What We Could Rationally Will

DEREK PARFIT

THE TANNER LECTURES ON HUMAN VALUES

Delivered at

University of California at Berkeley
November 4, 5, and 6, 2002

DEREK PARFIT is senior research fellow at All Souls College, Oxford. He regularly teaches there and is also affiliated with New York University and Harvard. He was educated at Oxford and was a Harkness Fellow at Columbia and Harvard. He has been a visiting professor at Princeton, Temple, Rice, and the University of Colorado at Boulder, and is a fellow of the British Academy and of the American Academy of Arts and Sciences. He has made major contributions to our understanding of personal identity, philosophy of the mind, and ethics, and he is thought to be one of the most important moral philosophers of the past century. His many academic articles include "Personal Identity" (1971), "Overpopulation and the Quality of Life" (1986), "The Unimportance of Identity" (1995), and "Equality and Priority" (1997). *Rationality and Morality* and *Rediscovering Reasons* are forthcoming from Oxford University Press. His book *Reasons and Persons* (1984) has been described by Alan Ryan of *The Sunday Times* as "something close to a work of genius."

I. RATIONAL CONSENT

1

According to Immanuel Kant's best-loved statement of his supreme moral law, often called

> *the Formula of Humanity:* We must treat all rational beings, and the rationality of these beings, never merely as a means, but always as ends-in-themselves.[1]

In calling rational beings "ends-in-themselves," Kant means in part that we must never treat such beings in ways to which they could not consent. For example, when he explains the wrongness of lying promises, Kant writes:

> he whom I want to use for my own purposes by such a promise cannot possibly agree to my way of treating him.[2]

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....[3]

Onora O'Neill similarly writes:

In writing these lectures I have been greatly helped by my commentators, Allen Wood, Thomas Scanlon, Susan Wolf, and Samuel Scheffler. I am also grateful for comments from Henry Allison, Elizabeth Ashford, Robert Audi, Bruce Aune, Jonathan Bennett, John Broome, Ruth Chang, G. A. Cohen, Mary Coleman, Roger Crisp, Jonathan Dancy, Stephen Darwall, David Enoch, Allen Gibbard, Bradford Hooker, Thomas Hurka, Shelly Kagan, Frances Kamm, Patricia Kitcher, Martha Klein, Christine Korsgaard, Jefferson McMahan, Ingmar Persson, Thomas Pogge, Peter Railton, Andrews Reath, Sophia Reibetanz, Tamar Schapiro, Jerome Schneewind, Philip Stratton-Lake, Roger Sullivan, David Sussman, Larry Temkin, and some people to whom I apologize for forgetting their names.

[1] *The Groundwork* (henceforth *G*), pp. 428–29. (Page references are to the page numbers of the Prussian Academy edition, which are given in most English translations.)

[2] *G,* p. 430.

[3] Christine Korsgaard, *Creating the Kingdom of Ends* (henceforth *CKE*) (Cambridge University Press, 1996), p. 139.

> if we coerce or deceive others, their dissent, and so their genuine consent, is in principle ruled out. Here we do indeed use others, treating them as mere props or tools in our own projects.[4]

Korsgaard concludes:

> According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrong-doing to others.[5]

These remarks suggest this argument:

> It is wrong to treat people in any way to which they cannot possibly consent.
>
> People cannot possibly consent to being coerced or deceived.

Therefore

> Coercion and deception are always wrong.

It can be right, however, to treat people in ways to which they cannot possibly consent. When people are unconscious, for example, they cannot consent to life-saving surgery, but that does not make such surgery wrong. And we can rightly make some decisions on behalf of people whose whereabouts we don't know.

Kant's objection, Korsgaard might say, applies only to those acts whose nature makes consent impossible. Deception, unlike surgery, is such an act. To be able to consent to someone's treating us in some way, we must know what this person would be doing. And, if we knew that this person would be trying to deceive us, we could not be deceived.

But consider

> *Deadly Knowledge:* You ask me whether *Grey* committed some murder. I know that, unless I tell you a lie, you would come to believe truly that Grey is the murderer. Since you could not conceal that belief from Grey, he would then, to protect himself, murder you as well.

[4] Onora O'Neill, *Constructions of Reason* (henceforth *CR*) (Cambridge University Press, 1989) p. 111.

[5] *CKE,* p. 140.

If I told you the truth, and Grey murdered you, you could reasonably complain, with your dying breath, that I should have lied to you. It would be no defense to reply that I could not have deceived you with your consent. This deception *would be* like life-saving surgery on some unconscious person. I may know that, just as some unconscious person would consent to such surgery, if she could, you would consent to my life-saving lie. It is a merely technical problem that, if I tried to get your consent, that would make my act impossible. We could solve this problem if you could make yourself lose particular memories. I could then get your consent to my deceiving you, and you could make this deception possible by forgetting our conversation. I would be a moral idiot if I believed that, because you lack this ability to lose particular memories, my life-saving lie would be wrong. Since you would consent to being deceived, if you could, this lie is morally as innocent as the lies that might be needed to give someone a surprise party.

Similar claims apply to coercion. We can sometimes freely and rationally consent, in advance, to being coerced in some way. Before the discovery of anaesthetics, people gave such consent to being coerced during painful surgery. And it may be true, even while we are being coerced, that we would consent to this coercion, if we could. Most of us would vote in favour of everyone's being coerced to pay their taxes, and to obey certain laws. Since people can rationally consent to being deceived or coerced, these are not acts whose nature makes consent impossible.

Nor, I believe, does Kant's view imply that deception and coercion are always wrong. Kant claims:

> (A) It is wrong to treat people in any way to which they cannot possibly consent.[6]

There are two ways to understand this claim. "People cannot assent," 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.[7]

[6] Though Kant does not explicitly make this claim, he treats his remark about lying promises as presenting what he calls "the principle of other human beings" (*G*, p. 430). (A) is the most straightforward statement of that principle.

[7] *CR*, p. 110.

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 denied them the opportunity to give or refuse consent.

If people know how we are treating them, they can refuse consent in the *declarative* sense, by protesting against our act. Korsgaard and O'Neill use "consent" in a stronger, *act-affecting* sense. On their reading, Kant accepts

> *the Choice-Giving Principle:* It is wrong to deny people the opportunity to choose how we treat them.[8]

This principle is in one way too permissive. When we cannot communicate with other people, we cannot either give or deny these people the opportunity to choose how we treat them. So, even if we treat these people in ways to which we know that they would not consent, the Choice-Giving Principle would not condemn our acts. To cover such cases, we might turn to

> *the Veto Principle*: It is wrong to treat people in any way to which they either do not consent, or if they had the opportunity they would not consent.

These two principles condemn much more than deception and coercion. When we do not tell people what we are proposing to do, these people will not have the knowledge that is needed for consent, but we may not be deceiving them. And, when we act without people's consent, or in ways to which they would not consent, we may not be coercing them.

Though these principles have some appeal, they are clearly false. Suppose that, in

> *Earthquake,* you and some stranger, *Black,* are trapped in slowly collapsing wreckage. We are rescuers, who could save either Black's life or your leg.

[8] As Korsgaard writes: "People cannot consent when they are given no chance to do so.... 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,* pp. 138–39).

If I saved Black's life without your consent, I would not be acting wrongly. And I might not be acting wrongly if, without your consent, I marry the person you love, recommend that your book not be published, tell some truth you wish me to conceal, or take the last life-jacket that you were hoping to reach before me. There are countless cases of this kind. It is often right to treat people in ways to which we deny them the opportunity to consent, or to which we know that they would not consent.

These principles fail in another way. When our acts would significantly affect several people, we cannot possibly give more than one of these people the opportunity to choose how we act. And, in many of these cases, there is no possible act to which all those affected would consent. The Choice-Giving and Veto Principles mistakenly imply that, in such cases, we cannot avoid acting wrongly.

Kant would have rejected both these principles. Though he condemns acts to which other people cannot possibly consent, Kant did not believe that we should always let other people choose how we treat them. Kant's claims can be understood in a different way. When people say, "I cannot possibly consent to that proposal," they do not mean that they have been denied the opportunity to consent. They mean that they have decisive reasons to refuse consent. Kant is appealing, I suggest, to

> *the Rational Consent Principle:* It is wrong to treat people in ways to which they cannot, or could not, *rationally* consent.

One fact, I admit, counts in favour of Korsgaard's and O'Neill's *choice-giving* reading. When Kant presents his principle about consent, he is discussing an act that involves deception. People cannot consent to being deceived, in most cases, not because they have decisive reasons to refuse consent, but because they do not even know how they are being treated. But Kant then gives, as clearer examples of acts that conflict with his principle, attacks on people's property or freedom.[9] When people are robbed or coerced, they often know how they are being treated. And, in most of these cases, these people are treated in ways to which they could not rationally consent.

Other facts count in favour of this *rational consent* reading. Even when discussing deception, Kant writes that the person whom I deceive

[9] *G*, p. 430.

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

And he writes:

> 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.[10]

If Kant believed that we ought to let other people choose how we treat them, he would have no reason to claim that, for our acts to be justified, other people must be able to share our ends, or aims. If we are letting other people choose whether we shall act in some way, we need not ask whether these people could share our aim. Kant must mean that, when *we* are choosing whether we shall act in some way, we should ask whether other people could share our aim. And it would not be enough to ask whether these people could conceivably share our aim. We should ask whether these people could rationally share our aims, so that they could rationally consent to our treatment of them.

On this reading, Kant's view is much more plausible. We cannot let everyone choose how we treat them; nor we can treat everyone only in ways to which they either do consent, or, if they had the opportunity, they would consent. But we might be able to treat everyone only in ways to which, if they had the opportunity, they could rationally consent. And, if that is possible, we can plausibly believe that this is how we ought to act.

To apply the Rational Consent Principle, we must appeal to our beliefs about rationality and reasons. On one widely accepted view, people could not rationally consent to acts that would be bad for them. On another common view, people could not rationally consent to acts that would leave them less able to achieve their aims, or fulfil their desires.

If we accept either of these views, we must reject the Consent Principle. In *Earthquake,* for example, we can plausibly suppose that, if we saved Black's life rather than your leg, that would be worse for you, and would leave you less able to achieve your aims. These views would then imply that you could not rationally consent to our saving Black's life, so the Consent Principle would mistakenly condemn this act. There are countless cases of this kind. When people's interests or aims conflict, we

[10] *G,* p. 430.

cannot avoid treating some people in ways that would be worse for them, or would leave them less able to achieve their aims. In such cases, on these views, there would be no act to which everyone could rationally consent, so the Consent Principle would mistakenly imply that every possible act would be wrong.

On what I believe to be the true view—which I shall describe more fully later—-our reasons for acting are provided, not by our aims or desires, but by the facts that give us reasons to have these aims or desires. These are facts about what is relevantly worth achieving or doing. And, though we have strong reasons to care about our own well-being, we have reasons to care as much about some other things, such as the well-being of others.

We cannot possibly respond to all of these reasons, by wanting and trying to achieve whatever is worth achieving. And there are often no precise truths about the relative strengths of different reasons. Given these facts, we could often rationally choose any of several aims. And, when we choose our aims, we can respond to many reasons both from an impartial point of view and from our own personal point of view. That greatly widens the range of aims that we could rationally choose. One point is especially relevant here. We often have sufficient reasons to choose, or to do, either what would be impartially best or what would be best from our own point of view. In such cases, we could rationally make either choice, or act in either way.

These remarks assume that we could rationally choose whatever we have sufficient reasons to choose. That is not always true. First, we can have reasons of which we are unaware. For example, if I ask my doctor whether I have any reason to avoid eating certain foods, and he knows that walnuts would kill me, he should tell me that I do have such a reason. It is irrelevant that I don't yet know the fact that gives me this reason. We can also have false beliefs whose truth would give us reasons. When we are ignorant, or have false beliefs, it can be rational for us to do what we have no reason to do, and strong reasons not to do. To avoid these complications, I shall use "could rationally" to mean "could rationally, if we knew all the relevant facts." On these assumptions, we could rationally choose, or do, whatever the relevant facts give us sufficient reasons to choose or do.

On these assumptions about rationality and reasons, the Rational Consent Principle can be plausibly applied to many kinds of case. Suppose that, in *Earthquake,* you had the power to choose whether we save

Black's life or your leg. You would have sufficient reason, I believe, to make either choice. You could rationally choose that we save your leg, since that would be much better for you. But you would not be rationally required to make this choice. You could rationally choose instead that we save Black's life, since you could rationally regard Black's well-being as mattering as much as yours, and dying is much worse than losing a leg. Black could also rationally choose that we save her life. But if Black is young, and can expect to live a good life, she could not rationally choose that we let her die, so that we can save your leg. Black would not have sufficient reason to make so great a sacrifice. On these assumptions, the Consent Principle rightly implies that we ought to save Black's life rather than your leg. That is the only act to which you could both rationally consent.

Return next to Kant's own examples: lying promises, robbery, and coercion. Those who are treated in such ways, Kant claims, cannot share the agent's aim. As Korsgaard points out, that may not be so. We might be willing to give someone the money that she falsely promises to repay. And, if some robber is much poorer than us, we might rationally share this robber's aim, and be willing to give him what he steals. But even if we shared these people's aims, Kant might claim, we could not rationally consent to being either tricked or forced into contributing to these aims. We could not rationally consent to being treated in these ways without our consent.

Someone might now argue:

> It is wrong to treat people in any way to which they could not rationally consent.
>
> People could not rationally consent to being treated in any way without their consent.

Therefore

> It is wrong to treat people in any way to which they either do not or would not consent.[11]

If this argument were sound, the Rational Consent Principle would imply the Veto Principle.

When applied to some acts, this argument is plausible. Suppose that

[11] This objection was suggested to me by Ingmar Persson.

some rapist claims 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 acts, she did not in fact consent. And she could not have rationally consented to being treated in this way without her actual consent.

As this argument rightly assumes, when we ask whether it would be wrong to treat people in a certain way, it is often morally important whether, while they are being treated in this way, these people do in fact or would in fact consent. We should not ignore that question, by asking only whether these people *could* rationally consent. If we are treating people in some way to which, at the time, they do not consent, we might ask whether they could have rationally consented, at this time, to being treated in this way without their actual consent. But that question is confusing, since people could not rationally, at the same time, both give and refuse consent. To make our question clearer, we could give the Consent Principle this revised form:

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

People give *unconditional* consent when they cannot later withdraw this consent, if they change their mind.

This principle rightly condemns almost all cases of rape. People could seldom rationally give unconditional consent, in advance, to sexual acts to which, at the time, they would not 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 acts to which we could rationally give, in advance, such unconditional consent. Before the discovery of anaesthetics, people could rationally give such consent to painful surgery, because they knew that, while they were in pain, their judgment would be distorted. There are other grounds on which people could rationally give such consent. In many cases, for example, other people need to know that someone's consent is binding, and cannot be withdrawn. Suppose that, in *Earthquake,* once we had started to save Black's life rather than your leg, it would be dangerous for us to stop. You could then rationally say, "Go ahead and save Black's life, even if I later change my mind." That would be rational in part because, if you later changed your mind, that would not make it significantly worse for you to lose your leg. Losing a leg is, in this respect, unlike being raped.

As before, I am not claiming that you would be rationally required to give such unconditional consent. You could also rationally refuse consent, or rationally give consent only on condition that you did not later change your mind. My claim is only that you could rationally choose, unconditionally, that we save Black's life rather than your leg. You would have sufficient reason to make that choice.

Since we could often rationally give such consent to being treated in some way without our consent, the Rational Consent Principle does not imply the Veto Principle. These principles often conflict. And, when our acts would affect more than one person, as is true in most important cases, it is only the Consent Principle to which we can plausibly appeal. Only this principle implies that, whether or not you actually consent, we ought to save Black's life rather than your leg.

When our acts would affect only one person, the Consent Principle may be claimed to be too paternalistic. We may believe that, in such cases, we ought to treat this person only as she chooses, even if her choice is irrational. I shall return to these cases, and to some other ways in which the Consent Principle can be challenged. First, however, we should discuss the other half of Kant's Formula. Rational beings, Kant claims, must never be treated merely as a means.

2

Using people, it is often said, is wrong. But this claim needs to be explained. If we are climbing together, I might use you as a ladder, by standing on your shoulders. Or I might use you as a dictionary, by asking you how some word is spelt. Such ways of using people are not wrong. What is wrong, Kant claims, is *merely* using people. As people say, "You were *just using* me."

How can we use people without *merely* using them? Compare how two scientists might treat their laboratory animals. One scientist experiments in the ways that best produce the data that she wants, regardless of the pain she causes her animals. This scientist treats her animals merely as a means. Another scientist 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 way of using them is restricted by a concern for their well-being.

