these theories, it would not be in this sense bad if some plague or earthquake killed many people, since this event would not be bad for everyone, nor would everyone have reasons to want this event not to occur. We ought, I have claimed, to reject these theories.

In what follows, I shall use 'best' in the impartial reason-involving sense. In many cases, there would be two or more possible outcomes that could be called 'equal-best'. But this phrase misleadingly suggests that there would be precise truths about the relative goodness of these 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 outcomes that would not be worse than any of the others. To save words, however, I shall use 'best' so that it applies to all such outcomes.

Those whom I call *consequentialists* believe that

> (1) the rightness of acts depends only on facts about how it would be best for things to go.

Consequentialists may differ in their 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.

On this view's hedonistic form, things go best when there is the greatest sum of happiness minus suffering. [372] Other consequentialists believe that the goodness of outcomes depends 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 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 these benefits would go to people who were worse off.

The word 'consequentialist' is misleading, since it suggests that on these views all that matters is the future, and the effects of our acts. Consequentialists can believe that the goodness of some outcomes depends in part on facts about the past. On two such views, for example, it would be better if benefits went to people who had earlier been worse off, and it would be bad if people were punished for crimes that they did not commit. That is one reason for talking, not of the goodness of outcomes, but of how well things go, or will have gone. Consequentialists can also believe that certain acts are in themselves good or bad.

As well as making conflicting claims about what is good, consequentialists can disagree in other ways. All consequentialists appeal to claims about what would make things go best. *Direct* Consequentialists apply this test to everything: not just to acts, but also to rules, laws, customs, habits, desires, emotions, moral beliefs, and anything else that could make things go better or worse. When these people apply this test to acts, they are called *Act* Consequentialists. Some of these people claim that

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

Others claim that

> (5) everyone ought to do whatever would be most likely to make things go best, or more precisely what would be *expectably-best*. [373]

As I have said, however, we ought to use 'ought' in several senses. If claim

(4) uses 'ought' in the knowledge-supposing sense and claim (5) uses 'ought' in either the evidence-relative or the belief-relative sense, these claims would not conflict, and Act Consequentialists could accept them both. Since we often don't know which acts would in fact make things go best, claim (5) is in practice more important. In most of what follows, however, 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 principles 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 not be right to do what would make things go best when such acts are condemned by one of the best principles. Similarly, according to some *Motive Consequentialists*, 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. [374] Such views overlap with some forms of *virtue ethics*: those which appeal to the character-traits and other dispositions that best promote human flourishing. There could be many other forms of Indirect Consequentialism. [375]

40 Consequentialist Maxims

Some consequentialists might apply their test directly to *maxims*, and only indirectly to acts. Some maxims are

> *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 she 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 act on these consequentialist maxims.

In assessing this argument, we must appeal to some view about reasons and rationality. According to wide value-based 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 very good 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's reasons would not be outweighed by any conflicting reasons.

When we apply Kant's formula, some writers claim, we ought to appeal only to a rational requirement to avoid inconsistency, or contradictions in our will. On this assumption, (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. [376] 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 would be best fulfilled.

(B) is also true on desire-based theories. 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. [377] And some of these people would have desires that would be best fulfilled in the world in which everyone acted on the optimific maxims. According to desire-based theories, these people could rationally will it to be true that everyone acts on these 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

non-optimific maxims, ones whose being acted on would give this person extra benefits, though imposing greater burdens on others. But this claim, I believe, is false. Of the fortunate people, some would care strongly about the well-being of others, and if things went in the ways that would be impartially best, that would also 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. On all such views, there would be at least 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 act on these consequentialist maxims.

It is an objection to Kant's formula that it permits only *some* people to act on these maxims, since such moral claims ought to apply to everyone. We can call this the *Relativism Objection*. There are, we have seen, other objections to Kant's formula. To avoid some of these objections, we ought to revise this 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 would apply to everyone.

We ought also to revise Kant's formulas so that they do not appeal to the agent's maxim, in the sense of 'maxim' that covers policies. And, rather than appealing to other possible maxims, these formulas should appeal to possible principles. These revisions lead us back to the Kantian Contractualist Formula. So we can now ask what this formula implies.

41 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. [378]

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 any principles that 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) These impartial reasons would not be decisively outweighed by any 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 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 Rule Consequentialist principles.

When we apply this 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 in all future centuries, or for the rest of human history. This formula appeals to the principles that, in these many thought-experiments, everyone could rationally choose.

According to premise (B), anyone could rationally choose any principles that they would have sufficient reasons to choose. What we have reason to choose depends on the facts, but what we can rationally choose depends on our beliefs. If we are ignorant, or have false beliefs, it may not be rational for us to choose what we have sufficient reasons to choose. But we should

suppose that, when making these imagined choices, everyone would know all of the relevant facts. That would make it more significant if there are some principles that everyone could rationally choose. And, if everyone knows the facts, premise (B) is true.

We can next suppose that, as (C) claims, there is some set of principles whose universal acceptance would make things go best in the impartial reason-involving sense. When we consider some kinds of case, there might be two or more optimific principles that were significantly different. That would raise questions of detail that would be best considered later.

If everyone accepted these UA-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. [379]

According to premise (E), these impartial reasons would not be decisively outweighed by any conflicting reasons. No one would have reasons that would count decisively *against* choosing that everyone accepts the optimific principles.

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 to whom we have close ties. So we might have strong personal and partial reasons *not* to choose the optimific principles.

To decide whether everyone could rationally choose these principles, we must know what the alternatives would be. It will be enough to consider those other principles that are *significantly* non-optimific, since their acceptance would make the rest of human history 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. We should first try to get the main outlines right. Details can wait.

42 Self-interested Reasons

Of our reasons not to choose the optimific 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. Such people would have strong self-interested reasons not to choose these principles.

You might be such a person. Suppose that, in

> *Lifeboat*, you are in White's position, since 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 what 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. [380]

If everyone accepted the Numbers Principle rather than the Nearness Principle, there would be many cases in which some people would act on this principle, and many more 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. [381] If instead you chose that everyone accepts the Nearness Principle, I would save your life. We can also suppose that the five are all strangers to you, and that, in dying, you would lose many happy years of life. You would have strong self-interested reasons not to choose the Numbers Principle, since you have strong reasons to choose that I save your life. According to premise (E), these reasons would not be decisive. Is that true?

On desire-based or aim-based 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 you have no reasons to care about the well-being of others for its own sake. Nor do you have such reasons to care about your own well-being. To adapt Hume's example, if after informed deliberation you preferred the destruction of the world to the scratching of your little finger,

you could rationally choose the destruction of the world. We ought, I have argued, to reject such desire-based or aim-based theories, and accept some value-based 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, which we can call

> *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 two relevantly similar strangers. If that were true, cases like *Lifeboat* would cast no doubt on premise (E). You would be rationally required to choose that everyone accepts some optimific principle, such as the Numbers Principle. [382] But we ought also, I believe, to reject this view.

According to

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

On such views, we are often rationally permitted but not rationally required to give strong priority both to our own well-being and to the well-being of certain other people, such as our close relatives and those we love. We ought, I believe, to accept some view of this kind.

On the most plausible views of this kind, if we could either save our own life, or save 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.

On some more egoistic value-based views, we are rationally required to give strong priority to our own well-being. You would have sufficient reasons to give up your life only if 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 to choose which principles everyone would accept. 'Everyone' here refers to all rational beings, both now and in all future centuries. The principles you chose would be accepted by many billions of people. So, as I have said, 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. [383]

Suppose next that I am mistaken. We ought, I have claimed, to reject Rational Egoism. But there is a more plausible, partly egoistic view that is relevant here. On this view, though we could often rationally choose to bear some great burden when we could thereby save many others 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 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 we are now considering, but also to the Kantian Contractualist Formula. On this view, you could not rationally choose any principle that required or permitted me to save the five, and 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. When applied to such cases, this formula would fail. If we could save either one stranger or a thousand others, this formula would not require us to save the thousand, which would be clearly what we ought to do.