Similar claims apply to our treatment of each other. We treat people as a means when we make any use of their abilities, activities, or other features. We do not treat people *merely* as a means if there are important ways in which we would not harm these people, or in Kant's phrase "act against them," because we believe that such treatment would be wrong. If our treatment of someone is relevantly and significantly governed by some such moral belief, that is enough to make it true that we do not treat this person merely as a means. That is enough, moreover, even if our moral belief is false, and we are acting wrongly. When my mother traveled on a Chinese river in the 1930s, her boat was held up by bandits, whose moral principles permitted them to take only half of anyone's 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.

There are other ways in which, when we treat someone as a means, we may not be treating them *merely* as a means. That is not true, for example, when we are also choosing to bear some burden for this other person's sake. When we choose to bear such burdens, we may be acting out of love, or sympathy, rather than on some moral belief.

As these remarks imply, whether we are treating someone merely as a means depends on our underlying policies and attitudes. And that is in part a matter of what we would have done, if the facts had been different. If my first scientist did some experiments that caused her animals no pain, she would still be treating her animals merely as a means, since it would still be true that, if it were even slightly more convenient, she would inflict any amount of pain on them. Similarly, when we are treating people well, we may be treating them merely as a means. Thus we might treat someone well with the sole aim of discovering her weaknesses, or inheriting her wealth.

According to Kant's Formula of Humanity, we must never treat any rational being merely as a means. On a similar but wider view, we must never treat any conscious being merely as a means. Taken as claims about our attitudes, we ought to accept both these views. It is wrong to regard any rational or conscious being as a mere tool, which we are free to use in whatever way would best achieve our aims. But, when Kant claims that we must never treat rational beings merely as a means, he seems to mean that, in acting in this way, 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 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. When he pays for his coffee, he does not act wrongly. Or consider Kant's remark:

> he who intends to make a lying promise…wants to make use of another human being merely as a means.[12]

We could similarly say:

> he who intends to keep a promise for self-interested reasons wants to make use of another human being merely as a means.

Though such a person has the wrong attitude to others, he does not, in keeping his promise, act wrongly.

To avoid condemning such acts, we could revise Kant's claim. According to what we can call

> *the Mere Means Principle:* It is wrong to regard anyone merely as a means, and wrong to harm anyone, without their consent, in any way that treats them merely as a means.

Since my gangster does not harm the coffee seller, this principle does not condemn his act. Kant would have accepted this principle, though it expresses only part of his view.

We can now combine the two halves of Kant's Formula. We do not treat someone merely as a means if our treatment of this person is relevantly governed by some significant moral constraint. The Rational Consent Principle is one such constraint. We do not treat people merely as a means if we would never treat them in any way to which they could not rationally consent. And, if our treatment of others is governed by this constraint, that is part of what Kant means by our treating others as ends-in-themselves.

Consider next these examples:

[12] *G,* p. 429.

> *Lifeboat:* A single person, *White,* is stranded on one rock, and five people are stranded on another. Before the rising tide covers both rocks, we could use a lifeboat to save either White or the five.
>
> *Tunnel:* A runaway train is headed for a tunnel, in which it would kill the same five people. As bystanders, we could save these people's lives by switching the points, thereby redirecting the train into another tunnel. Unfortunately, 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. Our 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 we save the five, White would die. But White's death would be differently related to our saving of the five. In *Lifeboat,* we would fail to save White so that, in the time available, we could save the five. In *Tunnel,* we would save the five in a way whose foreseen side-effect is that we kill White. In *Bridge,* we 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 relevant differences between them. (Similar claims apply to all my imagined cases.)

Of those who have considered these cases, almost everyone believes that, in *Lifeboat,* we either may or should save the five. If our duty not to kill outweighs our duty to save lives, as most of us assume, it may seem that, in *Tunnel,* it would be wrong for us to save the five. Most people believe, however, that our duty not to kill does not here have such priority, since it would not be wrong to redirect the train so that it kills White rather than the five. Most of these people also believe that, though it would not be wrong to kill White as a side-effect of redirecting the train, it *would* be wrong, in *Bridge,* to kill White as a means of stopping the train. Some other people reject this distinction, believing that in all three cases we ought to save as many lives as possible. My first aim here is not to resolve this disagreement, but to ask what is implied by the Consent and Mere Means Principles.

Suppose that, in deciding how to act in *Bridge,* we apply these principles. We could argue:

> On the Rational Consent Principle, we ought to treat people only in ways to which they could rationally consent.
>
> White could rationally consent to our killing her as a means of saving the five.

Therefore

> Even if White would not in fact consent, the Consent Principle permits this act.
>
> We do not treat people merely as a means if our treatment of them is governed by the Consent Principle.

Therefore

> The Mere Means Principle permits this act.

This argument, I believe, is sound. It may be wrong to kill one person, without her consent, as a means of saving several others. But that is not implied by these Kantian principles.

Of those who are not convinced by this argument, some would reject its second premise, denying that White could rationally consent to being killed as a means. And some would reject its third premise, giving a different account of what it is to treat people merely as a means.

3

I have claimed that, in *Earthquake,* if the choice were yours, you would have sufficient reason to save either your leg or Black's life. Since you could rationally act in either way, you could rationally consent to our saving Black's life rather than your leg. I believe that, in *Lifeboat,* similar claims apply. If the choice were White's, she could rationally save her own life, but she could also rationally save the five rather than herself, and she could rationally consent to our doing that. Since White could rationally give such consent, the Consent Principle rightly permits us to save the five, whether or not White actually consents.

Return next to *Tunnel.* Most of us would believe that, in this case, it would not be wrong for us to redirect the train, so that it would kill

White rather than the five. For the Consent Principle to permit this act, it must be true that White could rationally consent to being treated in this way. And that, I believe, is true. From White's point of view, there is no relevant difference between *Lifeboat* and *Tunnel.* White could rationally save the five rather than herself, and it makes no difference whether she would save the five by redirecting the train, so that it kills her instead. This way of dying would be no worse for White. And, since White could rationally redirect the train, she could rationally consent to our doing that.

Similar claims apply to *Bridge.* White could rationally jump in front of the train, so that it would kill her rather than the five. *Bridge* is not relevantly different from *Tunnel.* In both cases, White could rationally kill herself with an act that would save the five; and she would have no reason to prefer to kill herself as a side-effect of saving the five, rather than as a means. Since White could rationally kill herself as a means of saving the five, she could rationally consent to *our* killing her as a means. So the Consent Principle also permits this act. And, since White *could* rationally consent to this act, this principle permits this act even if White would not in fact consent.

The Rational Consent Principle here fails to support a widely held view. Many people would believe that, in killing White as a means without her consent, we would be acting wrongly. This view may be justified. But, to defend this view, we cannot appeal to the Consent Principle.

Nor can we appeal to the Mere Means Principle. I have claimed that

> (A) we do not treat someone merely as a means if our treatment of this person is relevantly governed by a significant moral constraint.

Even if we killed White, without her consent, as a means of saving the five, our treatment of White may be governed by the Rational Consent Principle. If that is true, we would not be treating White merely as a means. If this act would be wrong, this wrongness must be explained in some other way.

It may seem that, in making these claims, I must be misunderstanding what is involved in treating people merely as a means. Many people have believed that, to explain the wrongness of injuring one person as a means of benefiting others, we can appeal to the Mere Means Principle.

Robert Nozick, for example, writes:

> Side constraints upon action reflect the underlying Kantian principle that individuals are ends and not merely means; they may not be sacrificed or used for the achieving of other ends without their consent.[13]

Nozick here assumes that

> (B) if we harm someone, without her consent, as a means of achieving some aim, that is enough to make it true that we are treating this person merely as a means.

We ought, I believe, to reject this view. Consider

> *Accident:* Some malfunctioning machine threatens to kill you. You cannot protect yourself except by injuring *Blue,* without her consent. By causing Blue to lose one finger, you could save your life, but you would become completely paralysed. If you also caused Blue to lose a second finger, you would be unharmed.

Suppose you believe that, even to avoid becoming paralysed, it would be wrong for you to destroy Blue's second finger. Only the saving of a life, you assume, could justify inflicting such an injury. Acting on this belief, you save your life by causing Blue to lose one finger.

Your act harms Blue, without her consent, as a means of achieving your aim. On Nozick's view, you are treating Blue merely as a means. That is clearly false. If you were treating Blue merely as a means, you would cause her to lose two fingers, since you would thereby save yourself from becoming paralysed. We do not treat someone merely as a means if we allow ourselves to bear a great burden, so as to avoid imposing a much smaller burden on this other person.

Nozick might reply that, though you are not treating Blue merely as a means, that is because you are limiting the harm that you impose on Blue, in a way that is worse for you, or less effectively achieves your aim. That would not be true, in *Bridge,* if we killed White as a means of saving the five. We would have acted in the very same way, even if we had regarded White as a mere means. This may seem enough to justify the charge that, in acting in this way, we would be treating White merely as a means. On this suggestion,

> (C) we treat someone merely as a means if

[13] Robert Nozick, *Anarchy, State and Utopia* (Blackwell, 1974), p. 31.

> we harm this person, without their consent, as a means of achieving some aim,

and

> we do not limit the harm that we impose, in a way that makes our act significantly less effective in achieving our aim.

This account is also, I believe, mistaken. For us to be able to deny that we are treating someone merely as a means, on the ground that I have described, it must indeed be true that we *would* limit the harm that we imposed on this person, even if that would cause our aims to be significantly worse achieved. But that is enough. If we don't act in this way, because no such act is possible, that does not imply that we are treating this person merely as a means.

Suppose again that, in *Accident,* you have decided to save your life by causing Blue to lose only one finger, thereby letting yourself become paralysed. (C) allows that, in acting in this way, you would not be treating Blue merely as a means. Suppose next that, before you can act, the situation changes, since you are no longer threatened with paralysis. When you save your life, at the cost of Blue's finger, you are not now limiting the harm that you impose on Blue, so (C) implies that you are now treating Blue merely as a means. That is an indefensible distinction. It is still true that you would let yourself be paralysed rather than destroying Blue's second finger. It cannot make a moral difference that this way of acting has now become impossible. Nor could it make a difference if this act was never possible. If you would have let yourself be paralysed, rather than imposing a much smaller injury on Blue, that is enough to make it true that you are not treating Blue merely as a means.

For a simpler case, return to my scientist who, in nearly all of her experiments, uses less effective methods, so as to avoid causing her laboratory animals any pain. Suppose that, in one experiment, this scientist uses the most effective method, because this method causes no pain. It would be most implausible to claim that, in this one experiment, this scientist treats her animals merely as a means.

I have not claimed that, in *Accident,* you could justifiably save your life by destroying one of Blue's fingers. My point is only that, even if this act is wrong, you would not be treating Blue merely as a means. But it may help to illustrate this point with an act that most of us would believe to be justified. Suppose that, in

> *Catastrophe,* some nuclear power station is about to explode, in a way that would kill a million people. This explosion cannot be prevented except by your killing an innocent person, *Green.* Since Green is underground, she would survive the explosion.

If you would have killed yourself rather than Green, you would not, in killing her, be treating her merely as a means.

Similar claims apply to *Bridge.* Suppose that, in a variant of this case, 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 are saved. I intend, however, to try to save White's life, by running to the track so that I can throw myself in front of the train. It is clear that, if I succeed, I would not be treating White merely as a means. I would be killing myself for White's sake. And it would make no moral difference if I failed to reach the track in time. Nor would it make a difference if, though I would have sacrificed my life to avoid killing White, this was never possible.

We can now return to questions about which acts are wrong. In the sentence quoted above, Nozick claims:

> (D) If we harm people, without their consent, as a means of achieving some aim, we are acting wrongly, because we are treating these people merely as a means.

We ought, I have argued, to reject part of this claim. Even if we harm people in such a way, we may not be treating these people merely as a means. Nozick might still claim that

> (E) it is wrong to harm people, without their consent, as a means of achieving some aim.

This claim, however, is too strong. Though many such acts are wrong, there are also many exceptions. (E) should become

> *the Harm Principle:* It is wrong to harm people, without their consent, unless our act is the only way to achieve some good aim, and the harm we cause is not too great, or *disproportionate,* given the goodness of this aim.

Most of us would believe that, in *Catastrophe,* you could permissibly kill Green, without her consent. On this view, though Green's death is a great harm, this harm is not disproportionate, given the fact that your

act is the only way in which a million other people's lives can be saved. Some people would reject this view. According to Judith Thomson, for example, each person has absolute rights not to be killed or seriously injured, however many other people's lives such an act would save.[14] But, even on this more restrictive view, we can justifiably impose some lesser harms on people, without their consent, as a means of achieving certain aims. Thomson claims that, if it were the only way to save one person's life, we could permissibly bruise some other person's leg, causing her "a mild, short-lasting pain."[15]

According to Kant's Formula of Humanity, we must never treat any rational being merely as a means. This claim, I have argued, cannot be directly applied to our acts. When my gangster pays for his coffee, he treats the coffee seller merely as a means, but he is not acting wrongly. To meet this objection, I suggested, we might revise Kant's claim. According to

> *the Mere Means Principle:* It is wrong to regard anyone merely as a means, and wrong to harm anyone, without their consent, in any way that treats them merely as a means.

But consider

> *Accident II:* My gangster has a child, whose life is threatened by some malfunctioning machine. This man knows that, to save his child's life, he must bruise *Grey's* leg, without her consent, causing her a mild, short-lasting pain.

Since this harm is not disproportionate to the saving of a life, the Harm Principle permits my gangster to act in this way. But, since this man regards Grey as a mere means, he would be harming Grey, without her consent, and he would be treating Grey merely as a means. So, on the Mere Means Principle, his act is wrong. That is clearly false. Though this gangster has the wrong attitude to Grey, he could justifiably save his child's life by imposing this small harm on Grey.

We might revise the Mere Means Principle, so that it applies only to self-interested acts. On this view, though my gangster could justifiably save his child's life in this way, he could not justifiably save himself.

[14] Judith Thomson, *The Realm of Rights* (Harvard University Press, 1990), pp. 166–68. Thomson adds: "Where the numbers get very large, however, some people start to feel nervous. Hundreds! Billions! The whole population of Asia!"

[15] Ibid., p. 153.

Since this man regards Grey as a mere means, it would be wrong for him to save his own life by imposing even the slightest harm on Grey.

We ought, I believe, to reject this suggestion. It is better to keep Kant's distinction between whether some act is right or wrong, and whether, given the agent's motives and his other attitudes, this act has moral worth. When my gangster pays for his coffee, or saves his life in some way that would be justifiable for others, he does not act wrongly. But, since he regards other people as mere means, his acts have no moral worth.

On this view, the Mere Means Principle should be restricted to our attitudes. It is wrong to regard any rational being as a mere means. But, when we ask how much harm we could permissibly impose on others, we should appeal instead to the Harm Principle. It is often wrong to harm other people, without their consent. But what makes such acts wrong is not that we are harming others in ways that treat them merely as a means. Such acts are wrong when, and because, the harm that we impose on others cannot be claimed both to be necessary to the achievement of some good aim and to be harm that is not too great, given the goodness of this aim.

4

According to the other half of Kant's Formula of Humanity, we must treat all rational beings as ends-in-themselves. Part of what Kant means, I have claimed, is that we must follow the Consent Principle. We must treat people only in ways to which they could rationally consent, because they could rationally share our aims. Kant's Formula is often read in other ways. On Allen Wood's account, what Kant's Formula

> fundamentally demands of our actions is…that they express proper respect or reverence for the worth of humanity.[16]

It is impossible, Wood claims, to overestimate this idea's importance. Of the sixteen duties discussed in Kant's *Doctrine of Virtue,*[17] Kant defends eleven with appeals to the dignity or worth of "rational nature," or rational beings.

[16] Allen Wood, *Kant's Ethical Thought* (henceforth *KET*) (Cambridge University Press, 1999), p. 141.

[17] In the *Metaphysics of Morals* (henceforth *MM*).

Such remarks suggest

> *the Respect Principle:* It is wrong to treat people in ways that do not express respect for their dignity or worth as rational beings.

Taken in its ordinary sense, this principle mistakenly condemns many permissible acts. When my gangster pays for his coffee, he is not expressing his respect for the coffee seller's dignity as a rational being. He does not respect the coffee seller any more than he respects the vending machine. But he is not acting wrongly.