We ought, I believe, to reject High Stakes Egoism. But it is worth describing how, if this view were true, we could answer this objection to the Kantian Formula.

Contractualists appeal to the principles that it would be rational for everyone to choose, in some way that would make these choices sufficiently impartial. Rawls claims that, to achieve such impartiality, we should appeal to the principles that it would be rational for everyone to choose if we were all 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, I believe, be some relevant principle that, even with full knowledge, everyone could rationally choose.

In the kind of case that we are now discussing, we could save either of two groups of strangers, one of which contains more people. In such cases, according to High Stakes Egoism, everyone would be rationally required to give absolute priority to saving their own life. No one in the smaller group could rationally choose that we save the larger group, even if that group contained very many more people. The Kantian Formula would here fail 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. We could partly follow Rawls, by adding a *local* veil of ignorance. When we apply the Kantian Formula to such cases, we could ask which principles everyone could rationally choose if they did not know in which group they would be. On that assumption, everyone could rationally choose the Numbers Principle, which would require us to save the group that contained more people.

The Kantian Formula might be more sweepingly revised, by adding a *global* veil of ignorance. But that would make this formula much less appealing, for reasons some of which I gave 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, that 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, there would be some principle that, even with full knowledge, everyone could rationally choose.

Though High Stakes Egoism has some plausibility, we ought, I believe, to reject this view. We could rationally give up our life, if we would thereby save very many other people's lives. If that is true, the Kantian Formula does not need to be revised.

43 Altruistic and Deontic Reasons

Of our reasons not to choose the optimific principles, others might be provided by facts about 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 requires 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 the strangers rather than your child. If everyone accepted and many people followed such a requirement, things would go in one way better, since more people's lives would be saved. But these good effects would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would save several strangers rather than one of our own children, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and some other similar facts, the optimific principles would often permit us, and often require us, to give some kinds of strong priority to our own children's well-being. [384]

This objection could be transferred to a different case. Suppose that, in

> *Third Lifeboat*, it is I who could save either your child or five other children. These children are all strangers to me.

Any optimific principle would require *me* to save the other five children. We might claim that

> (I) you could not rationally choose that everyone accepts such a principle, since you would have decisive reasons to choose some principle that would require me to save your child.

You would have such decisive reasons, we might claim, because you would have a duty to protect your child.

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 and 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

> (a) some act is wrong because we could not all rationally choose any principle that permits such acts,

it would be pointless to claim that

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

It would be similarly pointless to claim both that

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

and that

> (d) these are the 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 are deciding whether we ought to accept this formula.

Since we cannot appeal to our belief that you have a duty to protect your child, could we defend (I) in some other way? We could most plausibly appeal, I believe, to the fact that you love your child. It may be hard not to be influenced by our belief that you have a duty to protect your child. So it will help to change our example. Suppose that, in

> *Fourth Lifeboat*, I could save either five strangers or one other person who is your most-loved friend. These people are all strangers to me.

Any optimific principle would require me to save the five rather than your 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 some principle that would require me to save the person whom you love most.

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

It may seem absurd to deny that we have decisive reasons to save the person whom we love most. Could Romeo or Isolde have rationally chosen to let Juliet or Tristan die? While discussing a similar example, Bernard 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. . . [385]

I am not, however, appealing to the kind of impartial system that Williams here rejects. First, on the optimific principles that we are considering, we would be often morally permitted or required to give some kinds of 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 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 could rationally give very great 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.

Suppose that, in *Fourth Lifeboat*, your friend is on the rock that is nearer to me. Your friend would know that, if she chose that everyone accepts some optimific principle, I would save the five strangers. If instead she chose that everyone accepts the Nearness Principle, I would save her life. She would have strong reasons to choose that I save her life. But, like your choice in *Lifeboat*, her choice would affect how people would later act in all other cases

of this kind. If your friend chose some optimific principle rather than the Nearness Principle, she would die, but she would indirectly save at least a million other people's lives. She would have sufficient reasons to give up her life, if she could thereby save these very many other people.

If your friend could rationally choose to bear some burden for the sake of benefits to others, that does not imply that *you* could also rationally choose that your *friend* bears such a burden. We may 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 rationally save two, or ten, or even a hundred people rather than the person whom we love most. But, if you chose some optimific principle rather than the Nearness Principle, your choice would also indirectly save at least a million other people. I suggest that, however much you love your friend, you could rationally make the choice that would save at least a million people.

Suppose next that my suggestion 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 this view were true, there would be no principle, in such cases, that everyone could rationally choose, so there would be no principle that, according to the Kantian Formula, everyone ought to follow. 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 our choice of principles would not be even weakly impartial. To avoid this objection, we could revise this formula, by adding 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, there would be some principle that, even with full knowledge, everyone could rationally choose.

44 Other Non-Deontic Reasons

On some value-based theories, there are some things that are worth doing, and some 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.' [386] We may have what can be called *perfectionist* reasons to act in these ways, and to try to achieve such aims.

On such views, it would be in itself good in the impartial reason-involving 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 such aims, and to help other people to do the same. So these perfectionist views do not conflict with the optimific principles. If we accept some such view, that would merely affect our beliefs about how it would be best for things to go.

On some views, however, we 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 our well-being. We may be forced to choose, in Yeats's phrase, between 'perfection of the life, or of the work'. [387] 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 manuscript of my great unfinished novel or the only manuscripts of similarly great unfinished novels by five other writers. I might have personal perfectionist reasons not to choose any optimific principle that would require me not to save my great unfinished book. But these reasons would not, I suggest, be decisive. I could rationally give up my book to save five other similarly great books. If my suggestion is mistaken, we could add another local veil of ignorance. That would make little difference, since such cases would be rare.

There is another, more important possibility. Suppose that some optimific principle would require some acts that we believe to be wrong. When we apply the Kantian Formula, as I have said, we cannot appeal to our belief that such 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 give us *non-deontic* reasons that would count decisively against acting in these ways, and against choosing that everyone accepts any optimific principle that would require such acts.

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 would give us strong non-deontic reasons not to act in these 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 bad for us to be someone who acts in such ways. Such facts would also give us decisive reasons not to choose any principle that would require such acts. Many other wrong-

making facts would give us similarly strong non-deontic reasons.

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. They would, for example, require some people to use artificial contraceptives. If we believe such acts to be wrong, could we claim that we had decisive non-deontic reasons not to use such contraceptives, and not to choose any principle that might require such acts? The answer is clearly No. If it would not be wrong to use artificial contraceptives, we would have no strong reason not to act in this way.

Consider next lying to some would-be murderer to protect his intended victim, and hastening our death to avoid suffering. The optimific principles would require all such lies, and would permit many such suicides. Suppose that, like Kant, we believe that all such acts are wrong. We might claim that we had decisive non-deontic reasons not to act in these ways, and not to choose any principles that would require or permit such acts. But this claim would be false. If such lies and suicides would not be wrong, we would have no strong reasons not to act in these ways. So we would not have decisive reasons not to choose such principles.

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 runaway train, so that White's body would trigger the train's automatic brake, I would kill White but would thereby save the five.

Since it would be better if fewer people die, the optimific principles might require me to act in this way. Suppose we believe that this act would be wrong. Such acts are wrong, we may believe, because they involve

> (1) killing someone as a means of saving other people.

We might claim that

> (2) if some act would be of the kind described by (1), this fact would give us a non-deontic reason that would count decisively against acting in this way, and against choosing any optimific principle that would require such acts.

Is this claim true? If it would *not* be wrong for me, in *Bridge*, to kill White as a means of saving the five, would I have decisive non-deontic reasons not to act in this way?