Wood interprets this principle in a less ordinary sense. He writes:

> We will be misled here…if we think of "valuing" or "respecting" humanity as some subjective state of mind…in dealing honestly with you, I treat you with respect whatever my inner state may be.[18]

This reading seems to me too narrow. Kant's greatness consists in part in the intensity of his belief that we must regard everyone, not as a mere thing or tool, but with unconditional respect. Rather than taking Kant's Formula to make no claim about our attitudes, it is better to understand this formula so that it covers both attitudes and acts. The Respect Principle could become

> *RP:* It is wrong to regard anyone with no respect for this person's dignity or worth as a rational being, and wrong to treat anyone in ways that are incompatible with such respect.

When my gangster pays for his coffee, his act is compatible with respect for the coffee seller's dignity, so RP condemns his attitude, but permits his act.

Is this a helpful principle, when applied to acts? Wood states this principle as

> FH: Always respect humanity, in one's own person as well as that of another, as an end-in-itself.

It may be objected, Wood writes, that FH

> is so empty and vague, and that its meaning is so flexible and disputable, that no determinate conclusions can be drawn from it.[19]

Wood replies:

[18] *KET,* p. 117.

[19] *KET,* p. 153.

> The meaning of FH is clear and determinate because the concepts of humanity (or rational nature) and existent end in itself are both reasonably clear and determinate.

Even if these concepts were reasonably clear, it would also need to be sufficiently clear what counts as *respecting* humanity or rational nature as an end-in-itself. And, that, I believe, is far from clear.

As one example, consider Kant's claim that it is wrong to shorten our lives to avoid suffering, since such an act would "degrade the humanity in our own person."[20] Kant believed that, as rational beings, we have an exalted status, with a worth that is above any price: a status that we must not abandon, or treat with disrespect, merely for the sake of a lower aim, the avoidance of suffering. We can accept part of Kant's thinking here. If someone chooses to endure great pain for the sake of continuing to use her rational powers—as Sigmund Freud refused morphine while he was dying so that he could still think clearly—such fortitude is admirable. But that does not show that, if Freud had taken morphine, or hastened his death, his act would have shown disrespect for his status as a rational being. Of those who have shown such fortitude when suffering, some were Stoics, who believed that we have more dignity if we make a rational choice about when and how we die. That is more plausible than the view that, in making such choices, we fail to respect our dignity as rational beings. Kant himself said:

> in the Stoic's principle concerning suicide there lay much sublimity of soul: that we may depart from life as we leave a smoky room.[21]

As Wood writes when he rejects Kant's claims about suicide,

> We disagree here because we justifiably believe we know more about what respect for humanity requires in these matters.[22]

Consider next Kant's claim that, in telling a lie even "to achieve a really good end"—such as deceiving a would-be murderer about where his intended victim is—the liar "violates the dignity of humanity in his

[20] *MM,* p. 423.

[21] *Lectures in Ethics* (henceforth *Lectures*), translated by Peter Heath (Cambridge University Press, 1997) p. 369.

[22] *Lectures,* p. 153.

own person."[23] Kant similarly claims that, in giving ourselves sexual pleasure, we defile our humanity.[24] In rejecting both these claims, we must again disagree with Kant about what respecting humanity requires. Kant is right to insist that we should always regard people with respect for their dignity as rational beings. But this claim is too vague to help us to answer difficult questions about the wrongness of acts.

There is another way to interpret this part of Kant's view. According to Kant's Formula, it is not only every rational being but also the rationality of these beings that we must always treat as ends-in-themselves. Our rationality, Kant claims, has a worth that is unconditional, and "exalted above all price." Thomas Hill takes such claims to imply

> a rather substantive value judgment with significant practical implications.... Kant's view implies that pleasure and the alleviation of pain, even gross misery, have mere price, never to be placed above the value of rationality in persons.[25]

On this view, Hill suggests, we act wrongly if we do anything that would reduce anyone's rationality, or interfere with rational activities, merely for the sake of preventing suffering.

According to Cardinal John Henry Newman, though both sin and pain are bad, sin is infinitely worse, so that, if all humankind suffered extremest agony for all eternity, that would be less bad than if one venial sin were committed. Though this view is horrific, we can understand why it has been held. We can see how sin could seem infinitely worse than pain.[26] If Kant had claimed that the relief of suffering should never be placed above a good will, his view would be as understandable as Newman's. But, on Hill's reading, what Kant claims to have such infinitely greater value is not moral goodness, but rationality. On this view, it would be wrong to save mankind from extremest agony at the cost of making one person less able to play chess or solve crossword puzzles. It is hard to believe that Kant held such a view.

[23] *MM,* pp. 430, 429.

[24] *MM,* pp. 424–25.

[25] Thomas Hill, *Dignity and Practical Reason* (Cornell University Press, 1992), pp. 55–77.

[26] Cardinal Henry Newman, *Certain Difficulties Felt by Anglicans in Catholic Teaching* (London, 1885), vol. 1, p. 204.

5

Return now to the Rational Consent Principle, which I believe to be the most valuable part of Kant's Formula. In applying this principle, I have assumed one view about rationality and reasons. Many people, as I have said, hold other views.

While reasons are provided by the facts, what we could rationally do depends on our beliefs. If we know the relevant facts, as I shall suppose throughout these lectures, we are rational insofar as we respond to reasons.

There are two main kinds of view about practical reasons. According to *desire-based* or *aim-based* theories, these reasons are provided by our present desires or aims. What we have most reason to do is whatever would best fulfil these desires or aims.

According to *value-based theories,* reasons are provided by facts about what is relevantly good, or worth achieving. Such theories can differ greatly. According to *the Self-interest Theory,* it is our own well-being that is most worth achieving, and we have most reason to do whatever would be best for ourselves. According to another simple theory, which we can call *rationalist utilitarianism,* we always have most reason to do whatever would, on the whole, benefit people most.

These two theories are *monistic,* appealing to only one kind of reason, and giving us a single ultimate aim. Henry Sidgwick combined these theories. On his view, we always have sufficient reason both to do whatever would be best for ourselves and to do whatever would benefit people most. When one act would be best for ourselves, but another would benefit people most, we have sufficient reason to act in either way. Either act would be rational. This view Sidgwick called "the Dualism of Practical Reason."

We ought, I believe, to accept a wider, pluralistic view. Though we have strong reasons to care about our own and other people's well-being, there are other practical reasons, and other things that are relevantly good, or worth achieving. Nor should we assume that only outcomes are relevantly good. Some things are worth doing for their own sakes. This view is pluralistic in another way. Though it is often clear what we have most reason to want, and to do, there are also many cases in which we have sufficient reason to have, and to try to achieve, any of two or more conflicting aims. In such cases, it would be rational for us to try to achieve any of these aims. We can call this a *wide value-based* view.

Many people now accept desire-based theories. In economics and the other social sciences, rationality is often defined in a desire-based or preference-based way. Of those who accept desire-based theories, many also accept the Self-interest Theory, since these people falsely assume that each of us would always care most about our own well-being.

If so many people believe that *all* reasons are provided by desires, how could it be true that, as I have claimed, *no* reasons are provided by desires? How could all these people be so mistaken?

One explanation is that, in most cases, these two kinds of theory partly agree. Even on value-based theories, we have some reason to fulfil most of our desires, since what we want is usually in some way worth achieving. But, though these theories agree that we have some reason to fulfil these desires, they make conflicting claims about what these reasons are, and how strong these reasons are. On desire-based theories, our reasons to fulfil these desires are provided by these desires. On value-based theories, these reasons are provided, not by our having these desires, but by the facts that give us reasons to have them. If some aim is worth achieving, we have a reason both to have this aim and to try to achieve it. Since our reason for acting is the same as our reason for having the desire on which we act, this desire is not itself part of this reason. And we would have this reason even if we didn't have this desire.

Second, even on value-based theories, there are certain other reasons that we *wouldn't* have if we didn't have certain desires. But though these reasons *depend* on our desires, they too are not *provided* by these desires. They are provided by other facts that depend on our having these desires. When we have some desire, for example, it may be true that this desire's fulfilment would give us pleasure, or that its non-fulfilment would be distressing, or distracting. In such cases, it would be these other facts, and not the fact that we had these desires, that gave us reasons to fulfil them.

On desire-based theories, we cannot have reasons to care about anything for its own sake. All reasons to have some desire must be provided by some desire. And this must be some *other* desire. We can have a reason to want some thing to happen if its happening would have effects that we want. But we cannot have any reason to have any intrinsic desire, or ultimate aim. We cannot have such reasons, for example, to want ourselves or others not to suffer or die.

This bleak view is seldom defended. Most desire-based theorists simply take it for granted that we cannot have reasons to want anything

for its own sake. There is, I think, one main argument for this view. Of those who hold this view, many are *naturalists,* who believe that there cannot be any irreducibly normative truths. These people give reductive accounts of desire-based reasons for acting. On these accounts, when we have a reason to act in some way, this normative truth is the same as the fact that this act would fulfil one of our desires, or the fact that, after informed deliberation, we would be motivated to act in this way. Desire-based reasons, so understood, merely involve causal or psychological facts. Value-based reasons cannot be so easily reduced to natural facts.

Reductive naturalism is, I believe, mistaken. But I cannot defend that belief here. So, if you accept naturalism, I must ask you to suppose that we can have reasons to care about some things for their own sake.

Though I call these reasons *value-based,* that is in a way misleading. As Thomas Scanlon claims, things are good or bad by having other properties that would, in certain contexts, give us certain reasons.[27] It is reason-giving properties or facts that are fundamental.

If we accept a desire-based theory, as I have said, we must reject the Consent Principle. It is clear that, in *Earthquake,* we ought to save Black's life rather than your leg. But on desire-based theories, if you preferred us to save your leg, you could not rationally consent to our saving Black's life, so the Consent Principle would mistakenly condemn this act. If we accept the Self-interest Theory, similar claims apply. On this view, if it would be worse for you if we failed to save your leg, you could not rationally consent to our saving Black's life. Similar claims apply to many other cases. If we accept either a desire-based theory or the Self-interest Theory, we must reject Kant's claim that we must treat people only in ways to which they could rationally consent. It is often right to treat people in ways to which, on these theories, these people could not rationally consent.

Suppose, however, that, as I believe, we ought to accept a wide value-based view. On this view, as I have argued, the Consent Principle rightly implies that we ought to save Black's life. And I believe, that, in many other kinds of cases, this principle has plausible implications.

Someone might now object: "Your claims about rational consent

[27] Thomas Scanlon, *What We Owe to Each Other* (Harvard University Press, 1998), pp. 95–100.

merely reflect your moral views. When you believe that people could rationally consent to being treated in certain ways, that is because you believe that these acts would be right. Those who disagree would claim that, in their opinion, people could not rationally consent to these acts."

If this objection were justified, the Consent Principle would be trivial. Whenever we asked whether people could rationally consent to some act, our answer would depend on whether we believed that this act was right. So this principle could not help us to decide which acts are right.

This objection is, I believe, mistaken. Remember first that, in applying this principle, we ask what people could rationally consent to, or choose, on grounds that do not include their beliefs about which acts are wrong.

Second, our beliefs about rationality may conflict with, or fail to support, our moral beliefs. Return to the comparison between *Tunnel* and *Bridge.* Many people would believe that, though it would not be wrong to save the five in a way whose side-effect is to kill White, it would be wrong to save the five by killing White. But these people should admit that, if White could rationally consent to being killed as a side-effect of our saving the five, White could also rationally consent to being killed as a means. White would have no reason to prefer one of these ways of being killed. If anything, it would be better for White to be killed as a means, since her death would then at least do some good. Though these people believe that it would be wrong to kill White as a means, they should admit that the Consent Principle does not condemn this act.

Since our beliefs about rational consent may conflict with our moral beliefs, the Consent Principle is far from trivial. This principle may help to support our moral beliefs. And, if it does that well enough, we could justifiably let this principle guide our beliefs, by revising some, and extending others.

In some cases, however, this principle has implications that may be hard to accept. That may be true in *Bridge.* Could the five rationally consent, on nonmoral grounds, to our failing to save their lives by killing White? It is not clear that they could. And, if they could not, the Consent Principle would not merely fail to condemn our killing White as a means. This principle would imply that we ought to kill White as a means. Many of us would find that hard to believe.

Return next to *Earthquake,* in which we could save either Black's life or your leg. While you could rationally consent to our saving Black's life, Black, I have claimed, could not rationally consent to our saving your leg. The Consent Principle gives the right answer here, since it is clear that we ought to save Black's life. But suppose that, in a variant of this case, it is *you* who could act in either of these ways. Could Black rationally consent to *your* saving your leg rather than Black's life?

The answer may be No. If the choice were Black's, she could not rationally save your leg rather than her own life. If Black could not rationally act in this way, it is not clear that she could rationally consent, on nonmoral grounds, to your acting in this way. And, if Black could not rationally give such consent, the Consent Principle implies that it would be wrong for you to save your leg rather than Black's life. We may find that hard to believe. While it is clear that *we* ought to save Black's life rather than your leg, since this act would cost us nothing, we may doubt that *you* ought to sacrifice your leg to save Black's life.

Suppose that, in a third version of this case, it is not you but your child whose leg is threatened. Could Black rationally consent to your saving your child's leg rather than Black's life? As before, it is not clear that Black could rationally give such consent. And, if she could not, the Consent Principle implies that you ought to save Black's life rather than your own child's leg. We may find that hard to believe.

If the Consent Principle conflicts too strongly with some of our moral beliefs, we could revise this principle, by weakening its claims. On this revised principle, if some act would treat people in ways to which they could not rationally consent, that counts strongly against this act, making it, in Ross's phrase, *prima facie* wrong. But, in some cases, such an act might be justified on other grounds. As a parent, you may have a special obligation to protect your own child from harm, and this might morally outweigh your reason to save Black's life. And, though you have no such obligation to protect yourself, there may be limits on the sacrifices that anyone has a duty to make for the greater good of others.

The Consent Principle is only part of Kant's Formula of Humanity. And Kant himself claims that, in trying to decide which acts are wrong, we do better to appeal, not to this formula, but to Kant's first statement of his Categorical Imperative, the Formula of Universal Law. That will be my subject in tomorrow's lecture.

II. UNIVERSAL LAWS

6

The rightness of our acts, Kant claims, depends on our *maxims,* by which he means, roughly, our intentions, policies, or underlying aims. Some of Kant's examples are: "Shorten my life to avoid suffering,"[28] "Let no insult pass unavenged,"[29] "Increase my wealth by every safe means,"[30] and "the maxim of self-love, or one's own happiness."[31]

According to what we can call Kant's

> *stated criterion of strict duties:* It is wrong to act on maxims that could not be universal laws.[32]

This criterion needs to be explained. In some passages, when Kant supposes that certain maxim are universal laws, he supposes that everyone is permitted to act on these maxims.[33] That may suggest that Kant's criterion is

> (A) It is wrong to act on some maxim unless it would be possible for everyone to be permitted to act upon it.

But Kant never appeals to (A). Nor would (A) be a helpful criterion, since it assumes that we have some other way of knowing which acts could be permitted.

O'Neill suggests that Kant's criterion is

[28] *G,* p. 422. In this version of the case, the person whom Kant discusses is merely sick of life, and in despair; but Kant is reported as saying that not even "the most excruciating pains and irremediable bodily suffering can give a man the authority to take his own life" (*Lectures,* p. 369).

[29] *Critique of Practical Reason* (henceforth *Second Critique)*, p. 19.

[30] Ibid., p. 27.

[31] Ibid., p. 34.

[32] *G,* p. 424, and surrounding text. Compare Kant's last statement of his Categorical Imperative: "We must act on maxims that can hold as universal laws" (*MM,* p. 225).

[33] 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,* p. 403); and he refers to "the universality of a law that everyone…could promise whatever he pleases with the intention of not keeping it" (*G,* p. 423). Similarly Kant refers elsewhere to "the law that everyone may deny a deposit which no one can prove has been made" (*Second Critique,* p. 27). And he writes of a maxim's being "a universal permissive law" (*MM,* p. 453).

> (B) It is wrong to act on some maxim unless it would be possible for everyone to accept this maxim.[34]

But, when Kant argues that certain maxims could not be universal laws, he does not claim that these maxims could not be universally accepted. He appeals to what would happen if these maxims were universally accepted.

Nor would (B) be a helpful criterion. If (B) used "possible" to mean "conceivable," this criterion would fail to condemn many wrong acts. We can easily conceive a world in which everyone accepts bad maxims, such as "Deceive, coerce, or injure others whenever that would benefit me." Such a world may be psychologically impossible, since there may be people who would be unable to accept these maxims. But, if (B) used "possible" in this sense, this criterion would fail in a different way, since there are many permissible or good maxims that some people would be psychologically unable to accept. Some of us, for example, could not accept the maxims of those who clean the windows of skyscrapers, or parachute from aeroplanes. And we have no reason to believe that maxims are more likely to be bad if they are harder to accept.

O'Neill also suggests that Kant's criterion is

> (C) It is wrong to act on some maxim unless it would be possible for everyone successfully to act upon it.[35]

This criterion, O'Neill argues, condemns deception and coercion, since deceivers and coercers make it impossible for their victims to act like them. But two people can deceive each other. And there can be mutual coercion. I might be coercing you, by making one credible threat, while you are coercing me, by making another.

[34] O'Neill writes: "I have interpreted FUL as a criterion for picking out maxims that could be universally adopted" (*CR,* p. 141), and she makes similar claims in many other passages.

[35] Though O'Neill often states Kant's criterion as requiring us to act only on maxims that could be universally adopted, or universally acted on, some of her arguments take Kant's requirement to be that these maxims could be universally successfully acted on. Thus she claims that deception is wrong because "[d]eception cannot be universally successfully practiced" (*CR,* p. 157). And she writes: "The reasonably foreseeable result of anything approaching universal commitment to coercion would ensure that there could not be universally available effective means to coerce: universal coercion is therefore an incoherent project" (*Autonomy and Trust in Bioethics,* pp. 86–87).

O'Neill is right, however, to claim that, if we appeal to (C), it is "remarkably easy" to derive significant moral conclusions.[36] Many wrong acts are condemned by this criterion. For example, we could not all successfully act on the maxim "Kill other people whenever that would benefit me." Some attempted murders would fail, and some successful murderers would be caught and punished. Similar claims apply to the maxims of self-interested deception and coercion. If we all acted on these maxims, some of us would fail to benefit ourselves.

Such arguments, though, are too easy. We could not all successfully act on the maxims "Rescue people who are in danger," "Avoid hurting other people's feelings," "Don't make decisions that I shall regret," or "Understand Kant's philosophy." (C) implies falsely that, in trying to achieve these aims, we would be acting wrongly.[37] As well as condemning many permissible acts, (C) has no plausibility. There is no reason to believe that, if we could not all successfully act on some maxim, no one should ever act upon it. Innocent or worthy aims can be hard to achieve. Nor is it wrong to make attempts some of which are bound to fail.

It makes a difference, O'Neill might answer, *why* we could not all succeed in acting on some maxim. Though we could not all succeed in rescuing people who are in danger, or avoiding hurting anyone's feelings, those who achieve these aims do not thereby make success impossible for others. That is what is wrong, O'Neill argues, with coercion. By coercing other people, we undercut their agency, thereby preventing them, "for at least some time," from acting in the same way as us.[38] This argument assumes

> (D) It is wrong to act on any maxim whose being acted on by some people would make some other people unable, for a time, successfully to act upon it.

This criterion, however, is much too strong. There are countless permissible acts that make some other people unable, for a time, to act

[36] *CR,* p. 95.

[37] O'Neill might reply that, even if it would be practically impossible for everyone to achieve these aims, such a world is conceivable. In the same way, however, it is conceivable that we could all deceive or coerce others.

[38] "A principle of coercion, whose enactment…undercuts the agency…of at least some others for at least some time, cannot be universally followed" (*CR,* p. 215). "It is not merely that victims do not in fact will the maxims of their…coercers: They are deliberately made unable to do so, or unable to do so for some period of time" (*CR,* p. 133).

successfully in the same way. It is not always wrong to buy the last ticket to some performance, or to use the last available tennis court.

O'Neill might turn to

> (E) It is wrong to act on any maxim whose being acted on by some people would prevent some others from ever successfully acting on it.

Though this criterion would not condemn temporary coercion, it would condemn killing. Murderers make their victims unable ever to commit murder.

This criterion is also much too strong. If we succeed in acting on the maxim "Win an Olympic gold medal," we make it impossible for others to succeed, so (E) implies that we have acted wrongly. O'Neill might accept that conclusion, since she remarks that, on Kant's view, it is wrong to play competitive games with the overriding aim of winning. But (E) condemns countless other permissible acts. It is not wrong to act on the maxim "Become a doctor." But, since we cannot all become doctors, those who succeed in achieving this aim make it impossible for some others to succeed. Or consider maxims like "Discover the causes of cancer" and "Find someone with whom I can happily live my life." It is not wrong to try to make some discovery, even if, by succeeding, we would make it impossible for others to succeed. Nor is it wrong to marry the only person with whom someone else could happily live their life.

We can now return to Kant. O'Neill reasonably assumes that, when Kant claims that certain maxims could not be universal laws, he means that we could not all accept these maxims, and successfully act upon them. But, when Kant condemns the maxim "Make lying promises," he does not claim that we could not *all* succeed in acting on this maxim. He claims that, if we all accepted this maxim, or we all believed that we were permitted to act upon it, *none* of us could successfully act upon it.[39] So we can say that, on Kant's

> *actual criterion of strict duties:* It is wrong to act on maxims whose being universally accepted, or believed to be permissible, would make it impossible for anyone successfully to act upon them.

[39] Kant more often appeals to the effects of our *being* permitted to act on some maxim. But these effects would be produced by our *believing* that such acts are permitted. And, at one point, Kant writes, "if everyone…*considered himself authorized* to shorten his life as soon as he was thoroughly weary of it" (*Second Critique,* p. 69; emphasis added).

All strict duties, Kant claims, depend on this principle. Is that so?

Consider first the maxims "Kill, injure, and coerce others whenever that would benefit me." If we all accepted and acted on these maxims, that would not make it impossible for any such act to succeed. So Kant's criterion does not condemn such acts.

Consider next lying. Kant's criterion, Barbara Herman writes,

> seems adequate for maxims of deception.... Universal deception would be held by Kant to make speech and thus deception impossible.[40]

Korsgaard similarly writes:

> lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive.[41]

On Kant's view, however, the wrongness of an act depends on the agent's maxim, and few liars act on the maxim "Always lie." As Kant assumes, most liars act on the maxim "Lie when that would benefit me."[42] Kant's criterion condemns such lies only if, in a world of self-interested liars, no such lies could succeed. And that would not be true. It would seldom be in our interests to deceive others. And, of the small proportion of cases in which deception would benefit us, there is only a small proportion in which lying would be likely to deceive. There are, in contrast, many cases in which we benefit from speaking truly. So, even if we were all self-interested liars, most of our statements would be true. Most people would know that fact. And, since we could not always tell when other people were lying, some lies would be believed, and would achieve the liar's aim.

To explain why theft is wrong, Kant claims:

> 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.[43]

[40] Barbara Herman, *The Practice of Moral Judgment* (henceforth *PMJ*) (Harvard University Press, 1993), p. 119.

[41] *CKE,* p. 136.

[42] "Suppose that someone were to have the maxim that he might tell an untruth whenever he could thereby obtain great advantage" (*Lectures,* p. 264).

[43] *Lectures,* p. 232.

As before, the relevant maxim isn't "Always steal." Most thieves act on the maxim "Steal when that would benefit me." If this maxim were universally accepted, 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.

Kant's criterion, I have argued, fails to condemn most of the acts that are most clearly wrong. This criterion does not condemn self-interested killing, injuring, coercing, lying, and stealing.

This may suggest that Kant's criterion condemns nothing. But we have not yet considered Kant's best example. This is the maxim "Make lying promises whenever that would benefit me."[44] Kant claims that, if such lying promises were universally believed to be permissible, that would make them impossible. 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.[45]

This prediction may be true. Kant could also claim that, if we all believed that there was nothing wrong in making lying promises, whenever that would be better for ourselves, we would not even understand the concept of a promise. In such a world, the practice of promising would not exist.

Now that we have found one kind of act that Kant's criterion condemns, we can ask whether this criterion is plausible. Kant's criterion is, in part,

> (F) It is wrong to act on maxims whose being universally believed to be permissible would make it impossible for anyone successfully to act upon them.

This claim condemns those acts whose success depends on other people's refraining from such acts, because they believe them to be wrong. And (F) may seem to condemn these acts for the right reason. These acts are wrong, we may think, because they exploit the conscientious self-restraint of others in ways that, if universal, would undermine the exis-

[44] *G*, pp. 402–3 and 422.

[45] *G*, p. 422.

tence of valuable practices, such as the practice of making and keeping promises.

Kant's criterion, however, seems more plausible than it really is. Kant applies (F) to acts that many people believe to be wrong; and, of the acts that are widely believed to be wrong, many *are* wrong. But Kant's criterion is not intended merely to appeal to our moral beliefs. To judge this criterion, we should turn to actual or imagined cases in which people's moral beliefs are mistaken.

Suppose that, because my nation's tyrannical ruler is waging an unjust war, I adopt the maxim "Kill this tyrant to end this war." Most of my fellow-citizens would be appalled by such an act, since they accept Kant's view that we should never try to overthrow any established government. Because this tyrant's bodyguards know that nearly everyone accepts this view, they do not expect any attempt on the tyrant's life. That makes them inattentive, which allows me to kill this tyrant, thereby ending this war. If everyone believed that my maxim was permissible, the bodyguards would be more alert, making it impossible for any such attempt to succeed. On these assumptions, (F) condemns my act.

This example counts against this criterion. First, though Kant believed that killing our nation's ruler is always wrong,[46] that belief is false. It would have been right for a German to kill Hitler during the Second World War. Second, even if tyrannicide were always wrong, (F) could not provide the reason why. The objection to tyrannicide could not be that, if we all believed that tyrannicide could be justified, that would make it impossible for any such act to succeed.

Suppose next that, during this war, some German civilian knows that Jews are being rounded up and killed. This person acts on the maxim "Tell lies to the police when I could thereby help any Jews to escape." It might also have been true that, if all Germans had believed that such lies were permissible, that would have made it impossible for anyone to help Jews in this way. The German policemen would not have believed what civilians told them about the whereabouts of Jews. On these assumptions, (F) condemns this life-saving act.

Kant might have accepted this conclusion, since he condemned lying to a would-be murderer about where his intended victim is. But such lies are clearly justified. And, in this example, (F) has no

[46] *MM*, p. 320.

plausibility. It is no objection to this way of saving people's lives that, if we all believed such acts to be permissible, that would make them impossible.

This example is intentionally similar to that of Kant's lying promise. This promise succeeds because there are many people who can be trusted to keep their promises, since they believe that breaking promises is wrong. If everyone was known to believe that lying promises are not wrong, that would make it impossible, Kant claims, for anyone to act successfully on this lying promiser's maxim. In the same way, when this German civilian lies to help some Jews to escape, this act succeeds because there are many people who can be trusted not to lie to the police, since they believe that such lies are wrong. If everyone was known to believe that such lies are not wrong, that would make it impossible for anyone to act successfully on this person's life-saving maxim. The difference between these maxims is only in what these lies are intended to achieve; and this difference is ignored by (F).

Suppose next that German soldiers of this period could be relied upon to obey orders, because they believed that disobedience would be wrong. That might have allowed some soldier to act on the maxim "Disobey orders when that would help any Jews to escape." We can also suppose that, if all German soldiers had been known to believe that such disobedience was permissible, their officers would not have given orders whose being disobeyed would allow Jews to escape. (F) would then mistakenly condemn this soldier's act.

As these cases show, (F) is wholly unacceptable. This criterion condemns some acts that are clearly right; and though it condemns some wrong acts, it condemns these acts for the wrong reason.

Kant's criterion is also, in part,

> (G) It is wrong to act on maxims whose being universally accepted, or acted upon, would make it impossible for anyone successfully to act upon them.

There are many maxims that, even if they were universally believed to be permissible, would not be universally accepted. Acting on such maxims, though not condemned by (F), might be condemned by (G).

According to some writers, Kant's criterion mistakenly condemns several good maxims, such as "Give to the poor" and "Refuse to accept bribes." If these maxims were universally accepted, that would make it impossible to act upon them, since there would cease to be any poor

people, and no one would offer bribes. That could not show, these objectors claim, that acting on these maxims is wrong. Korsgaard answers this objection well. If we all accepted the maxim "Give to the poor," we would be trying to abolish poverty. That is our maxim's aim. If this maxim's universal acceptance made it impossible to act upon it, because poverty would be abolished, that would not defeat but achieve this aim.[47]

There are other cases, though, to which such claims do not apply. Consider the men who accepted codes of honour, like the one that led Pushkin to his death. Some of these men accepted the maxim "Fight duels to preserve my honour, but always shoot to miss." If everyone had accepted this maxim, the practice of duelling would have become farcical, and would not have survived. That would have defeated this maxim's aim. It may seem that (G) is right to condemn this maxim, since duelling is wrong. But (G) does not condemn the maxim "Fight duels to preserve my honour, and always shoot to kill." And, of these maxims, the second is clearly worse. As that shows, (G) condemns the first maxim for the wrong reason. It is no objection to this maxim that, if it were universally accepted, the practice of duelling would not survive.

Turn next 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, none of us would achieve our aims. Cakes would never get eaten, conversations would never get started, and journeys on narrow roads 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 (G) condemns this maxim, in a way that is clearly mistaken.

It might next be said that, of the claims that I have rejected, some could be defended in a weaker form. It goes too far to claim that, if we were all self-interested, speech would be impossible, and there would be no property. But these claims remind us that some valuable practices and institutions depend in part, for their full effectiveness, on moral motivation.

[47] *CKE,* pp. 86–87.

This reply, however, cannot defend Kant's criterion. Since Kant condemns maxims that could not be universal laws, his criterion cannot be given a more moderate reading. This may be why some writers move between claims at opposite extremes. For example, Herman writes that Kant's criterion "seems adequate" for maxims of deception and coercion. But, while she condemns maxims of deception with the claim that, if they were universally accepted, *no one* could successfully act upon them, she condemns maxims of coercion with the claim that we could not *all* successfully act upon them.[48]

Most important maxims, whether good or bad, come in between these two extremes. These are maxims on which, even in the most favourable circumstances, we couldn't all successfully act, but they are also maxims on which, even in the least favourable circumstances, some of us could successfully act. Almost all important maxims are condemned by the version of Kant's criterion that requires universal success. Hardly any important maxims are condemned by the version that appeals to universal failure. So Kant's criterion is either much too weak or much too strong.

In neither version, moreover, does Kant's criterion appeal to a morally relevant idea. For another illustration of this point, we can imagine a version of Kant's criterion that tries to occupy the middle ground. This might be

> (H) It is wrong to act on maxims whose being universally accepted, or acted upon, would greatly reduce the number of people who could successfully act upon them.

By appealing to (H) rather than (F) or (G), we would be able to condemn a larger number of wrong acts. Thus, if everyone accepted the maxim of a self-interested liar, that might reduce the number of people who could lie successfully, if they tried. But (H) could not explain why lying is wrong. There are many innocent maxims that (H) mistakenly condemns. It was not wrong for romantic poets to try to have the experience of being the only human being in some wilderness. Nor is it wrong to buy only secondhand books, or give surprise parties.[49]

[48] *PMJ,* p. 119.

[49] This is Herman's example: *PMJ,* p. 141.

7

We can now turn to Kant's famous

> *Formula of Universal Law:* It is wrong to act on some maxim unless we could also rationally will it to be true that this maxim is a universal law.

Kant sometimes suggests that, if some maxim fails this test, that gives us only an *unstrict* or *imperfect* duty not to act upon it. On this reading of Kant's Formula, we would be permitted sometimes to act on such maxims. But we should ignore this reading, as Kant often does. As we have seen, Kant's criterion of strict duties fails to condemn nearly all of the acts that any adequate criterion must condemn. We should ask whether Kant's Formula can fill this gap, by implying that some kinds of act are always wrong—or at least *prima facie* wrong, though permissible in special conditions.

Kant often claims that, in applying his formula, we should imagine that our maxim would become a universal law of nature.[50] On this version of Kant's Formula, which we can call

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

The word "everyone" here refers to all of those people to whom some maxim applies. Thus the maxim "Care for my children" applies only to parents.

In other passages, Kant appeals to what we can call

> *the Permissibility Formula:* It is wrong to act on some maxim unless we could also rationally will it to be true that everyone is morally permitted to act in this way.[51]

Kant assumes that, if it were permissible to act on the maxims that he discusses, people would be more likely to act in these ways. These effects would be produced, not by its *being* permissible to act in these ways, but by people's *believing* that such acts are permissible. So Kant is really appealing here to what we can call

[50] As on *G,* p. 421, and *Second Critique,* pp. 69–70.

[51] This reading of Kant's Formula has been, until recently, surprisingly neglected. Scanlon proposed it many years ago in some unpublished lectures.

> *the Moral Belief Formula:* It is wrong to act on some maxim unless we could also rationally will it to be true that everyone believes such acts to be permissible.