It might be objected that, if some kind of act is wrong, we cannot usefully ask what would be true if such acts were *not* wrong. In any possible world, such acts would be wrong. But, even if we can't imagine that wrong acts are not wrong, we can imagine changing some of our beliefs about which acts are wrong. Remember that, in

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

Suppose we believe that I would be morally permitted to save the five in this way. Suppose next that in a variant of *Tunnel*, which we can call

> *Loop*, the other track loops round and comes back into the tunnel that contains the five. I could redirect the train so that it would kill White rather the five. But, if White weren't on this other track, White's body would not stop the train, which would then loop round and kill the five. So my act would also save the five by killing White.

Given our beliefs about *Bridge* and *Tunnel*, we may find *Loop* puzzling. If I acted in this way in *Loop*, I would be

> (1) killing someone as a means of saving other people,

which we believe to be wrong in *Bridge*. But I would also be

> (3) redirecting the train so that it would kill fewer people,

which we believe to be permissible in *Tunnel*. Though my act in *Loop* would be of the kind described by (1), we may be uncertain whether this fact would make this act wrong.

If we had such doubts, could we rationally claim that, since my act in *Loop* would be of the kind described by (1), this fact would give me a decisive *non*-deontic reason not to act in this way? The answer, I believe, is No. I would have some strong reasons to want not to kill White. That may be why many people believe that, in *Tunnel*, I would be merely morally permitted rather than required to redirect the train and thereby kill White. But these reasons would not be decisive. If my act would *not* be wrong in *Loop*, the fact that I would be causing the train to kill four fewer people would give me a sufficient reason to act in this way.

Since I would have such a reason, (2) is false. If some act is of the kind described by (1), this fact would give us a decisive reason only if and because this fact would make this act wrong. So we cannot claim that we have decisive *non*-deontic reasons not to choose any optimific principle that would

require such acts.

Consider finally this argument:

> (4) The optimific principles might sometimes require us to torture someone.
>
> (5) The awfulness of inflicting torture gives us a decisive non-deontic reason never to act in this way, and not to choose any principle that might require such acts.

Therefore

> In such cases premise (E) would be false, since our impartial reasons to choose the optimific principles would be decisively outweighed by this conflicting reason. [388]

On some plausible assumptions, (4) is false, since the optimific principles would never require torture. Principles are optimific if their universal acceptance would make things go best. If everyone accepted principles that sometimes required or permitted the use of torture, it would be likely that many people would use torture when such acts would make things go worse. Things would be likely to go best if everyone believed that torture is always wrong.

We should note that, in making that claim, we may not be appealing only to the effects of torture. Acts of torture may be in themselves very bad events. For the optimific principles to require the use of torture in certain kinds of case, it would have to be true that, from an impartial point of view, everyone would have most reason to want or hope that, in such cases, torture *would* be used. If instead everyone would have most reason to want it to be true that, in these cases, torture would *not* be used, things would go best if torture was not used, and the optimific principles would condemn such acts.

Let us suppose, however, that (4) is true, since some optimific principle would in certain extreme cases require torture to be used. That might be true, for example, when torturing some terrorist would be the only way to prevent some nuclear explosion that would kill as many as a million people. According to (5), the awfulness of inflicting torture would give us decisive non-deontic reasons not to use torture even in these cases, and not to choose any principle that would require such acts.

This claim, I believe, is false. If such an act would be wrong, that might give us decisive reasons not to act in this way. But if we would be morally

permitted to use torture in such cases, the fact that we would be saving a million people's lives would give us sufficient reasons to act in this way. So we could not truly claim that we had decisive *non*-deontic reasons not to choose any principle that would require such acts.

Since (5) is false, this argument provides no objection to premise (E). These are not cases in which our impartial reasons to choose some optimific principle would be decisively outweighed by relevant conflicting reasons.

There are, I have said, many wrong-making facts that give us decisive non-deontic reasons not to act in some way. But these various examples suggest that

> (6) 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 (6) 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 were 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 kinds of facts that might give us such reasons, two of the most obvious examples are the facts that some act would involve torture, or would kill some innocent person. But, in the cases that I have discussed, these facts would not, I have claimed, give us such reasons. If even these facts would not give us such reasons, the same would be true of the other facts that can make acts wrong, such as the facts that can make it wrong to lie, steal, or break some promise. When everyone would have impartial reasons to want us to act in these ways, these facts would give us decisive reasons only if, and because, they would make these acts wrong.

We might strengthen my example. We might suppose that, to save a million people's lives we would have to torture, not some terrorist, but some innocent person, who might even be our own child. The optimific principles would not, I believe, require such an act. And, even if they did, this objection would fail, I believe, in other ways. But it is worth asking what would follow if I am mistaken, so I discuss that question in Appendix X. It is

enough to say here that such objections could at most show that some of this chapter's main claims would need to be slightly qualified.

There are, I believe, no other plausible objections to premise (E). So we ought to accept this premise. Everyone would have 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 premises (B) to (E), 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.

45 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 much worse in the impartial reason-involving sense. Things would also go very badly for some unfortunate people. 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. [389] In *Earthquake*, for example, Black could not rationally choose any principle that required or permitted 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 or permitted me to save you rather than saving all of the five. Similar remarks apply to all such cases. So, as (G) claims, there are no significantly non-optimific principles whose universal acceptance everyone could rationally choose.

(F) and (G) 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 in a shorter form. 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.

This argument, we may suspect, must have at least one consequentialist premise. If that were true, this argument might be uninteresting, and have no importance. 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-involving 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 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 mistaken, and ought to be rejected.

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-involving 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 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 everyone would accept whatever principles she chose. Impartiality is here achieved, without any need to reach agreement, 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 arguments with some contractualist premise could have consequentialist conclusions.

CHAPTER 13 CONCLUSIONS

46 Kantian Consequentialism

Return next to Act Consequentialism, or

> AC: Everyone ought always 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. [390] 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. For example, it would be easy 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. And, if everyone had the motives of an Act Consequentialist, that would be bad in other ways. For it to be true that everyone accepted and 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 permit 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, 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 follow 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 in Chapter 12 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. [391]

(N) is not, I believe, true. When we ask what would happen if everyone followed AC, we ought to consider all of the ways in which such a world would differ from some other possible worlds. 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. [392] For some of the reasons Sidgwick gave, we can claim that

> (O) if everyone followed AC, things would go worse than they would go if everyone 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. [393] 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. [394] So this version of the Kantian Formula comes closer to supporting Act Consequentialism. [395]

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. As before, to avoid the New Ideal World Objection, 'universally' might here mean 'by everyone, and by any other number of people'.

Since these different versions of KC 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 it is worth mentioning 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. 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*. 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, we may *not* need to choose between some of the different versions of KC, since we might appeal to these versions of KC, and use these different senses of 'ought' and 'wrong', in such different parts of our moral theory. Similar claims apply to consequentialist moral theories.

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 have not yet thought about them, and 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 claimed, are best revised and combined in Kantian Contractualism. So Kant might have argued:

> 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 gives Kant's moral theory what I have called its most unified and harmonious form. Kant's Formula of the Greatest Good describes a single ultimate end or aim which everyone ought to try to achieve, and Kantian Contractualism describes 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 might 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 believe, to revise (Q). More exactly, we ought to revise Kant's claims about which other states of the world would be closer to the ideal. 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 the spatio-temporal world, no one could deserve to suffer, or to be less happy, because of what they did.

We ought, I believe, to accept this claim. We can add

> (S) All of our decisions *are* merely such events.

Therefore

> (T) No one could deserve to suffer, or to be less happy, because of what they did.

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.

Kant's claims about his ideal world raise another question. In asking how we could get closest to Kant's ideal, we must compare the goodness of virtue and happiness. On one view, the goodness of virtue is infinitely greater, so that if anyone became slightly more virtuous, or slightly less vicious, this change would be better than the achievement of any amount of happiness, however great, or the prevention of any amount of suffering. For this view to seem plausible, we must assume that we have some kind of freedom that could make us responsible for our acts in a desert-implying way. If there could be no such freedom, as I have claimed, [396] we ought to accept a very

different view. If someone is morally bad, by being a cruel murderer for example, that is bad for 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, similar in kind to the badness of someone's being insane. Such badness can be 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. [397]

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.