Kant remarks that he is not proposing a "new principle," but only a more precise statement of the principle that "common human reason… has always before its eyes."[52] This remark understates Kant's originality. But these two versions of Kant's Formula can be claimed to develop the ideas that are implied by two familiar questions: "What if everyone did that?" and "What if everyone thought like you?"

The wrongness of an act, Kant claims, depends on the agent's maxim. Kant sometimes uses "maxim" to refer the policy on which someone acts. On a second use, a maxim consists both of someone's policy and of this person's underlying aim. Suppose that two merchants both act on the policy "Never cheat my customers." One merchant acts in this way because he believes it to be his duty, while the other's motive is to preserve his reputation and his profits. These merchants, we might say, both have the same *policy maxim,* but they have different *deep maxims.*

Kant's Formula should not, I believe, appeal to either kind of maxim. Consider some egoist who has only one policy and underlying aim, "Do whatever would be best for me." This man could not rationally will that his maxim be universal. Egoists have strong reasons to want other people to accept and follow, not their egoistic maxim, but various moral principles. Egoists suffer from the egoism of others.

Since this egoist cannot will that his maxim be universal, Kant's Formula implies that, whatever this man does, he acts wrongly. He acts wrongly, not only when he steals, breaks promises, and harms other people, but also when, for self-interested reasons, he acts honestly, keeps his promises, and helps other people. These are unacceptable conclusions. When this egoist saves a drowning child, because he hopes to get some reward, he is not acting wrongly.

Turn now to someone who has mistaken moral beliefs. Our example can be Kant himself, during the period when he accepted the maxim "Never lie." This maxim is condemned by Kant's Formula, since Kant could not rationally will that no one ever tells lies, not even to prevent would-be murderers from finding their victims. So Kant's Formula implies, not only that Kant would have acted wrongly if he had told such a

[52] *G,* p. 403.

murderer the truth, but also that he did act wrongly whenever he acted on his maxim "Never lie." Since this maxim cannot be universalized, Kant acted wrongly whenever he told anyone the truth. That is another unacceptable conclusion.

The problem here is this. On Kant's Formula, if some maxim cannot be rationally willed to be universal, it is *always* wrong to act upon it. There are some maxims to which this claim applies. One example might be a sadist's maxim "Torture others for my own amusement." Such maxims, we might say, are wholly bad. But, of the maxims that cannot be willed to be universal, some are not wholly bad. While it is sometimes wrong to act upon these maxims, that is not always true. Since these maxims are neither wholly bad, nor wholly good, we can call them *mixed.*

Of the maxims that are actually accepted, many are of this kind. That is true of both the egoist's maxim "Do whatever would be best for me" and Kant's maxim "Never lie." Though we should not always act upon these maxims, doing so is often permissible or right. Kant overlooks the fact some maxims are in this sense mixed. His formula assumes that, if we could not rationally will it to be true either that everyone always acts on some maxim, or that everyone believes that all such acts are right, that shows that no one should ever act on this maxim, or that no such acts are right. But *not always* does not imply *never.* When we apply Kant's Formula to these mixed maxims, it has implications that are clearly false. We can call this the *mixed maxims objection.*

According to some writers, Kant distinguishes between some act's being wrong and its being contrary to duty. On this account, when my egoist keeps his promises and saves people's lives, Kant's formula implies that these acts are wrong, without implying that they are contrary to duty. Kant, I believe, does not draw this distinction. And Kant often claims that, by applying his formula, we can answer the question whether some act would be contrary to duty.

For Kant's Formula to answer such questions, it must be revised. An act's wrongness depends, not on the agent's policy or underlying aim, but on what this person is intentionally doing. In the morally relevant description of someone's act, we should include what this person intends her act to achieve, and the other morally relevant effects that she foresees. It is often irrelevant, however, on what policy this person acts. When Kant told someone the truth, it is irrelevant that he was acting on the policy "Never lie," so that he would have told the truth even to

the would-be murderer. And when my egoist saves someone's life, it is irrelevant that his underlying aim is only to get some reward. As Kant would claim, this man's act is not wrong, or contrary to duty, though it has no moral worth.

Since an act's rightness depends on what the agent intentionally does, we could drop all references to maxims. Of the two versions of Kant's Formula that I described above, the Law of Nature Formula could become

> *RLN:* It is wrong to act in some way unless we could rationally will it to be true that everyone does whatever, in acting in this way, we would be intentionally doing.

The Moral Belief Formula could become

> *RMB:* It is wrong to act in such a way unless we could rationally will that everyone believes such acts to be permissible.

To save words, I shall continue to discuss the unrevised versions of Kant's Formula. Kant sometimes uses "maxim" in a thinner sense, as when he discusses the maxim "Shorten my life to avoid suffering." This maxim is not a policy or principle. And, when someone acts on this maxim, what he is intentionally doing is shortening his life to avoid suffering. In most of what follows, we could understand "maxim" in this thinner sense.

To apply Kant's Formula, we must make some assumptions about what we could rationally will. On one view, it is always irrational to act wrongly. This view, even if true, is irrelevant here. For Kant's Formula to succeed, it must provide a criterion for the wrongness of acts that does not itself assume that these acts are wrong. It would be pointless to claim both that

> our act is wrong unless we could rationally will that everyone acts in this way

and that

> we could not rationally will that everyone acts in this way because such acts are wrong.

According to another view, we are rationally required to give significant weight to other people's well-being. It is irrational to benefit ourselves

in ways that impose great burdens on others. This view, even if true, is also irrelevant here. Kant's Formula does not assume that such acts are irrational. The main idea behind this formula is that, though it may be rational for us to act in ways that are wrong, we could not rationally will either that everyone acts in these ways, or that everyone believes such acts to be permissible. So, when we apply Kant's Formula, we should not appeal either to the irrationality of acting wrongly, or to the view that it would be irrational to give little weight to other people's well-being. We should appeal to claims about *nonmoral* and *nonaltruistic* rationality.

On one such view, we could not rationally will that everyone acts in some way, if such a world would be bad for us. That may seem not to be Kant's view, since Kant calls the principle of prudence, or "self-love," a merely *hypothetical imperative,* which applies to us only insofar as we care about our future well-being. But if we care about our future, as Kant assumes that we all do, it would be instrumentally irrational for us to will that other people act in ways that would be bad for us. Some Kantians prefer to claim that we could not rationally will a world in which our true needs would be worse met, or our rational agency would be frustrated, or our aims and purposes would be harder to fulfil. Though I shall appeal to the claim that we could not rationally will what would be worse for ourselves, most of my arguments could be restated in these other ways. I shall also appeal to claims about what would be *likely* to be good or bad for us. Though Kant seldom appeals to such claims, they are consistent with Kant's view, and they make his formula more successful.

8

Kant's 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 who act in this way, the effects of each act would be the same or roughly similar,

(3) these effects would be randomly or roughly equally distributed between different people.

These claims apply to many of the acts that are most clearly wrong, such as acts of self-interested injuring, coercing, or deceiving. Most people could often act in these ways. Whatever the number who act in these ways, most of these acts would have similar effects, since they would benefit their agents but impose greater burdens on others. And, in many cases, these burdens would be likely to be randomly or roughly equally distributed. In such cases, it would be likely to be worse for most of us if everyone rather than no one acted in these ways. Even if each of us would gain from acting in these ways, each would be likely to lose more from the similar acts of others. Kant's Formula condemns these acts, since we could not rationally will that such acts be universal.

In some of these cases, though any such act would impose burdens on others, it is also true that

> (4) since these burdens would be spread over many people, each act's effects on each person would be either trivial or imperceptible.

Some examples could involve pollution, soil-erosion, overfishing, overgrazing, and overpopulation. In such cases, if each of us considers only the effects of our own acts, we may believe that we are not acting wrongly. When applied to such cases, Kant's Formula is much more successful than most other relevant principles, such as the act utilitarian principle, ordinary principles about harming others, or the Golden Rule. Though each of these acts would impose only trivial burdens on others, we could not rationally will that everyone acts in these ways, since these acts would together impose on everyone, including us, great burdens. Though these are not the cases that Kant had in mind, they count strongly in favour of his formula.

When conditions (1) to (3) are not all met, however, Kant's Formula works less well. This formula here faces several objections, of which I have time to discuss only one.

There are some wrong acts whose bad effects are not randomly or equally distributed between different people. These acts impose burdens only on the people who are in certain groups.

The Golden Rule makes us impartial by telling us to treat others as we would want or will that others treat us. Kant's Formula makes us impartial in a less direct way. When we apply this formula, rather than asking "What if they did that to me?", we ask "What if everyone did that?" This kind of impartiality has great importance. If it is permissible for us to act in some way, it must be permissible for everyone else to act in the

same way as us, in the same circumstances. And when we act wrongly, as Kant points out, we often make unfair exceptions for ourselves, allowing ourselves to do things that we would not want or will other people to do.[53] On Kant's Formula, it is wrong to do what we could not rationally will everyone to do.

This kind of impartiality is not, however, enough. Like the Golden Rule, Kant's Formula applies best to those wrong acts with which we benefit ourselves in ways that impose greater burdens on others. We could not rationally will that other people do these things to us, since we would then have to bear these greater burdens. But, on Kant's Formula, we don't ask whether we could will that other people do these things *to us.* We ask whether we could will 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. Kant's Formula may then fail to condemn these wrong acts. This we can call *the impartiality objection.* If Kant's Formula cannot condemn these acts, it does not ensure the kind of impartiality that, as Kant assumed, moral reasoning requires.

Consider first some white racist, in the age of segregation. This man might have claimed to be following the two versions of Kant's Formula of Universal Law. He might have said:

> 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 and commies changed their mind, that would be fine with me.

In making these claims, would this man have misunderstood Kant's Formula? 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.[54] My question is only what is implied by Kant's Formula.

Kant did not consider cases of this kind. When he imagines some wrongdoer asking, "Could I will that my maxim be a universal law?" Kant assumes that this person's maxim *isn't* such a law. But in some cases, like that of this racist, a wrongdoer's maxim may already be universal,

[53] *G,* p. 424.

[54] See *KET,* pp. 3 and 7.

since it may already be acted on by all those people to whom it applies. Kant's Formula permits these people's acts if they could rationally will that they and others continue to act as they are now doing. If it is bad for these people that others are acting in the same way as them—as would be true, for example, in some anarchic war of all against all—these people could not rationally will the continuation of the status quo. But, if the status quo is good for these people, we may face the following problem. The status quo may be good for these people in part because their bad maxim is universal. Those to whom some maxim applies may be some powerful and privileged group, who are oppressing other people.

Kant's Formula condemns these people's acts only if they could not rationally will that they keep their privileged position. And, for the reasons given above, we cannot defend such claims in ways that assume that these people's acts are wrong. Nor can we appeal to the claim that these people are rationally required to give significant weight to other people's well-being. When we apply Kant's Formula, we must claim that it would be nonmorally irrational for such people to will that they keep their privileged position. Such claims may be hard to defend.

Nor would it help to turn to the moral belief version of Kant's Formula. If these people could rationally will that everyone acts in the same way as them, they could rationally will that everyone believes such acts to be permissible. 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 them various rights and privileges, and giving less weight to their well-being. On Kant's Formula, it is wrong for men to act in this way unless they could rationally will it to be true that everyone acts in this way, and that everyone believes such acts to be justified. That is not a useful claim. For most of history, most people, including most women, have treated women as inferior, and believed such treatment to be justified. Many men could rationally will that they keep their privileged position, and that everyone believes that position to be justified. Similarly, in condemning slavery, it would not help to claim that slave-owners acted wrongly unless they could have rationally willed, on nonmoral grounds, that they keep their slaves, and willed that everyone, including slaves, believes slavery to be justified. In considering such cases, we would do better to appeal to the Golden Rule, which Kant contemptuously dismissed. Men and slave-owners would not will that

they be treated as inferior, or as mere property, if they supposed that they themselves were going to be women or slaves.

For another example, consider global inequality. On any plausible moral view, those who control most of the world's resources ought to transfer some of their wealth to the billion poorest people in the world. Many rich people now transfer nothing to the poor. Kant's Formula does not condemn these people's acts if they could rationally will it to be true that all rich people act like them, and that everyone, including the poor, believes such acts to be justified. As before, Kant's Formula here achieves nothing.

It might be suggested that, if we redescribe these kinds of acts, Kant's Formula would do better. For example, when some man treats women as inferior, he is treating members of the opposite sex as inferior. In willing that everyone acts in this way, this man would be willing, not only that all men treat women as inferior, but also that all women treat men as inferior. Since men have more power, however, that redescription may not make enough difference. And there are many cases in which such redescriptions would be no help. We might say that, when the rich transfer nothing to the poor, they are giving nothing to those whose financial position is the opposite of theirs. That would allow us to claim that, in willing that everyone acts in this way, the rich would in part be willing that the poor give nothing to them. But the rich could happily will such a world, since the poor have nothing to give. Similarly, when the strong exploit the weak, it would not help to say that the strong are exploiting others, and that, if everyone acted in this way, others would exploit them. If the weak tried to exploit the strong, they would not succeed.

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, mean-ness, 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.[55]

[55] *CKE,* p. 101.

What Kant's view handles best is not, I have argued, all kinds of selfishness or advantage-taking. 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.

9

Some may think that, in presenting this objection, I have misinterpreted Kant's Formula. Thomas Nagel suggests that, when we ask whether we could rationally will that everyone acts in the same way as us, Kant intends us to imagine that we ourselves are going to be in everyone else's position.[56] This suggestion makes Kant's Formula more like the Golden Rule.

None of Kant's claims about his formula support Nagel's reading.[57] And there are contrary passages, such as Kant's discussion of the self-reliant man who has the maxim of not helping others who are in need. When he explains why this man could not rationally will that his maxim be a universal law, Kant 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*.[58]

If Kant intended this man to imagine that he would be in the positions of all of the people who would need help, it would be hard to explain why Kant doesn't say that here.

Nagel defends his reading with the claim that, if Kant did not intend us to imagine being in everyone else's position, Kant's Formula would be open to serious objections. But even the greatest philosophers

[56] Thomas Nagel, *Equality and Partiality* (Oxford University Press, 1991), pp. 42–43.

[57] Kant does write: "every rational being…must always take his maxims from the point of view of himself, and likewise every other rational being" (*G,* p. 438). But this remark comes in Kant's discussion of the Formula of the Kingdom of Ends, to which I shall return.

[58] *G,* p. 423 (my emphases).

can overlook possible objections. We should not assume that, when great philosophers seem to make some mistake, they cannot have meant what they wrote.

John Rawls proposes another reading of Kant's Formula. When we apply this formula, Rawls suggests, Kant intends us to imagine that we don't know anything 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 his reading with the claim that it seems needed to defend Kant's Formula from objections. 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….[59]

But as before, even if Kant ought to have made such an assumption, that doesn't show that he did. In his discussions of his formula, Kant never suggests that we should imagine being behind a veil of ignorance.

Scanlon proposes a third reading. Kant writes:

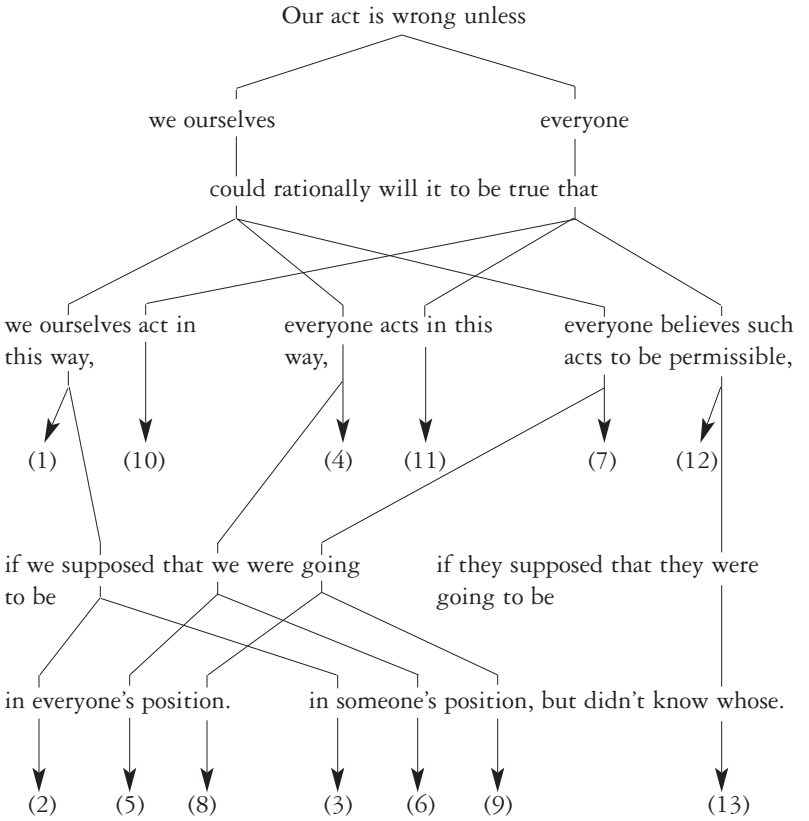
> I ought never to act except in such a way that I could also will that my maxim be a universal law.