47 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 other 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.

If this argument is sound, Kantian and Scanlonian Contractualism can be combined. The principles that no one could reasonably reject are all and only 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 admissible or reasonable grounds for rejecting some 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 only in Appendix X.

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. [398]

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 guilt, 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. 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 can be claimed to make these acts wrong. All 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 some kinds of act are wrong, we do 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 by 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, and by one of the only principles whose being universal laws everyone could rationally will, and any principle permitting these acts could be reasonably rejected, these facts may 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 condemn 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'. [399]

For a moral theory to succeed, it must have plausible implications. The Triple Theory has many such implications. But suppose we 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 justifiably reject this theory. But if these conflicts are significantly less deep, or less common, we could justifiably revise these intuitive beliefs. [400]

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, a 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, that can be plausibly claimed to 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 claims:

> 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 can 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 may 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. [401]

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. If we reject (Q), that is because this claim's implications conflict too strongly with some of our intuitive beliefs about which acts are wrong. Rule Consequentialist principles conflict much less often 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 condemn, 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 know 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 wrong because they are disallowed by one of the only principles whose being universal laws everyone could rationally will.

If Kantian Contractualism implies Rule Consequentialism, as I have claimed, that does not make the resulting view wholly consequentialist. Though this view is consequentialist in its claims about which 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-involving 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 the spatio-temporal world, we cannot be responsible for our acts in any way that could make us deserve to suffer, or to be less happy, because of what we did.
>
> 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 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 moral questions must have answers. But some deep 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.

102,157 words

---

[1] I follow Thomas Scanlon, *What We Owe to Each Other*, henceforth WWO, (Harvard University Press, 1998) Chapter 1. Instead of saying that some fact *gives us* a reason, we might say that this fact *is* a reason. We would then need to distinguish between this fact itself, and the fact *that* this fact is a reason. It seems clearer to say that some fact gives us a reason, or provides a reason, since that reminds us of the distinction between the fact itself and its property of being reason-giving, or of being a reason.

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 facts about our desires, or about the wrongness of some act.

[2] Some people claim that, in such cases, *there is* a reason for us to act in some way, but that, because we don't know the reason-giving facts, we don't *have* this reason. These claims do not state a different view, but are merely another way of stating the same view.

[3] When we claim that we have *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 the strength of some reason or set of reasons, without distinguishing between these reasons. Similar remarks apply to the claim that we have *sufficient reason* to act in some way.

[4] I use 'respond' to mean 'respond positively or appropriately'. For example, when I am aware of facts that give me a decisive reason to help someone, I would respond to this reason only if I help this person. I might be said to respond to this reason in a wider sense if I ignore it, or respond inappropriately, by standing on my head.

[5] Like the concept of *a reason*, this version of the concept *ought* cannot, I believe, be helpfully defined in other terms. But we can partly identify this concept as the one that applies to some act when and because this act is what we have most reason to do.

It might be suggested that, when we claim that we ought to do something, we mean that this is what we have most reason to do. This sense of 'ought' could then be fully explained by appealing to the concept of a reason. Perhaps some people use 'ought' in this sense. But, when I claim that we ought to do what we have most reason 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 concept *ought*. I doubt whether this project 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 might claim that the concept of *a reason* and this version of the concept *ought* have the same indefinable common element, differing only in the way that comparatives like *more* or *greater* differ from superlatives like *most* and *greatest*.)

[6] As we shall see in Section 5, this claim does not apply to some of our normative beliefs. The word 'rational' can also be used more thinly, to mean 'not irrational'. Some act might be in this sense rational, though we have no reason to act in this way, if we also have no reason not to act in this way.

[7] 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 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 you ran away from the snake, you might say either 'To save my life', or 'Because my aim was to save my life'.

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 belief in which 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.

[8] WWO 97.

[9] 'Self-interested' does not here mean 'selfish' or 'egoistic'.

[10] From this impartial point of view, we would all have reasons, I believe, to care equally about everyone's well-being. But that is a substantive belief, not something that is true by definition. It might be claimed that, from this impartial point of view, we would all have reason to care more about the well-being of certain people, such as those who are morally best, or those who have the greatest abilities. I should add that, even when our *viewpoint* is impartial, that does not ensure that *we* shall be impartial. We might, for example, care more about the well-being of those strangers who are members of our sex, or race. But we would have no *reasons* to care more about what happens to these people.

[11] As before, I follow Scanlon, WWO Chapter 1. See also Joseph Raz, *Engaging Reason* (Oxford University Press, 199) Chapter 2.

[12] We may also have extrinsic telic reasons, which are provided by some event's non-instrumental relations to other events or things. See Christine Korsgaard, 'Two Distinctions in Goodness', in *Creating the Kingdom of Ends*, henceforth *CKE* (Cambridge University Press, 1996). We can ignore such reasons here.

[13] Credit for such cases may be due to Gregory Kavka 'The Toxin Puzzle', *Analysis*, 43 (1986).

[14] See Niko Kolodny 'Why be Rational?', *Mind*, 2005 Volume 114.

[15] 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 discussion I failed to make it clear that, in my view, the most rational attitude is temporal neutrality.)

[16] A footnote to be added commenting on Frankfurt's contrary claims.

[17] There might seem to be other ways in which we could have desire-based reasons to want avoid this period of agony. Even if our avoiding this agony would not have effects that we want, our *wanting* to avoid this agony might have such effects. And we might *want* to want to avoid this agony. But these claims appeal to state-given reasons; and, as I have argued, there are no such reasons. If that argument fails, we could turn to cases in which we would not have such alleged reasons. We can suppose that, as well as

having no desire to avoid some period of future agony, and no desire whose fulfilment this agony would prevent, we also have no desire to want to avoid this agony, and our wanting to avoid this agony would not have effects that we want. Even if we appeal to the category of state-given reasons, desire-based theories would then imply that we have no reason to want to avoid this future agony.

[18] John Rawls *A Theory of Justice*, (Harvard University Press, 1971) henceforth TJ, 395.

[19] TJ 417.

[20] As Henry 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 this 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 emphases.)

[21] TJ 408 (my emphasis).

[22] TJ 184-5, RE 161.

[23] TJ 432. Rawls describes his imagined man as someone 'whose only pleasure' is to count blades of grass. On a hedonistic theory of the good, it would be best for this imagined man to spend his life counting blades of grass, since that would give him most pleasure. But Rawls then writes that his 'definition of the good forces us to admit that the good for this man is indeed counting blades of grass'. And, on Rawls's definition, the best life for someone is not the life that would give him most pleasure, but the life that conforms to the plan that this person would choose after carefully considering the facts.

[24] TJ 401. Cf 111 and 451.

[25] 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.

[26] There could also be *non*-reductive desire-based theories about reasons. 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.

[27] In these remarks, I follow Thomas Nagel, *The View from Nowhere*, henceforth *VFN,* (Oxford University Press, 1986) 141, and *The Last Word*.

[28] As Scanlon argues, in WWO, Chapter 1.

[29] Scanlon discusses these people (without their names) in WWO 29-31.

[30] At one point Scanlon suggests that we should use the word 'irrational', not in the ordinary sense that applies to people who are open to the strongest rational criticisms, but in a restricted sense, which applies only to apply to people whose preferences fail to match their beliefs about reasons. But, if some of these people are not open to the strongest criticisms, and some other people are, this use of 'irrational' would be misleading.

[31] As Thomas Schelling argues, there are many such cases.

[32] Comment on contrary claims and arguments made by Temkin and Rachels.

[33] Desire-based theories could be revised

[34] 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' (my italics, ME xxviii).

[35] ME, Concluding Chapter p 000. This is only part of Sidgwick's view. Sidgwick makes other claims, to which I shall turn in Section 7.

[36] 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).

[37] John Findley, *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. If we did not have reasons to be specially concerned about our own well-being, it would matter much less how benefits and burdens are distributed between different people. 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.

[38] 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.

[39] 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).

[40] 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).

[41] ME 508.

[42] In some cases, there would be some possible act whose outcome would be in between what would be impartially best and what would be best for ourselves. 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 ourselves from one hour of pain and save the stranger from five hours of pain. Though (3) would be neither impartially best, nor best for ourselves, Sidgwick's argument implies that we could rationally do (3). We are then assessing the strength of our reasons from both these points of view, and choosing a compromise.

---

[43] Sidgwick's view also implies that no impartial reason could be outweighed by any self-interested reason. We could rationally do what we knew would be impartially only slightly better, even if this act would be very much worse for ourselves. This claim may also seem too extreme, since it permits us to give absolute priority to what would be impartially best. But this part of Sidgwick's view is not extreme, because our own well-being counts from an impartial point of view. Suppose that I could either

(1) save myself from 99 days of pain,

or

(2) save some stranger from 100 days of pain.

Of these acts, (2) would be impartially only slightly better, and this act would be much worse for me, since I would be failing to save myself from 99 days of pain. But, if I chose to bear this burden, I would be saving someone else from an even greater burden. In permitting us to act in such ways, Sidgwick's view does not imply that we could rationally give absolute priority to the well-being of other people. This view implies only that we could rationally give equal weight to everyone's well-being.

[44] VFN 160.

[45] Some of us would add 'unless this person deserves to be in pain'.

[46] I am here following Scanlon.

[47] When we compare two kinds of reason from the same point of view, that does not ensure that these reasons are at least partly comparable. Epistemic and practical reasons are, I believe, wholly incomparable. My claim is about partial and impartial reasons provided by facts about people's well-being.

[48] VFN 161.

[49] 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. But some act's being generous and fine does give us some reason to act in this way.

[50] Jefferson McMahan points out that, since my act would be generous and fine, this act would make things go impartially better. 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.

[51] When we are choosing between more than two possible acts, we could often rationally do what would be neither best for ourselves nor best for these strangers, but some compromise in between. [Compare this view with Scheffler's agent-centred prerogative.]

[52] ME 386 note 4.

[53] ME 508. On 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 view, because he believes that our duty is always to do what would be impartially best.

[54] 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'.

[55] TJ 575.

---

[56] This is well-argued, for example, by Niko Kolodny in 'Why be Rational?', Mind, 2005 Volume 114.

[57] It might be suggested that, if this sense of 'wrong' can be expressed with the phrase 'mustn't-be-done', it *can* be helpfully explained in other terms. Something mustn't-be-done when we have decisive reasons not to do it. But that isn't the concept that is relevant here. I might say that you mustn't touch some live electric wire, because doing that would injure you. Your touching this wire would then something that mustn't-be-done. But that doesn't make this act morally wrong, or something that mustn't-be-done, in the distinctive but undefinable moral sense that I am discussing here.

[58] To exclude Rational Egoism, we might use 'our duty' to mean 'what we have decisive *moral* reasons to do'. This definition tells us little, since it uses the word 'moral'. And, to use this sense of 'duty', we must be Moral Rationalists, who believe that no one ever has sufficient reasons to act wrongly. This view is not, I believe, true.

[59] ME 382-3.

[60] ME 200,403.

[61] Or, more precisely, the greatest sum of happiness minus suffering. Rather than claiming that we ought to maximize happiness, these consequentialists would do better to claim that we ought to minimize suffering, or more precisely to minimize the sum of suffering minus happiness. These claims are equivalent, in the way that minimizing net losses is equivalent to maximizing net profits. But, by telling us to minimize suffering, these consequentialists 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 conveys what makes it plausible.

[62] When we consider some kinds of case, we would need to extend and qualify these rough definitions. It may be objected that these definitions are circular, since their explanations of these senses of 'right' all use the word 'right'. But these definitions are not intended to explain what 'right' means in other terms. Their aim is only to draw a distinction between three senses of 'right'. These senses coincide when they are applied to cases in which people know all of the relevant facts. In such cases, the same acts would be right in all three senses.

---

[63] TJ 52. Nagel similarly writes that 'ethical theory. . . is in its infancy', *Other Minds* (Oxford University Press, 1995) 182.

[64] *The Groundwork*, henceforth *G* 392. Page references are to the page numbers of the Prussian Academy edition, which are given in most English translations.

[65] 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).

[66] G 430.

[67] CKE 139.

[68] Onora O'Neill, *Constructions of Reason*, henceforth *CR*, (Cambridge University Press, 1989), 111.

[69] CKE 140.

[70] I here follow CKE 295-6.

[71] After saying that the person whom he deceives 'cannot possibly consent to my way of treating him', Kant treats this remark as having introduced what he calls 'the principle of other human beings' (G 430). (A) is the simplest statement of this principle.

[72] CR 110.

[73] 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).

[74] We can add 'or be of concern to such people'. We can have personal reasons to be concerned about whether someone acts in some way, even if these acts would have no effects on us. One example are acts that affect our children after we are dead. To save words, I shall ignore such reasons.

[75] Two people could not have such power, as would be shown if they made choices that conflicted.

[76] Other writers have assumed or claimed that this is what Kant means. One example is Thomas Hill, in his *Dignity and Practical Reason*, henceforth *DPR*, (Cornell University Press, 1992) 45.

[77] That is true, for example, when he claims that we could not will that some maxim be a universal law.

[78] G 429-30, my italics.

[79] Of these two conditions, the second is more important. Even if 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. And, even if these people could *not* rationally share our aim, there may be no objection to our act if these people *could* rationally consent to our way of treating them.

[80] 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.

[81] G 436.

[82] Rawls, *Lectures*, 191 and 182-3.

[83] *Critique of Practical Reason*, henceforth *Second Critique*, note on p.8.

[84] Barbara Herman, *The Practice of Moral Judgment*, henceforth *PMJ* (Harvard University Press, 1993) vii.

[85] In considering this and my other imagined cases, we should assume that there are no other morally relevant facts. For example, we should suppose that, in Earthquake, there is no other rescuer whom I can rely upon to save Blue's life, that Blue and Grey are strangers both to me and to each other, and that neither is responsible for the risks to them.

[86] 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 do that. So this principle would mistakenly imply that I cannot avoid acting wrongly.

[87] Others would rightly try to persuade Blue not to act in such a way. (Things might be different if Blue was old and Grey was a young professional dancer. In that variant of the case, Blue might have sufficient reasons to

consent to our saving Grey's leg rather than Blue's life, since Blue's loss might here be less than Grey's, or be not much greater.)

[88] I say 'nearly always' because some principle can be defensible, even when it fails in some cases, if these counter-examples are sufficiently rare, or they have some feature that makes them indecisive, or if the principle can be revised in a way that preserves its importance.

[89] This argument was suggested to me by Ingmar Persson.

[90] A note to be added here, discussing some possible exceptions to this claim.

[91] This principle would condemn some act if some people could not rationally consent to it. If this act is required, all of its alternatives would be wrong. The wrongness of these alternatives would give everyone sufficient reasons to consent to our failing to act in these other ways. So the wrongness of these acts would also give everyone sufficient reasons to consent to our doing what is morally required.

[92] There are, of course, other alternatives. This person would have sufficient reasons to consent to my giving this money to some agency which would use it to save someone else from some similarly great harm. This fact does not affect my claim that, in such cases, the Consent Principle requires me to make such a gift.

[93] *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.

[94] *Lectures on Ethics*, edited by Peter Heath and J.B. Schneewind, henceforth *Lectures* (Cambridge University Press, 1997) 179 (Prussian Edition, 27: 416).

[95] References to Murphy, Mulgan, Cullity, and others.

[96] which I shall discuss in this unwritten note.

[97] G 392.

[98] 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.

[99] Kamm writes: 'But because this requirement that we must never violate is so weak (i.e., so easily satisfied) it allows, if it is not supplemented, much (e.g., going so far as to blind someone for trivial reasons) that my admittedly overrideable constraint on treating someone as a mere means need not allow' (Frances Kamm, *Intricate Ethics* (OUP, forthcoming) 000.)