Scanlon suggests that, when we apply this test, Kant intends us to ask whether *everyone* could rationally will that our maxim be a universal law. To defend this reading, Scanlon writes that, if Kant were merely telling us to ask what we ourselves could rationally will, he would be wrong to claim that his different formulas have the same implications.[60] Like the other two proposals, Scanlon's proposal cannot, I believe, be what Kant meant. Kant gives nearly twenty different statements of his Formula of Universal Law, none of which refer to what everyone could will.

These proposals are best regarded, not as interpretations, but as ways of revising Kant's Formula so that it avoids the impartiality objection. These and other such proposals can be shown as in Diagram 1. According to (1), to decide whether some act is wrong, it is enough to ask whether we could rationally will it to be true that we ourselves act in this

[59] John Rawls, *Lectures on the History of Moral Philosophy,* edited by Barbara Herman (Harvard University Press, 2000), p. 175.

[60] Scanlon, *What We Owe to Each Other,* pp. 170–71, and in unpublished summaries of lectures.



way. Even Kant does not make that assumption. According to (4), the Law of Nature version of Kant's Formula of Universal Law, we should ask whether we could rationally will that everyone acts in this way. According to (7), the Moral Belief version, we should ask whether we could rationally will that everyone believes such acts to be permissible.

These formulas, I have argued, fail. Many wrongdoers could rationally will both that everyone acts like them, and that everyone believes such acts to be permissible. These people may know that, even if everyone did to others what they are doing, no one would do these things to them. Many men, for example, could rationally will that everyone treats women as inferior.

(2), the Golden Rule, avoids this objection. These men could not rationally will such treatment if they were going to be women. According to (5), Nagel's proposal, we revise Kant's Law of Nature Formula so that it becomes like the Golden Rule. We ask what we could rationally will everyone to do, if we supposed that we ourselves were going to be in

everyone's position. (8) is a similar revision of Kant's Moral Belief Formula.

When revised in this way, these formulas avoid the impartiality objection. It is hard, however, to imagine ourselves in the positions of every other person. When Rawls discusses Richard Hare's version of (5), he suggests another objection. In imagining ourselves in all these positions, we shall think of ourselves as living all of these people's lives. That may lead us to ignore the separateness of these lives, and the fact that one person's burdens cannot be compensated by benefits to other people. We may thus be led to ignore the grounds for accepting principles of distributive justice.[61]

Rawls suggests that, to avoid this objection, we should imagine that we are behind a veil of ignorance. We should suppose that, rather than being in everyone's position, we shall be in one person's position, but we don't know whose. When revised in this way, Kant's Law of Nature and Moral Belief Formulas become (6) and (9).

Scanlon suggests that, rather than asking what we ourselves could rationally will, we should ask what everyone could rationally will. According to Kant's Law of Nature Formula, or

> (4) Our act is wrong unless we ourselves could rationally will it to be true that everyone acts in this way.

On Scanlon's proposal, this would become

> (11) An act is wrong unless everyone could rationally will that everyone acts in this way.

It is worth making a further revision. If we appeal to what everyone could will, that is enough to achieve impartiality. We need not ask whether everyone could will that *everyone* acts in some way. And there are some acts that are right, though we could not rationally will that everyone acts in these ways. Kant did not act wrongly, for example, in having no children. So we do better to turn to (10), or

> *the Formula of Universally Willed Acts:* An act is wrong unless it could be rationally willed by everyone.

This formula is a wider version of Kant's Consent Principle. On that principle, we ought to treat everyone only in ways to which they could

[61] John Rawls, *Theory of Justice* (henceforth *TJ*), section 30.

rationally consent. On this wider formula, an act is wrong unless everyone, if they had the choice, could rationally choose that this act be done.

According to Kant's Moral Belief Formula, or

> (7) Our act is wrong unless we ourselves could rationally will it to be true that everyone believes such acts to be permissible.

On Scanlon's proposal, this becomes

> (12) An act is wrong unless *everyone* could rationally will it to be true that everyone believes such acts to be permissible.

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?"[62]

This formula is strongly suggested by several of Kant's claims about his other two main principles, the Formulas of Autonomy and of the Realm of Ends. For example, Kant refers to

> the concept of every rational being as one who must regard himself as giving universal law....

Kant never explicitly appeals to what everyone could rationally will. The phrase just quoted, for example, ends

> through all the maxims of his will.[63]

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. But, as I have argued, this formula needs to be revised; and (12) is the revision that seems closest to Kant's own view. Though Kant never appeals to (12), that, I suggest, is only because Kant assumes that (7) and (12) coincide, since what each of us could rationally will must be the same as what everyone else could will.

(12) might be called the *Formula of Universally Willed Moral Beliefs.* But we can restate this formula, and give it a shorter name, as

[62] Scanlon, *What We Owe to Each Other,* p. 171.

[63] *G,* p. 433.

> *Kant's Contractualist Formula:* We ought to act on the principles whose universal acceptance everyone could rationally will.

(13) differs from (12) by including a veil of ignorance. On the best known version of (13), or

> *Rawls's Formula:* We ought to act on the principles that it would be rational for everyone to choose, as the principles that we would all accept, if no one knew anything about themselves or their circumstances.

In tomorrow's lecture, my main subject will be Kant's Contractualist Formula. If we combine this formula with the right view about reasons, we shall reach what may be the best version of contractualism. We shall also be led to some surprising conclusions.

III. CONTRACTUALISM

10

Most contractualists ask us to imagine that we are all trying to reach agreement on which moral principles we shall all accept. According to what we can call

> *the Rational Agreement Formula:* We ought to act on the principles to whose acceptance it would be rational for everyone to agree.

I shall say that people *choose* the principles to whose acceptance they agree. People choose rationally, most contractualists assume, if their choices would be likely to be best for themselves. We can start by making this assumption.

Though there are some principles whose acceptance would be likely to be best for everyone, there are others whose acceptance would be best only for certain people. What would be best for the rich, for example, would not be best for the poor. It may seem that, in such cases, there would be no principle whose choice would be rational for everyone in self-interested terms. But everyone would know that everyone would accept only the principles that everyone chose. So what each person

ought rationally to choose would depend on what others 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 everyone else.

What we ought rationally to choose would also depend on the effects of our failing to reach agreement. Most contractualists tell us to suppose that, if we failed to agree, no one would accept any moral principles, so no one would believe that any acts were wrong. This *no-agreement world* would be likely to be bad for everyone. That would give everyone strong reasons to try to reach agreement. And it might be rational for everyone to choose, not the principles whose acceptance would be best for themselves, but the principles that other people would be most likely to choose. This might be everyone's best hope of avoiding the horrors of the no-agreement world.

Such reasoning is, for some writers, the essence of contractualism. On this view, we should regard morality as if it were a mutually advantageous bargain. When people's interests conflict, it would be rational for everyone to agree on certain principles to resolve these conflicts. And 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.

To make this imagined agreement easier to achieve, we can suppose that there would be discussions, and a series of votes. But there would have to be some final vote. It must be true 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 would be rational for us to try to reach agreement on terms that favoured ourselves. Only in the decisive final vote would it be rational for us to make our full concessions to others.

There is now a complication. The no-agreement world would be less bad for certain people, such as those who control more resources, or have greater abilities. In a world without morality, such people would be better able to fend for themselves. These people would have less need to reach agreement, and that would give them greater bargaining power. These people could declare that they would accept only principles that gave special advantages to them. Such warnings or threats might be credible, since these people would be more prepared to run the risk of no agreement.

In some cases, moreover, it would 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 transfer 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 would be fine with the rich. For these and similar reasons, those who had greater bargaining power could win agreement on principles that gave special advantages to them, since it would be rational for others to give in to their threats.

Some writers accept these implications of the Rational Agreement Formula. These Hobbesian contractualists defend a minimal version of morality. On David Gauthier's view, for example, since morality presupposes mutual benefit, it would not be wrong for us to act in ways that injure or kill other people, if it would have been no worse for us if these people had never existed.

Kantian contractualists, like Rawls, reject these implications. As Rawls writes, "to each according to his threat advantage is not a conception of justice."[64] But Rawls's version of contractualism is not, I believe, Kantian enough.

11

In considering Rawls's view, we can start with his assumptions about rationality. Rawls accepts the Deliberative Theory, according to which we ought rationally to do whatever would best achieve what we most want after informed deliberation. Of those who accept this theory, many believe that it coincides with the Self-interest Theory, according to which we ought rationally 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.

Rawls does not make that assumption. He considers cases in which justice requires us to act in ways that would be very bad for ourselves. 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.[65]

[64] *TJ,* p. 134, revised edition (henceforth *RE*), p. 116.

[65] *TJ,* p. 569, *RE,* p. 498.

Since the Deliberative Theory is desire-based, however, Rawls cannot claim that it would be rational for everyone to act justly. When he discusses people who would benefit from injustice, 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.[66]

On desire-based theories, we cannot have reasons to want anything for its own sake. If people don't care about something, and would not care even after informed 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.[67]

Similarly, when he rejects the view that

> something is right…when an ideally rational and impartial spectator would approve of it,

Rawls comments,

> Since this definition makes no specific psychological assumptions about the impartial spectator, it yields no principles to account for his approvals….[68]

This comment assumes that we have no reasons to care about the well-being of others. 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 reasons to approve. For example, he would approve of acts that relieved suffering, or saved people's lives.

We can now turn to Rawls's suggested account of morality, which he calls *rightness as fairness.* 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 want. Since Rawls cannot predict what everyone would want, he adds a motivational assumption. He tells us to

[66] *TJ,* p. 575, *RE,* pp. 503–4.

[67] *Political Liberalism* (Columbia University Press, 1996), p. 49.

[68] *TJ,* p. 185, *RE,* p. 161.

suppose that, when we were choosing moral principles, everyone's main aim would be to promote their own well-being. On this simplifying assumption, the Deliberative and Self-interest Theories coincide. If we cared most about our own well-being, it would be deliberatively rational for us to make the choices that we could expect to be best for ourselves. So, though Rawls rejects the Self-interest Theory, his motivational assumption allows him to appeal to claims about self-interested rationality.

Rawls revises the Rational Agreement Formula by adding a veil of ignorance. According to

> *Rawls's Formula:* We ought to act on the principles that it would be rational for everyone to choose, if we had to choose these principles without knowing anything about ourselves or our circumstances.

Rawls gives several reasons for his veil of ignorance. If we had full information, Rawls claims, we would often be unable to reach agreement. And, if we knew nothing about ourselves, that would make us impartial. No one would know the facts that would give some people greater bargaining power. Nor could anyone choose principles that were biassed in their own favour. Though we would be choosing principles for self-interested reasons, our ignorance of who we are would mean that everyone's well-being would, in effect, be taken into account.[69]

Rawls believes that his contractualism provides a systematic alternative to all forms of utilitarianism. This belief is surprising. If we appeal to a combination of self-interested rationality and impartiality, it is hard to avoid utilitarian conclusions. Utilitarianism is, roughly, self-interested rationality plus impartiality.

Rawls is aware of this problem. He compares two versions of his formula. On what we can call the *equal-chance formula,* if we were behind Rawls's veil of ignorance, we should assume that we had an equal chance of being in anyone's position. Rawls admits that, on this assumption, it would be rational for us to choose a principle whose acceptance would make the average level of well-being as high as possible.[70] By choosing this *utilitarian average principle,* we would maximize our own expected

[69] *TJ,* section 24.

[70] *TJ,* pp. 165–66, *RE,* pp. 143–44. Rawls might argue that it would be rational to choose a principle that was more cautious than this 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 average principle.

level of well-being. But Rawls rejects this equal-chance formula. If we were behind the veil of ignorance, Rawls claims, we should not assume that we had an equal chance of being in anyone's position. On his preferred *no-knowledge formula,* we would have no knowledge of the probabilities. That would make it rational for us, Rawls argues, to choose certain nonutilitarian principles.

For this argument to succeed, 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.[71]

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 is wrong with that assumption? Rawls himself points out that, since there are different contractualist formulas, he must explain why we should appeal to his 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.[72]

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 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, these formulas are equally plausible.[73]

Remember next that, as Rawls claims, the equal-chance formula "leads naturally" to the utilitarian average principle.[74] Since Rawls can-

[71] *TJ,* p. 168, *RE,* p. 145.

[72] *TJ,* pp. 122 and 121, *RE,* p. 105.

[73] This objection to Rawls's argument I take from Thomas Nagel's "Rawls on Justice," *Philosophical Review* (April 1973), reprinted in *Reading Rawls,* edited by Norman Daniels (Blackwell, 1975), p. 11.

[74] *TJ,* p. 166, *RE,* p. 143.

not justify his rejection of this formula, Rawls's contractualism does not, as he believes, provide an argument against utilitarianism.

As Rawls points out, there is another ground on which we might justifiably reject some formula. We may be right to reject some formula, however plausible it seems, if this formula's implications conflict with some of our strongest moral beliefs. Since Rawls assumes that utilitarianism conflicts with such beliefs, he might claim that we can justifiably reject the equal-chance formula on the ground that it leads to an unacceptable conclusion.

If Rawls made this claim, however, his contractualism would still provide no argument against utilitarianism. Rawls would be appealing to our nonutilitarian 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 nonutilitarian beliefs. If we defend some argument 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 acceptable nonutilitarian 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. But Rawls's Formula cannot, I believe, achieve this aim.

When he appeals to 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 equally worst-off people as well off as possible. Since this argument tells us to *max*imize the *min*imum level, we can call it the *Maximin Argument.*[75]

[75] 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,* p. 98, *RE,* p. 84). But, if the Maximin Argument were sound, it would support 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 definition, we should support policies that 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.

This argument has been widely criticised. Even if it were sound, however, it would not support an acceptable moral view. Suppose that we must choose how to use some scarce medical resources. In one of the two possible outcomes,

> Green would live to twenty-five, and a thousand other people would live to eighty.

In the other outcome,

> Green would live to twenty-six, and these other people would live to thirty.

On the Maximin Argument, we ought to choose this second outcome, giving Green her extra year of life. That is the wrong conclusion. We can plausibly give some priority to benefiting those who are worse off. But this priority should not be absolute. It would be wrong to give Green one extra year of life, rather than giving fifty extra years to each of a thousand other people—people who, without these years, would die almost as young as Green.

Rawls accepts what I have just claimed. Though he applies his Maximin Argument to the basic structure of society, Rawls agrees that, when we consider other questions about distributive justice, this argument has implications that are unacceptably extreme. He rejects utilitarian theories because they fail to provide an acceptable general principle of distributive justice. But, as Rawls admits, his version of contractualism also fails to provide such a principle.

We can now turn to other moral questions, such as whether and when it would be right to break promises, or tell lies, or impose certain risks on others. On Rawls's Maximin Argument, when we choose between different principles about such acts, we ought rationally to choose the principles whose acceptance would make the worst-off people as well off as possible.

There are here three objections to this argument. First, it is hard to apply. In the case of promises, for example, whom should we regard as the worst-off people? Are these the people who would lose most from particular acts of breaking or keeping promises, or those who would benefit least from the practice of promising, or those who would be, on the whole, worst off all things considered?

Second, however we answer such questions, this cannot be the right way to choose between different principles. If one of two forms of the practice of promising would give much greater benefits to most

people, that is not, as the Maximin Argument implies, morally irrelevant.

Third, this argument forces us to ignore most nonutilitarian 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 relevant considerations. Some examples are certain facts about responsibility, desert, deception, coercion, fairness, gratitude, and autonomy. Other examples are distinctions between positive and negative duties, such as the distinction between harming and failing to benefit. 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.[76]

Rawls merely denies these people most of the knowledge that self-interested calculations need. There is no way in which nonutilitarian 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.[77]

He also writes:

> for the contract view, which is the traditional alternative to utilitarianism, such a conclusion would be a disaster.[78]

Rawls, I believe, could deny that his theory justifies utilitarianism. But his claim would have to be that, though his theory leads to utilitarian conclusions, it is not plausible enough to support these conclusions.[79]

[76] *TJ,* p. 584, *RE,* p. 512.

[77] *TJ,* p. 29, *RE,* pp. 25–26.

[78] "Distributive Justice: Some Addenda," 1968, republished in John Rawls, *Collected Papers,* edited by Samuel Freeman (Harvard University Press, 2001), p. 174.

[79] 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], p. 106). But nothing more is said.

12

To reach a more successful version of contractualism, we should turn to a different formula, and a different view about rationality and reasons. According to

> *the Rational Agreement Formula:* We ought to act on the principles to whose acceptance it would be rational for everyone to agree.

According to what I have called

> *Kant's Contractualist Formula:* We ought to act on the principles whose universal acceptance everyone could rationally will.

These formulas both require unanimity, since they both appeal to the principles that everyone could rationally choose. But, unlike the Rational Agreement Formula, Kant's Formula does not use the idea of an agreement. According to the Agreement Formula, more fully stated:

> We ought to act on the principles that it would be rational for everyone to choose, if each person supposed that everyone would accept all and only the principles that *everyone* chose.

According to Kant's Formula:

> We ought to act on 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 the Agreement Formula, we conduct a single thought-experiment, in which we imagine that we are all trying to reach agree-

---

Rawls adds some other stipulations that 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,* pp. 181–82 and note 31, *RE,* pp. 158–59 and 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.

I should emphasize that, in criticizing Rawls's appeal to his contractualist formula, I am not criticizing his theory as a whole. Several of Rawls's claims, such as his claims about the moral arbitrariness of the natural lottery, I believe to be both true and of great importance.

ment on which principles everyone would accept. In applying Kant's Formula, we conduct many thought-experiments, one for each person. We imagine that each of us applies Kant's Moral Belief Formula, by asking which principles she could rationally choose, if she had the power to choose which principles everyone would accept. Kant's Contractualist Formula appeals to the principles that, in these separate thought-experiments, everyone could rationally choose.

Kant's Formula, I believe, better achieves Rawls's aims. One of these aims is to eliminate inequalities in bargaining power. On Kant's Formula, what it would be rational for each of us to choose does not depend on what other people would be likely to choose. 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. Justice, Rawls writes,

> does not allow that the sacrifices imposed on the few are outweighed by the larger sum of advantages enjoyed by the many.[80]

According to several Kantian contractualists, utilitarianism goes astray because of the way in which it adds together different people's benefits and burdens. 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. According to these contractualists, to protect people from having such great burdens imposed on them, we should deny that the numbers count, and should appeal instead to the idea of a unanimous agreement. By requiring unanimity, we give everyone a veto against being made to bear such burdens. We can call this contractualism's *protective aim.*

The Rational Agreement Formula, as we have seen, fails to achieve this aim. Precisely by requiring a 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 also fails 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 people on whom utilitarian principles would impose great burdens. And, since Rawls

[80] *TJ,* p. 4, *RE,* p. 3.

appeals to the principles whose choice would be rational in self-interested terms, he cannot plausibly deny that we could rationally choose rule utilitarian principles, running the risks of bearing some burdens for the sake of possible benefits.

Since Kant's Formula neither requires an agreement nor has a veil of ignorance, it better achieves contractualism's protective aim. On this formula, we ought not to impose burdens on anyone unless our act is permitted by some principle that this person could rationally choose, even when she knows that she is one of the people who would bear such burdens. And such people, we can argue, could not rationally choose rule utilitarian principles.

Kant's 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, unanimity is guaranteed. In the thought-experiments to which Kant's Formula appeals, everyone knows how their interests conflict with the interests of others. Since unanimity is not guaranteed, it would be significant if unanimity could be achieved.

Whether unanimity could be achieved depends on what we ought to believe about rationality and reasons. If we ought to accept a desire-based theory, or the Self-interest Theory, Kant's Formula could 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 theories. And though the Self-interest Theory is of the right kind, in being value-based, this theory is too narrow. On the wide value-based view 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 the supremely rational ultimate aim. We can rationally care as much, or more, about some other things, such as justice, and the well-being of others.

If we combine Kant's Formula with this view about reasons, does this formula succeed? Ought we to act on the principles whose universal acceptance everyone could rationally will?

I cannot answer that question here. But there are two other, prior questions. *Are* there such principles? If there are, what do they imply?

For Kant's Formula to succeed, what we can call its *uniqueness condition* must be sufficiently often met. It must be true that, at least in most cases, there is some relevant principle, and only one such principle, that everyone could rationally choose. If there was either no such principle, or everyone could rationally choose two or more seriously conflicting principles, Kant's Formula would have unacceptable implications, since it would either permit or condemn too many acts. It might not matter, though, if everyone could rationally choose any of several similar principles. Such principles might 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.

This formula's uniqueness condition is, I believe, often met.

Consider first cases in which

> 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 being worse off than others,

and

> however these goods are distributed, the total sum of benefits would be the same.

Most of us believe that, in such cases, everyone should get equal shares.

Kant's Formula appeals to the principles that everyone could rationally choose, if each person supposed that everyone would accept whatever principles she chose. We can 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 the Self-interest Theory, we must deny that everyone could rationally choose the principle of equal shares. 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 informed 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.

According to (B), no one could rationally choose any principle that gave them less than equal shares. We can note a difference here between consenting to acts and choosing principles. We could sometimes rationally consent to being given a less than equal share, so that some stranger would get a greater share. Such choices would be generous and fine. But that does not imply that we could rationally choose some *principle* that gave us less than equal shares. Such a principle would apply not only to us but to all the members of some group. One example would be a principle that gave smaller shares to women. If we are women, we could not rationally choose this principle. We would have no reason to want all other women to get smaller shares; nor would this choice be generous and fine.

Consider next *Act Utilitarianism,* or *AU,* according to which acts are right just in case they maximize the sum of benefits. In the cases we are now considering, the sum of benefits would be the same whatever their distribution, so AU implies that it does not matter how we distribute these benefits. We could permissibly give some people no benefits at all.

Even the greatest Act Utilitarian rejected this conclusion. In such cases, Sidgwick writes, we ought to supplement AU with an egalitarian principle.[81] If Sidgwick had applied Kant's Formula, he would have admitted that no one could rationally choose AU, since no one could rationally choose any principle that permits them to be given no benefits at all.

According to this argument's remaining premise, only the principle

[81] *The Methods of Ethics* (London: Macmillan, 1907), pp. 416–17.

of equal shares gives no one less than equal shares. That is clearly true. So, as this argument claims, this is the only principle that everyone could rationally choose. Kant's Formula implies that, in these cases, everyone should get equal shares.

In most cases, the sum of benefits depends in part on how goods are distributed between different people. I believe that, in some of these cases, Kant's Formula can be used to defend other nonutilitarian distributive principles. But, given their complexity, I shall not discuss these cases here.

Someone may again object: "When we ask these questions about what people could rationally choose, our answers depend on our moral beliefs. You believe that you could rationally choose this principle because you believe that it would be permissible to act in this way. If people believed that such acts are wrong, they would not believe that they could rationally choose this principle. Since our answers to these questions depend on our moral beliefs, Kant's Formula achieves nothing."

This objection, as I have said, seems to me mistaken. When we combine Kant's principles with our beliefs about rationality, we can reach conclusions that conflict with our moral beliefs. But these beliefs are, as this objection claims, closely related. By considering these relations, we can reach some wider conclusions.

13

Our beliefs about rationality and reasons may partly depend on our moral beliefs, since we have moral reasons for acting, and some of these reasons are provided by facts about what is right or wrong.

This dependence can also go the other way. Our moral beliefs may partly depend on our beliefs about rationality and reasons. That can be true in two main ways. First, we may be contractualists, who believe that the rightness of acts depends on which are the principles that everyone could rationally choose. Second, our moral beliefs may partly depend on our beliefs about what is nonmorally good or bad, and these beliefs may in turn depend on our beliefs about reasons.

This dependence is less obvious. Pain is bad, most of us believe, in the sense of being something that we have reason to avoid. But some great philosophers did not have this belief. David Hume, for example, does not use "good" or "bad" in reason-involving senses. That is why he

claims that it cannot be contrary to reason to prefer our own acknowledged lesser good. Hume uses "good" and "evil" to mean "pleasure" and "pain." When Hume calls pain bad, he means that pain is pain.

While Hume would have thought it trivial to claim that pain is bad, Kant believed this claim to be false. Kant writes:

> good or evil is, strictly speaking, applied to actions, not to the person's state of feeling.... Thus one may laugh at the Stoic who in the most intense pains of gout cried out, "Pain, however you torment me, I will still never admit that you are something evil (*kakon*, *malum*)," nevertheless, he was right.[82]

Kant misunderstands this Stoic, taking the Stoic to mean that his pain is not morally bad.[83] Feelings, Kant agrees, cannot be morally bad. But the Stoic is denying that his pain is bad in the sense of being something that he has reasons to avoid. Kant ignores this kind of badness.

Another such philosopher is David Ross. Though Ross believes that pain is bad, he assumes that, if some outcome would be bad, we have a *prima facie* duty to prevent this outcome, if we can. Because Ross believes that we have no duty to prevent our own pain, he concludes that our own pain is not bad. Each person's pain is bad, he claims, only for *other* people.[84] Ross reaches this strange conclusion because he ignores the reason-involving senses in which pain is bad.

In claiming that it is bad to be in pain, I mean that we each have personal reasons to want not to be in pain. Personal reasons are *agent-relative,* in the sense of being reasons for only one person. As well as being bad *for* the person who is in pain, pain is also *impersonally* bad. If more people suffer, that would be worse, even though it may not be worse for any of these people. Impersonal badness involves reasons that are *agent-neutral,* in the sense of being reasons for everyone. Each of us has reasons to prevent or relieve any conscious being's pain, if we can.

Rather than distinguishing between these two kinds of reason, we can distinguish between the weights that our reasons have from two different points of view. From an impartial point of view, everyone matters

[82] *Second Critique,* p. 60.

[83] As Terence Irwin notes ("Kant's Criticisms of Eudaemonism," in *Aristotle, Kant, and the Stoics,* edited by Stephen Engstrom and Jennifer Whiting [Cambridge University Press, 1996], p. 80).

[84] Sir David Ross, *The Foundations of Ethics* (Oxford University Press, reprinted 2000), pp. 272–84. (Though Ross makes these claims about pleasure, he intends them to apply to pain.)

equally, so we have equal reasons to care about everyone's well-being. If some stranger's pain is worse than ours, our reason to relieve that person's pain is impersonally stronger. From our own, personal point of view, though we have some reason to care about everyone's well-being, we have stronger reasons to care about our own well-being and the well-being of those we love.[85]

If the strengths of our reasons differ in this way, what we have most reason to want, or do, may be different from these two points of view. In such cases, we can ask what, all things considered, we have most reason to do. Since these points of view cannot be easily combined, we should not expect that there would always be, even in principle, a precise answer. That is why, as I have claimed, we often have sufficient reasons to do what, from either point of view, we have most reason to do.

Return now to the ways in which our moral beliefs may depend on our beliefs about reasons. We can call some outcome

> *best* in the *impartial reason-involving sense* if this outcome has features that give everyone the strongest impartial *outcome-given* reasons to prefer this outcome, and to bring it about if they can.

This definition does not assume that all impartial reasons are outcome-given. We may have other impartial reasons, such as reasons provided by people's rights. And such reasons might outweigh our outcome-given reasons. Suppose that some poor person gives us some wallet, bulging with bank notes, that she has just found. We may believe both that we would make the outcome best if we told this person to keep this wallet and that our reasons to produce this outcome are outweighed by our reasons to try to return the wallet to its rich owner. On this view, we may have the strongest impartial reasons both to try to find this owner and to hope that this attempt will fail, so that we could then permissibly produce the best outcome, by returning the wallet to its poor discoverer.

The view just sketched is *semi-consequentialist,* in the sense that it gives some moral weight to the goodness of outcomes. On views that are *consequentialist,* the rightness of acts depends entirely on certain facts about what is nonmorally good or bad. According to

> *Act Consequentialists:* Acts are right just in case they make things go best.

[85] See Thomas Nagel, *Equality and Partiality* (Oxford University Press, 1991), chapter 2.

According to

> *Rule Consequentialists:* Acts are right just in case they are permitted by one of the principles whose acceptance would make things go best.

If these people use "best" in the impartial reason-involving sense, we can call them *Reason Consequentialists.*

These consequentialists believe that

> (1) to answer moral questions, we must answer questions about rationality and reasons.

Contractualists also believe (1). While consequentialists appeal to claims about which outcomes would be best, contractualists appeal to claims about what everyone could rationally choose. But, when outcomes are best in the impartial reason-involving sense, they may be the outcomes that, from an impartial point of view, everyone could rationally choose. So, if consequentialists and contractualists make the same assumptions about rationality and reasons, their theories may coincide. Though consequentialists appeal to what is best, and contractualists appeal to what everyone could rationally choose, these people may reach the same conclusions. They may find that, in John Stuart Mill's phrase, they have been climbing the same mountain on opposite sides.

This possibility has been widely overlooked. But that is not surprising. Many people do not use "best" in the impartial reason-involving sense. Some people make no use of the concept of a normative reason. Others accept desire-based theories about reasons, or the Self-interest Theory. On these theories, there are no outcomes that everyone has some reason to want, since there are no outcomes that would be good for everyone, or that would do something to fulfil everyone's informed desires.

There are other ways in which Reason Consequentialism has been overlooked. Rawls writes that, on consequentialist theories,

> the good is defined independently from the right, and then the right is defined as that which maximizes the good.[86]

Such theories are not worth considering. In calling some act "right" in this *consequentialist sense,* we would mean that this act maximizes the

[86] *TJ,* p. 24, *RE,* pp. 21–22.

good, by producing the best outcome. According to Act Consequentialism, or

> AC: Acts are right just in case they produce the best outcome.

If consequentialists used "right" in this sense, AC would mean

> Acts produce the best outcomes just in case they produce the best outcomes.

No one could deny this trivial claim.

Rather than defining the right in terms of the good, Kant sometimes defines the good in terms of the right. In calling some outcome "best" in this *deontological sense,* we would mean that this is the outcome that it would be right for us to produce. If consequentialists used "best" in this sense, AC would mean

> Our acts are right just in case they produce the outcomes that it would be right for us to produce.

As before, no one could deny this trivial claim. For consequentialists to make significant claims, they must neither use "right" in the consequentialist sense nor use "best" in the deontological sense.

Consequentialists might use "right" and "best" in what Ross and G. E. Moore claim to be their indefinable senses. That would make AC significant, though somewhat obscure. Reason Consequentialists use "best" in the impartial reason-involving sense. These people might use "right" in any of several reason-involving senses, which we need not consider here. On this version of AC,

> Acts are right just in case they produce the outcomes that everyone has the strongest impartial outcome-given reasons to prefer, and to bring about, if they can.

That is another significant claim.

We should next distinguish between consequentialism and utilitarianism. Consequentialists believe that

> (2) the rightness of acts depends only on facts about what would make things go best.

Utilitarians believe that

> (3) the rightness of acts depends only on facts about what would benefit people most.

Some utilitarians are consequentialists, since they believe that

> (4) things go best when they go in the ways that would benefit people most.

There are also some utilitarians who make no appeals to the goodness of outcomes. According to one such view, we ought to do what would benefit people most because that is how we can best treat everyone with equal respect and concern.

Just as some utilitarians are not consequentialists, some consequentialists are not utilitarians. These people accept (2), believing that the rightness of acts depends only on facts about what would make things go best. But they reject both (3) and (4). They believe that

> (5) it would often be best if things went in some way that did not benefit people most.

On the view that is most relevant here,

> (6) how well things go depends in part on how benefits and burdens are distributed between different people.

On this view, one of two outcomes could be better, though it would involve a smaller sum of benefits, because these benefits would be more equally distributed, or because more of these benefits would go to people who are worse off. In their beliefs about the goodness of outcomes, these consequentialists accept distributive principles.

Rawls claims that, if a moral theory includes such principles, it is not consequentialist. In his words:

> the problem of distribution falls under the concept of right as one intuitively understands it, and so the theory lacks an independent definition of the good.[87]

As we have seen, this may not be true. Reason Consequentialists use "good" in the impartial reason-involving sense. When these people claim that

> (2) the rightness of acts depends only on facts about what would make things go best,

[87] *TJ,* p. 25, *RE,* p. 22.

they mean that

> (7) the rightness of acts depends only on facts about what everyone has most reason to want, from an impartial point of view.

These people may claim that

> (6) the goodness of outcomes depends in part on the distribution of benefits and burdens,

because they believe that

> (8) everyone has such impartial reasons to want benefits to be more equally distributed, or to want benefits to go to people who are worse off.