[100] G 423. Kant discusses someone for whom 'things are going well', and who 'contributes nothing' to those who are in need.

[101] MM 443.

[102] As this example also suggests, the moral belief mentioned in condition (1) need not be true.

[103] G 429.

[104] I am not assuming here that, whatever our motives, it cannot be wrong for us to save someone's life. As Marcia Baron suggests, 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.

[105] For a further defence of these claims, see pages 000 below.

[106] 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.

[107] 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).

[108] See page 00 above.

[109] 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 is, I believe, mistaken. Suppose that my doctor presses my chest, saying 'Tell me where it hurts', as a means of finding out whether I have a broken rib. My doctor is hurting me as a means of getting this information, but she isn't *using* me, or treating *me* as a means.

And, as I have said, she gives me this painful treatment *merely* as a means, and not at all as an end, or for the sake of hurting me. This distinction can be easily missed, since it is sometimes drawn only by emphasizing different words. When my doctor gives me this painful medical treatment, she *treats* me merely as a means, but she doesn't treat *me* merely as a means.

[110] CR 111 and 114.

[111] CKE 347. Korsgaard may be intending only to describe Kant's view.

[112] 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' (CR 138). Deception and coercion are not, however, appropriate ways of treating things, since neither is even possible. And there are some appropriate ways of treating both things and persons. O'Neill's suggestion might become: 'To treat people as a mere means is to treat them in ways that are inappropriate to persons.' On this definition, when we claim it to be wrong to treat people as a mere means, we would mean that it is wrong to treat people in inappropriate ways. That would not be a helpful claim.

[113] CKE 142.

[114] CKE 93.

[115] TJ, 111. and 184

[116] John Rawls *Collected Papers*, edited by Samuel Freeman (Harvard University Press 2001) 355.

[117] Since Rawls makes no use of these proposed senses of 'right' and 'true', my remarks are no objection to his moral theory.

[118] As when he claims that, if we kill ourselves to avoid suffering, or give ourselves sexual pleasure, we thereby treat ourselves merely as a means (G 429, and MM, 425).

[119] 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.

[120] 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!'

[121] 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.

[122] We might claim that it would be wrong for him to save his own life in this way.

[123] G 428.

[124] KET 152-5.

[125] KET 117.

[126] MM 462-8.

[127] Kant's *Lectures*, 58-9 (Prussian edition 27: 264-5), and Sidgwick's ME 74-5. According to some writers, we fail to respect persons if we act in ways that are incompatible with respect for people's rights. That is another unhelpful claim, since it does not help us to decide which rights people have.

[128] KET 141, and Shelly Kagan, 'Kantianism for Consequentialists', in Allen Wood's translation of the *Groundwork* (Yale University Press, 2002) 000.

[129] KET 155.

[130] KET 139. (This longest book is *the Metaphysics of Morals*.)

[131] MM 444 and 392.

[132] MM 423-5.

[133] MM 429-30.

[134] KET 154, and 371, note 32.

[135] TJ 31, note 16.

[136] KET 141.

[137] Herman, PMJ 208, 153.

[138] I here follow Scanlon, WWO Chapters 1 and 2.

---

[139] Possible events would be intrinsically good as ends when our reasons to want them to be actual are provided by their intrinsic features. On most views, happiness is such an end. Intrinsic goodness is often contrasted with instrumental goodness, or goodness as a means. But, as Korsgaard points out (CKE, Chapter 9), there are really two distinctions here. One is between goodness as an end and as a means, the other is between intrinsic and extrinsic goodness. Some events may be extrinsically good as ends. Such events would have extrinsic features, such as relational properties, that would give us reasons to want these events to happen for their own sake, and not merely as a means to some good end. On Kant's view, happiness is only extrinsically good as an end, because someone's happiness is good only when and because this person is morally good, and therefore deserves to be happy.

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 some event would be good as a means if it would be an effective means to some end, but that we have no reason to want some event that would be good as a means to some end if this end is not good, or good for us, since it is not an end that we have any reasons to want.

[140] *Principia Ethica* 171. (At the end of this paragraph he seems to contradict this claim.)

[141] WWO 99.

[142] 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.

[143] WWO 104.

[144] WWO 105.

[145] 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.

---

[146] G 435-6.

[147] Herman writes, 'the domain of the good is rational activity and agency, that is willing' (PMJ 213).

[148] G 396-7.

[149] 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.

[150] 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). My translation 'the Greatest Good' better suggests what Kant means.

[151] For references, see the notes near the start of Section 22.

[152] *Second Critique* 119.

[153] G 428.

[154] *The Critique of Judgment* 442-3.

[155] PMJ 238.

[156] KET 133.

[157] 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.

[158] PMJ 213.

[159] 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).

---

[160] G 435.

[161] 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.]

[162] Hill says that this view 'reflects an extreme moral stand that few of us, I suspect, could accept without modification' (DPR 38). (It might be objected that, even on this view, we ought to save these people from pain, since their pain would prevent them from engaging in rational activities. But this is not a good enough reply. For example, it would not apply to cases in which we could relieve people's pain only by giving them anaesthetics that would make them unconscious.)

[163] DPR 50-57.

[164] G 435.

[165] 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 he 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. We should admit that, on Kant's view, there are several kinds of thing that have supreme or unsurpassed value.

[166] MM 427.

[167] PMJ 215.

[168] PMJ 210.

[169] See, for example, the *Second Critique* 20.

---

[170] As I have said, there are other kinds of value which are not kinds of goodness, such as economic value. That is irrelevant here.

[171] PMJ 129.

[172] 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).

[173] *Second Critique* 125, Kant writes 'We', but he means 'all of us' or 'everyone'. And he 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).

[174] *Second Critique* 129.

[175] 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).

[176] 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).

[177] 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).

[178] My statement of HAU may seem to conflate two incompatible versions of this view, of which one tells us what we ought to do, and the other tells us what we ought to *try* to do. But, as I have claimed in Section 8, we ought to use 'ought' in different senses, and that allows a moral theory to tell us *both* what we ought to do, *and* what we ought to try to do.

---

[179] It might seem better to use 'telic' more widely, so that this word applies to any view or theory which claims that we ought to try to achieve certain ends, or aims. We could then distinguish between agent-neutral telic theories which give everyone common aims and agent-relative telic theories which give different people different aims. This wider use of 'telic' applies plausibly to Rational Egoism, which is just like Act Utilitarianism except that it tells each person to do what best promotes, not everyone's well-being, but her own well-being. But all other theories could also be described as giving people agent-relative aims, such as the aims of keeping their promises, promoting the well-being of their children, or obeying the Ten Commandments. Since every theory can be called 'telic' in this wider sense, this sense is not worth using.

[180] I contrast *good* with *ought* rather than *right* because of the fact that 'right' can mean 'not wrong', so that there can be many possible acts that would be right, and none of these acts would be something that we ought to do. *Ought* is a more important concept. (So is *wrong*. But we don't need to discuss both these concepts, since we ought to do something just when every alternative would be wrong, and acts are wrong when we ought to do something else.)

[181] Moore declares that the word '"right". . . does and can mean nothing but "cause of a good result" (*Principia Ethica* 196). 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.)

If Moore is using 'right' to mean 'cause of a good result'---or, as he must intend 'cause of the best result'---Moore's version of act consequentialism would not be a substantive moral view, but a *concealed tautology*. In claiming that an act is right if and only if it would cause the best result, Moore would mean that an act is the cause of the best result if and only if this act is the cause of the best result. Everyone could accept that trivial claim. 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, 109).

[182] *Second Critique*, 63-4.

[183] G 413. Explain why the word 'ought' is a mistranslation.

[184] G 412.

[185] In these ordinary senses, for example, some act may be good, though some other act would be even better. In such cases, the first act would not be practically necessary.

[186] 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.

[187] *Second Critique* 64.

[188] I haven't yet argued this..