There is no useful sense in which this view is not consequentialist.

14

Kant was not a consequentialist. But we are considering, not Kant's moral beliefs, but the implications of his principles.

According to Kant's

> *Consent Principle:* We ought to treat people only in ways to which they could rationally consent.

Act Consequentialists can argue:

> (1) Everyone could rationally consent to being treated in any way that would make things go best.

Therefore

> The Consent Principle never implies that such acts are wrong.

This argument's premise, I believe, is true. I believe that in *Lifeboat,* for example, White could rationally consent to our leaving her to die, so that, in the time available, we could save the five. If that belief is true, that strongly supports (1). If we could rationally consent to being left to die, when and because that is how things could go best, we could rationally give such consent to having lesser burdens imposed on us.

Note that, to accept (1), we need not assume that everyone could

rationally consent to being treated in any way that would benefit people most. When such an act would impose a great burden on one person, for the sake of a greater sum of benefits to people who are better off, we may believe both that this person could not rationally consent to this act and that this act would make things go worse. As I have said, consequentialists can reject utilitarianism.

If this argument is sound, as I believe, Kant's Consent Principle could be accepted by Act Consequentialists.

Kant's Contractualist Formula, however, provides one premise of an argument against AC. In the thought-experiments to which this formula appeals, each person supposes that she has the power to choose which principles we would all accept. We can argue:

> We ought to act on the principles that everyone could rationally choose.
>
> We could not all rationally choose that everyone accepts the Act Consequentialist principle.

Therefore

> This is not the principle on which we ought to act.

This argument, I believe, is also sound. Sidgwick concluded that we could not rationally want it to be true that everyone accepts Act Utilitarianism. Of Sidgwick's reasons for reaching this conclusion, most would apply to Act Consequentialism. If everyone believed that it was right to do whatever would make things go best, that would be unlikely to make things go best. Things would be likely to go better if everyone had certain other moral beliefs. If we knew that to be true, some of us could not rationally choose that everyone accepts Act Consequentialism. Kant's Formula would then imply that we should reject this view.

Return now to Rule Consequentialism. On this view,

> we ought to act on the principles whose acceptance would make things go best.

This view is very different from Act Consequentialism. Partly for the reasons Sidgwick gave, Rule Consequentialism supports principles that are much closer to most people's moral beliefs.

Rule Consequentialists can appeal to Kant's Contractualist Formula. They can argue:

(A) We ought to act on the principles whose universal acceptance everyone could rationally choose.

(B) Everyone could rationally choose whatever they would have sufficient reason to choose.

(C) Everyone would have sufficient reason to choose that everyone accepts the principles whose acceptance would make things go best.

(D) These are the only principles that everyone would have sufficient reasons to choose.

Therefore

These are the principles on which we ought to act.

Premise (A) is Kant's Contractualist Formula. If the other premises are true, this formula implies Rule Consequentialism.

According to premise (B), everyone could rationally choose whatever they would have sufficient reason to choose. That is not always true. 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 can be rational to choose what we have no reason to choose, or vice versa. So, when we apply Kant's Formula, we should suppose that everyone knows the relevant facts.

We should suppose, in particular, that people have no false beliefs about reasons. If we have sufficient reasons to make some choice, but we falsely believe that we have stronger reasons not to make this choice, we could not rationally make this choice. It is irrational to choose what we believe that we have stronger reasons not to choose. For Kant's Formula to be plausible, we must suppose that everyone knows what they have reasons to choose. On that assumption, premise (B) is true. Everyone could rationally choose whatever, as they know, they have sufficient reason to choose.

Kant's Formula appeals to the principles that everyone could rationally choose, if each person supposed that she could choose which principles everyone would accept. According to premise (C), each person would have sufficient reason to choose the principles whose acceptance would make things go best.

Some people would reject (C) because they believe that there is no intelligible sense in which things might go better or worse. But, if these people have the concept of a normative reason, as some of them seem to do, they must understand the claim that some outcomes would be best in the impartial reason-involving sense. These people might claim that there are no such outcomes, since there are no outcomes that everyone has reasons to want. And this must be claimed by those who accept either desire-based theories about reasons or the Self-interest Theory. I am assuming, though, that we should reject these theories. We all have reasons to want some outcomes, such as those in which fewer people suffer, or die young. And there are some principles whose acceptance would, in this impartial sense, make things go best or equal-best.

We might also challenge (C) by appealing to our nonconsequentialist moral beliefs. We may believe that, if everyone accepted the principles whose acceptance would make things go best, that would sometimes lead people to act wrongly. This may seem to give us an overriding reason not to choose these principles. But this objection misunderstands contractualism. When we apply some contractualist formula, we should not appeal to our beliefs about the wrongness of acts. According to contractualists, which acts are wrong depends upon which principles we could rationally choose. That claim would achieve nothing if which principles we could rationally choose depended on our beliefs about which acts are wrong. In applying a contractualist formula, as Rawls says, we should "bracket" these beliefs. We should appeal to our moral beliefs only at a different stage, when we are deciding whether to accept this formula.

There is another, similar point. We may believe that, in many actual cases, it would be wrong to choose what would make things go best. In making such choices, we might be violating someone's rights, or failing to fulfil some obligation. And we may believe that, since such choices would be wrong, they would not be the choices that we had most reason to make, from an impartial point of view. But such claims do not apply to the imagined thought-experiments to which Kant's Formula appeals. When we ask which principles we and others could rationally choose, in these imagined cases, we should assume that no one would have any moral obligations either to choose or not to choose certain principles. As before, in applying Kant's Formula, we should bracket our beliefs about which acts are wrong.

If each person chose the principles whose acceptance would make

things go best, this choice would itself make things go best, in the impartial reason-involving sense. In other words, this is the choice that each person would have the strongest impartial outcome-given reasons to make. And, as we have just seen, these reasons would not here be outweighed by any other impartial reasons. We can therefore claim that

> (E) What each person would have most reason to choose, from an impartial point of view, is that everyone accepts the principles whose acceptance would make things best.

According to

> (C) Each person would have sufficient reasons to choose that everyone accepts these principles.

(C) is not implied by (E). Though we can all respond to reasons from an impartial point of view, we also have our own, personal points of view. And the same reasons may have different strengths from these points of view. In deciding whether we have sufficient reasons to make some choice, all things considered, we must take into account both points of view. To use a different metaphor, after assessing the weights of our reasons on both the impartial and the personal scales, we must try to compare their weights on some neutral scale.

Sidgwick believed that there is no such scale. On his view, if some reasons have more weight on the impartial scale, and other reasons have more weight on the personal scale, neither set of reasons could outweigh the other. In such cases, practical reason would be divided against itself, and would have nothing more to say. Sidgwick considered this "the profoundest problem in ethics."[88]

There are two other simple views. According to desire-based theories, or the Self-interest Theory, no reasons have any weight on the impartial scale. There are no outcomes that everyone has some reason to want. According to some moral theorists, all reasons have weight only on the impartial scale. We could not rationally save our own life, or the life of someone we love, rather than saving two relevantly similar strangers. In saving ourselves, or the person we love, we would be doing what, all things considered, we had less reason to do.

These views are both, I believe, mistaken. We should not ignore or deny the strength of people's reasons from their own point of view. Such

[88] *The Methods of Ethics,* pp. 508 and 386, note 4.

impartialist theories give too little weight to the separateness of persons, or the fact that we each have only one life to live. Nor, however, should we ignore the impartial point of view. Our ability to think from this point of view is, as Nagel writes,

> an essential aspect of ourselves. Suppression of the full force of the impersonal standpoint is denial of our full humanity, and of the basis for a full recognition of the value of our own lives.[89]

We need not conclude that, as Sidgwick assumed, personal and impartial reasons cannot be compared. We should reject Sidgwick's claim that we always have sufficient reason to do whatever would be best for ourselves. We have non-self-interested personal reasons. Nor is there always a relevant distinction between the impartial point of view and our own point of view. And, when some choice would be best from our own point of view, but a different choice would be best from the impartial point of view, our personal reasons may be outweighed, all things considered. In such cases, we may not have sufficient reasons to choose what would be best, considered only from our own point of view.

There are many difficult questions about the comparability of different kinds of reason. Fortunately, we can ignore most of these questions here. We are asking which principles everyone could rationally choose, in the thought-experiments to which Kant's Formula appeals. According to premise (C), each person would have sufficient reason to choose the principles whose acceptance would make things go best.

This choice, I have argued, would be best from the impartial point of view. And I believe that

> (F) We always have sufficient reason to make the choice that is impartially best.

On my view, we have sufficient reason to give up our life, if we could thereby save two strangers. To defend (C), however, we need not appeal to (F). It is enough to appeal to a much weaker and less controversial claim.

Suppose first that, in

> *Case One,* you are rich, and you could give ten thousand dollars to some of the worst-off people in the world.

[89] *Equality and Partiality,* pp. 19–20.

If you made this gift, the loss to your well-being would be much less than the benefits to these other people. Ten thousand dollars could, for example, save many lives. Every rich person, I believe, has sufficient reason to act in this way.

Some of us would reject this claim. Even for a rich person, we might say, ten thousand dollars is a large sum, and some rich people would not have sufficient reasons to make so large a gift. But we might concede that everyone has reasons to give *some* weight to everyone else's well-being. According to what we can call

> the *strongly self-regarding* value-based view, each of us has sufficient reason to give, to the well-being of any stranger, at least one-millionth of the weight that we give to our own well-being.

Suppose next that, in

> *Case Two,* if you gave up your ten thousand dollars, the worst-off people would receive a thousand billion dollars.

Compared with the loss to you, the benefits to the worst-off people would here be more than a hundred million times as great. Even on the strongly self-regarding view, any rich person would have sufficient reason to act in this way.

This case may seem too unreal to be worth considering. But that is not so. Suppose that you live in some community that contains a hundred million rich people. Someone proposes that each of these people should pay a tax of ten thousand dollars, which would be transferred to some of the world's worst-off people. This proposal gains enough support to be put forward in a referendum. If you vote in favour, and your vote makes a difference, you would be giving up ten thousand dollars, but the worst-off people would receive a hundred million times as much.

Kant's Formula appeals to imagined choices of this kind. We ask what everyone could rationally choose, if each person supposed that she had the power to choose which principles everyone would accept. In these imagined cases, the multiplicatory factor would be much larger than a hundred million. "Everyone" here refers to all rational beings, both now and in the future. So each person should suppose that the principles she chose would be accepted by many billions of people. Consider next a choice between two sets of principles. One set of principles are *optimific,* in the sense that their acceptance would make things go

best. The other set are significantly different, since their acceptance would make things go much worse. Things would go much worse, we can assume, for many millions of people.

According to premise (C), given this choice, each person would have sufficient reason to choose the optimific principles. Since some people's interests deeply conflict, it is true of every set of principles that, compared with some alternative, their acceptance would be much worse for certain people. So there would be some people who would have strong personal reasons *not* to choose these optimific principles. Suppose that *Blue* is one such person. Blue would know that, if she chose that we all accept these principles, that would be much worse from her point of view. Both she and those she loves would die young. But Blue would also know that, if everyone accepted these principles, things would go much better for many millions of people. On plausible assumptions, several millions of these people would be saved from dying young. So, even on the strongly self-regarding view, Blue would have sufficient reasons to choose these principles.

As before, I am not claiming that Blue would be rationally required to choose these principles. Perhaps Blue could rationally choose the principles whose acceptance would be best for her and those she loves. My claim is only that Blue *could* rationally choose the optimific principles, since she would have sufficient reasons to make this choice. Blue has strong reasons to care about her own well-being, and the well-being of those she loves. But she could rationally choose these principles if, as I believe, she has sufficient reasons to care at least one-millionth as much about the well-being of other people. Blue could rationally give up her life, if she could thereby save a million other people.

To reject premise (C), we would have to appeal to a purely desire-based theory, or the Self-interest Theory. We would have to claim that some people have no reason to care at all—not even very slightly—about the suffering and deaths of strangers. That, I believe, is not true.

Consider next those sets of principles that are close to being optimific, since their acceptance would make things go only slightly worse. The claims just made may not apply to such principles. But this point would not have much importance. We could revise premise (C), so that it applied to the principles whose acceptance would make things go best, or be close to doing that.

According to premise (D), with the same revision, it is only such principles that everyone could rationally choose. (D) compares such

principles with those that are significantly different, since their acceptance would make things go much worse. According to (D), there is no such set of principles that everyone could rationally choose.

It is true of any such set of principles that, compared with the optimific principles, their acceptance would be much worse for certain people, and for those they love. These people would have strong personal reasons not to choose these principles. These people would also know that, compared with the optimific principles, the acceptance of these other principles would, on the whole, make things go much worse, by being much worse for many more people. So these people would also have strong impartial reasons not to choose these principles. Since these people would have both personal and impartial reasons not to choose these principles, they could not rationally choose these principles.

Premises (B) to (D) are all, I conclude, true. Everyone could rationally choose the principles whose acceptance would make things go best. And these are the only principles that everyone could rationally choose. So, on Kant's Formula, these are the principles on which we ought to act. That is what Rule Consequentialists believe.

As well as providing an argument for Rule Consequentialism, Kant's Formula may give this view a stronger foundation. According to some Rule Consequentialists, all that ultimately matters is how well things go. There is a strong objection to this view. If all that matters is how well things go, why is it wrong to act in ways that make things go best, when and because such acts are not permitted by the optimific principles?

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 act on the principles whose universal acceptance everyone could rationally choose. According to these Contractualist Rule Consequentialists, what everyone could rationally choose is that everyone accepts the principles whose acceptance would make things go best. Because this view does not assume that all that ultimately matters is how well things go, it may better answer the objection that it cannot be wrong to do what would make things go best. Such acts may be wrong, not because they contravene the principles whose acceptance would make things go best, but because they contravene the principles whose acceptance everyone could rationally choose.

I have not claimed that we should become Rule Consequentialists.

My aim has been, in part, cartographical. On one standard moral map, most of us are shown as holding pluralist views, of the kind that Sidgwick called "common sense morality." This map also shows a few kinds of systematic theory. One kind are consequentialist, with utilitarian theories as the main examples. Two others are Kantian and contractualist theories. These theories are often claimed to be the main systematic rivals to consequentialism. That, I have argued, is not true. Of the different ways of thinking about morality, it is Kantian and contractualist theories that do most to support consequentialism. So this moral map should be redrawn.

This fact is not surprising. Rawls writes that, in adopting his contractualist formula, we "have substituted for an ethical judgment a judgment of rational prudence."[90] In applying Rawls's Formula, we cannot appeal to our beliefs about which acts are wrong. So we should expect to reach utilitarian conclusions. As I have said, utilitarianism is, roughly, rational prudence plus impartiality.

Rawls's aim is, in part, to provide an alternative to utilitarianism. By turning to Kant's Contractualist Formula, I have argued, we can achieve this aim. On Kant's Formula, we appeal to the principles whose choice would be rational, not in purely self-interested terms, but in response to every kind of reason. These principles, we can argue, are not utilitarian. By appealing to Kant's Formula, I believe, we can support various distributive principles. And this formula may allow us to defend other nonutilitarian principles, such as principles about retributive justice, or some kinds of perfectionism.

Though Kant's Formula counts against utilitarianism, it supports consequentialism. As before, that is not surprising. On this formula, we achieve impartiality by appealing to the principles whose universal acceptance everyone could rationally choose. And, in applying this formula, we cannot appeal to our beliefs about which acts are wrong. So we should expect to reach consequentialist conclusions. What everyone could rationally choose are the principles whose acceptance would make things go best. Consequentialism is, roughly, wide value-based rationality plus impartiality.

If we believe that we ought to avoid consequentialist conclusions, the obvious course is to think about morality not in a Kantian or contractualist way, but in Sidgwick's way. That claim may seem surprising,

[90] *TJ,* p. 44, *RE,* p. 39.

since Sidgwick is an Act Utilitarian. But Sidgwick assumes that we must appeal to our beliefs about which acts are wrong. This appeal should be critical, since we should try to clarify these beliefs, and we should consider all good arguments for and against them. But we cannot hope to answer all moral questions by appealing only to claims about what everyone could rationally choose.

There is an alternative. Scanlon claims that, rather than appealing to the principles that everyone could rationally choose, we should appeal to the principles that no one could reasonably reject. Scanlon's contractualist formula appeals to claims about what is reasonable in a moral sense. This does not, as some suggest, make this formula circular, or trivial. And Scanlon's version of contractualism may support some nonconsequentialist principles. Compared with the Kantian version of contractualism that I have been discussing, only Scanlon's version, I believe, may be as good. But that is a subject for another day.