[189] Lectures 440-1 (27:717). This 'highest end' is the Greatest Good.

[190] *Religion Within the Boundaries of Mere Reason*, 6: 8.

[191] 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.

[192] As I argue in *Reasons and Persons*, Section 25, we ought to reject this view. But that fact is irrelevant here.

[193] *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* he 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.

[194] 'On a supposed right to lie from philanthropy' (8: 425-30).

[195] 8: 426.

[196] *Lectures* 388 (27:651).

[197] *Metaphysik* L1,28:337, cited in Paul Guyer *Kant on Freedom, Law, and Happiness* (Cambridge University Press, 2000) 94.

[198] See, for example, the quotations in note 178 above.

[199] 'There is nothing in the CI-procedure that can generate precepts requiring us to proportion happiness to virtue' (*Lectures*, 316.)

[200] *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).

[201] *Second Critique* 27.

[202] *Second Critique*, 19.

[203] G 423.

[204] 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 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 be such laws. That is true of 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.

[205] MM 453. He also refers to the universality of a law that everyone *could* act in certain ways (G 422, my emphasis).

[206] It is unclear in what sense it might be impossible for everyone to be permitted to act on some maxim. This might be claimed to be impossible if it would be wrong for everyone to act on this maxim. But (A) would not then help us to decide whether such acts would be wrong. Kant elsewhere claims

that it is wrong to act on some maxim, not if it would be *impossible* that everyone be permitted to act upon it, but if we could not rationally *will* that everyone be so permitted. That is a much more plausible claim, to which I shall return.

[207] See, for example, O'Neill, CR, 157. (O'Neill's view has since changed. See, for example, *Towards Justice and Virtue*, 59.) We successfully act in some way if our act achieves our aim.

[208] This is most clearly shown in Kant's discussion of lying promises on G 422.

[209] PMJ 118-119.

[210] PMJ 119.

[211] CKE 136.

[212] *Lectures* 232-3 (29:609).

[213] *Second Critique*, 19.

[214] G 402-3, and 422.

[215] G 422.

[216] Rawls *Lectures* 169.

[217] The point is not only that people would not trust each other's promises. 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.

[218] 'On a supposed right to lie from philanthropy' 8 425-30.

[219] These imagined cases might be claimed to be unrealistic, because 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 (F). 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.

[220] CKE 95.

---

[221] CR, 133 and 215 and elsewhere.

[222] CR 138-9.

[223] CR 215-6.

[224] As O'Neill points out, CR 102-3.

[225] CKE 92-3.

[226] This is Herman's example (PMJ 138-9).

[227] I take this example from Simon Blackburn *Ruling Passions* (OUP, 1998) 218.

[228] Herman again (PMJ 141).

[229] 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 he first proposes this formula, Kant 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).

[230] PMJ 123.

[231] 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.

[232] 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 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.

[233] 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 he also writes: 'if everyone . . *considered* himself authorized to shorten his life as soon as he was thoroughly weary of it' (*Second Critique* 69). (In both quotations the emphases are mine.)

[234] 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. That 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 believers in God 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.

[235] G 403.

[236] Rawls, Lectures, 166-70. who attributes this point to Herman.

[237] 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, it thereby implies that this act is morally permitted,

[238] CR 85.

[239] See Wood's excellent discussion, KET 103-5.

[240] Lectures, 187

[241] MM 455-7.

[242] *The Second Critique* 34.

[243] 'What is Kantian Ethics?' in *Groundwork for the Metaphysics of Morals*, translated by Allen Wood, (Yale University Press, 2002) 172.

[244] PMJ 104, 132.

[245] Onora O'Neill, *Acting on Principle*, henceforth AOP (Columbia University Press, 1975) 129, 125. See also CR 130

[246] *Human Welfare and Moral Worth*, 122.

---

[247] PMJ 117.

[248] CR 86, 98, 103. O'Neill is here appealing to Kant's claim that. .

[249] G 403.

[250] G 404, 424.

[251] *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.'

[252] 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 could often shorten our lives. Smokers might do that every time they smoke. But Kant is here discussing a single act of suicide.

[253] O'Neill , Herman, Pogge, and Shelly Kagan all make 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.

[254] It might be objected that this maxim is not really a policy, because I can't expect to be able to act on it again. But this objection to LN2 could appeal to possible maxims that are policies, because they could be acted on more than once, but which could be rationally willed to be universal because such acts would be rare.

[255] AOP 112.

[256] 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).

[257] CR 87.

[258] AOP 112-117, and 124-143, and CR 130. Herman makes similar claims in PMJ Chapters 4 and 10.

[259] 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 we don't need to appeal to the agent's maxim. If Kant told the truth (1) because he believed this 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.)

[260] As is suggested by his remarks about the maxim of never helping others who are in need (G 423). On the other most plausible reading, the implied alternative is that only we act on the maxim that we are considering. It would seldom make a difference whether the alternative is only we or no one. [Check]

[261] Add a comment on what are wrongly called Repeated Prisoner's Dilemmas.

[262] There are also probabilistic each-we dilemmas, which appeal to the likely effects of different acts. 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.

[263] When we help strangers, even if that takes time and uses some of our resources, these may not be true costs. We may be glad to help. Such cases are irrelevant here.

[264] There is a further distinction between those goods which in fact benefit even those 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 be in the first category, clean air in the second.

[265] In claiming that Kant's formula condemns these acts, I am not assuming that we could never rationally will it to be true that we all act in some way that would be worse for all of us. My claim is restricted to each-we dilemmas, when these involve benefits to ourselves or our M-related people. Here is another way to show what makes such cases special. If we and the other members of the relevant group could all communicate, and knew each other to be trustworthy, we would all be rationally and morally required to make a joint conditional promise that we shall always act differently, by giving the greater benefits to others. If this joint promise would become binding only if everyone makes it, this fact would *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 and 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. The conflict with Kant's formula would then disappear. 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.

[266] MM 393.

[267] In a different way, however, this solution may be indirectly collectively self-defeating. See page 000 below.

[268] This requirement may also seem unfair. Thus, in an unsolved Parent's Dilemma, 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.

[269] For a suggestion about when that might be morally permissible, see my *Reasons and Persons*, 100-1.

[270] 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.

[271] 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 the aim is to benefit the agent. And, in many cases, the points we are making are not affected by the agent's aim.

[272] 'The Categorical Imperative', op.cit., 190. Pogge is here following an unpublished lecture given by Scanlon in 1983.

[273] 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.)

[274] 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.

[275] 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.

[276] Latest collection, 66.

[277] Similar but more complicated claims would apply to other cases, those in which it would be best, not if everyone acted in the same way, but if everyone played her part in the best possible patterns of acts.

[278] 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 does not affect my argument.

[279] This rule is not in fact ideal, for reasons that I describe in Section 46, but this point is irrelevant here.

[280] For the best recent statement and defence of Rule Consequentialism, see Bradford Hooker, *Ideal Code, Real World* (Oxford University Press, 2000).

[281] See Hooker's discussion of this question in *Ideal Code, Real World*, op.cit. See also Liam Murphy, Timothy Mulgan, and Garret Cullity .

[282] As Herman notes, PMJ Chapter 7.

[283] 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.

[284] MM 451. I have changed 'benevolent' to 'beneficent', since that must be what Kant means.

[285] The ancient Near East, India, and China. Add references.

[286] G 430 note.

[287] G. 423 (my italics).

[288] We might claim that the Golden Rule does not here require me to save Blue from going blind. Even if I were going to be in Blue's position, I could rationally choose that I be left to go blind, so that White's life would be saved. If we make such claims, however, Kant's objection to the Golden Rule could be similarly answered. Even if Kant's judge were going to be in the criminal's position, he could rationally choose that he be punished, so as to protect others by deterring future crimes.

[289] Thomas Nagel, *The Possibility of Altruism* (Oxford University Press, 000) 000, and Equality and Partiality 000-000.

[290] R.M.Hare, *Freedom and Reason*, 000.

[291] TJ, *passim*.

[292] As Leibniz pointed out. See *Leibniz: Political Writings* 2nd edition translated by Patrick Riley (Cambridge, Cambridge Univeristy Press, 1988) 56. (I owe this reference to D.D.Raphael *Concepts of Justice* (Oxford University Press: 2001) 84-5.)

[293] MM 450-1.

[294] Kant similarly claims, '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.

[295] CR 94.

[296] We don't even need to decide what is the morally relevant description of these acts.

[297] TJ, section 30.

[298] G 424.

[299] See Allen Wood's KET op.cit. 3 and 7.

[300] See, for example, G422.

[301] CKE, 101.

[302] Thomas Nagel, *Equality and Partiality* (Oxford University Press, 1991) 42-3.

[303] 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.

[304] G 423, (my emphases).

[305] 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. Comment on his remark about Kant's mistakes, also on the passage in the *Second Critique*.

[306] Quote and discuss the passage from the *Second Critique* to which Rawls appeals.

[307] T.C.Williams, *The Concept of the Categorical Imperative*, (OUP, 1968), 123-131.

[308] Thomas Scanlon, WWO 170-1, and in unpublished summaries of lectures.

[309] G 402.

[310] For example, Kant 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.

[311] 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.'

[312] WWO 171.

[313]KET 172, PMJ 104 and 132, AOP 125 and 129.

[314] To avoid the Ideal World Objection, 'universal' could again mean 'by everyone and by any other number of people'.

[315] or something similar, such as steadily increasing penalties for failure to agree.

[316] See Brian Barry,

[317] TJ 134, Revised Edition, henceforth RE, 116.

[318] 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.

[319] TJ 569, RE 498.

[320] TJ 575, RE 503-4.

[321] *Political Liberalism*, (Columbia University Press, 1996) 49.

[322] 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.

[323] TJ sections 18-9.

[324] 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.

[325] TJ 142.

[326] TJ 140.

[327] 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.

[328] TJ 22.

[329] Add references to Brian Barry.

[330] He writes, for example, 'the utilitarian extends to society the principle of choice for one man'(TJ 28).

[331] 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.

[332] TJ 168, RE 145.

[333] TJ 122 and 121, RE 105.

---

[334] In giving this objection, I again follow Nagel in his 'Rawls on Justice', op.cit.

[335] John Rawls, *Selected Papers*, edited by Samuel Freeman (Harvard University Press, 1999) 335-6. See also TJ Section 40.

[336] As Rawls claims, TJ 397

[337] Add some remarks about G. A. Cohen's discussion of this question.

[338] *Selected Papers* op.cit 265.

[339] TJ 166, RE, 143.

[340] This objection to Rawls's argument I take from Nagel's 'Rawls on Justice', op.cit. 11.

[341] Even when applied to the basic structure of society, the Maximin Argument may be unable to avoid such extreme conclusions. 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, 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 they are not better off than anyone else.

[342] TJ 584, RE 512.

[343] TJ 29, RE 25-6.

[344] 'Distributive Justice: Some Addenda' 1968, republished in John Rawls, *Collected Papers*, edited by Samuel Freeman (Harvard University Press 2001), 174.

[345] 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 supposes 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. And he supposes 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.

[346] TJ 4, RE 3.

[347] As I have said, though it might be rational to choose principles which guaranteed some minimum, or gave greater weight to avoiding what would be the worst outcomes for ourselves, the resulting principles would be fairly close to rule utilitarian principles.

[348] 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.

[349] See Scanlon's discussion in WWO 333-342.

[350] WWO 4-5 (and elsewhere).

[351] 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.

[352] Scanlon appeals to this restriction (though not with this name) on WWO 4-5, 194, and 213-6.

[353] As before, we impose an injury on someone as a means of benefiting others if we do something to someone as a means which predictably also injures this person.

[354] 'Contractualism and Utilitarianism', in *Moral Discourse and Practice*, edited by Stephen Darwall, Allan Gibbard and Peter Railton (OUP, 1997) page 272.

[355] 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 might be called 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.

[356] WWO 215.

[357] 'Contractualism and Utilitarianism', 267.

[358] This anxiety might not be rational, but that does not undermine these claims.

[359] 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 how such anxiety and mistrust would be bad for very many people. 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 have argued in 'Justifiability to Others', in *On What We Owe to Others*, edited by Philip Stratton-Lake (Blackwell, 2004). In some of his more recent writings, Scanlon is less committed to this restriction [add references].

[360] These emergencies do not include intended threats to people's lives, such as threats by terrorists. Such cases have special features, such as the need not to encourage such threats, and must therefore be covered by some other principle.

[361] *Selected Papers* 344.

[362] 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.

[363] *Selected Papers*, 351.

[364] Scanlon, 'Rawls on Justification', in *The Cambridge Companion to Rawls*, edited by Samuel Freeman, (Cambridge University Press, 2003) 149.

[365] 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.

[366] As when he writes, 'Besides good and evil, or in other words, pain and pleasure. . . ' 439.

[367] 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.

[368] 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.

[369] As Ross admits, what he calls *prima facie duties* are neither prima facie (which means 'at first appearance') nor duties. They are, roughly, moral reasons to act in some way that would make this act a duty in the absence of similarly strong conflicting reasons.

[370] 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.)

[371] The *View from Nowhere*, 161.

[372] There are also some utilitarians who are not in my sense consequentialists, since these people make no claims about what would be best in the impartial reason-involving sense, or any other reason-involving sense. Most hedonistic utilitarians would broaden (2), so that it covered benefits and burdens to all sentient beings, and most non-hedonistic utilitarians would include the pleasure and pain of non-rational animals in their claims about about how it would be best for things to go.

[373] In the sense explained in Section 8 above.

[374] Explain how Adams has been misunderstood.

[375] 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. 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.

[376] See, for example, Rawls, *Lectures*,173-6 and 232-4.

[377] If these people themselves accept a desire-based theory, 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.

[378] To avoid the New Ideal World Objection, as I suggest in Section 30, 'universal acceptance' could here mean 'acceptance by everyone and by any other number of people'.

[379] 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), 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.

[380] 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).

[381] We should not assume that, if everyone accepted some moral principle, everyone would always act upon it. But you should assume that, if I accept the Numbers Principle, I shall save the five rather than you. I would have no reason not to act on this principle.

[382] 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.

[383] 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 not be irrational.

[384] If I am wrong, since the optimific principles would require you to save the five strangers rather than your child, this case would raise an objection like the one that I am about to discuss.

[385] Bernard Williams, *Moral Luck* (Cambridge University Press, 1981) 18.

[386] 125.

[387] *The Collected Works of W.B. Yeats, Vol III,* edited by Douglas Archibald and William O'Donnell, 246.

[388] This argument was suggested to me by Robert Adams and Garrett Cullity.

[389] 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.

[390] ME, Book IV, Chapters III to V.

[391] 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 to will this world in self-interested terms, 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.

[392] It is easy to overlook our reasons to consider these other effects. Kagan may have thought that AC is the maxim whose being universally followed would make things go best. But it is not enough to consider only the effects of this maxim's being followed, since we shall then take into account only the effects of people's acts. This point does not apply when we ask which are the maxims or principles whose universal acceptance would make things go best.

[393] 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.

[394] That is mainly because, in asking which are the principles whose being universally followed would make things go best, we can ignore the various ways in which, when people try to make things go best, they often go astray, through miscalculation, self-deception, and the like.

[395] We can remember next that, to answer the New Ideal World Objection, Rule Consequentialists should appeal to the principles whose being accepted or followed by *any number of people* would make things go best. As I have said, that makes such theories closer to Act Consequentialism.

[396] In [the unwritten] Section 24 above.

[397] *The First Critique*, A 851 B 879.

[398] 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 makes 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).

[399] WWO, 11.

[400] 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.

[401] Though Kantian Rule Consequentialism has different versions, which may conflict, these conflicts are not between the Kantian and Rule Consequentialist parts of this view.