How I am not a Kantian

HOW I AM NOT A KANTIAN [1]

T. M. SCANLON

On What Matters begins with a vigorous defense of a cognitivist and value-based account of reasons. It ends with a striking claim of a convergence between Kantian, Consequentialist and Contractualist moral theories. In these comments I will concentrate on the relation between these two parts of Parfit's rich and provocative book.

Questions about reasons are fundamental to Parfit's conclusion because the theories whose convergence is in question all characterize right and wrong in terms of what people have reason to want, or could rationally do. The three theories Parfit is considering are:

> *The Kantian Contractualist Formula*: Everyone ought to follow the principles whose universal acceptance everyone could rationally will.
>
> *Scanlon's Formula*: An act is wrong if it would be disallowed by any principle that no one could reasonably reject.
>
> *Kantian Rule Consequentialism*: Everyone ought to follow the principles that are optimific, because these are the only principles that everyone could rationally will to be universal laws.

Parfit acknowledges that the two theories he labels "Kantian" diverge from what Kant himself said. But he regards this as no objection to what he is doing. "We are asking," he writes, "whether Kant's ideas can help us to decide which acts are wrong, and help to explain why these acts are wrong. If we can revise Kant's formulas in a way that improves them, we are developing a Kantian moral theory" (p. 000).

I agree that it can be a valuable project to develop a moral theory that is similar to Kant's in some ways but departs from it in others. But I believe that one of the ways in which the theories Parfit lays out diverge from Kant's own view deserves attention. The degree to which Parfit's conclusion should seem surprising depends to a certain extent on how close the theories he is discussing are to Kant's. More important, an examination of one way in which these theories differ from Kant's will bring out some of the difficulties faced by an account of reasons of the kind that Parfit and I favor, and hence also by a moral theory based on such an account.

I will not engage in detailed exegesis of Kant's texts, but will base my discussion of these issues on a few broad claims about Kant's view of rationality and

---

[1] I am grateful to Derek Parfit for many discussions of these issues as well as for helpful comments on an earlier version of this paper.

morality which I hope are relatively uncontroversial. For simplicity, I will concentrate on Kant's Formula of Universal Law, and on Kant's discussion of this formula in his *Groundwork of the Metaphysics of Morals*. A full discussion would need to take into account other formulations of the Categorical Imperative as well as what Kant says in other works. But this will suffice for the mainly comparative points that I want to make.

I begin with an observation about the way in which Kant sees the Categorical Imperative as authoritative for us. What he says in Section 3 of the *Groundwork* is that when we are deciding what to do we must *see* the Categorical Imperative as our highest level principle of practical reasoning insofar as we see ourselves as acting at all. If we take any other principle to be fundamental for us, then we cannot see ourselves as acting but only as the slaves of factors acting on us. This claim depends in turn on Kant's argument, in Section 2 of the *Groundwork*, that there can be only one categorical imperative (that is, that any principle other than the one he has presented could influence an agent only though its appeal to his or her inclinations.) Thus, in Kant's view it is only if one takes the Categorical Imperative as the fundamental principle of practical reasoning that one can see oneself as *deciding* what to do rather than merely being determined by one's inclinations.

Turning now from the authority of the Categorical Imperative to its content, the Formula of Universal Law says that one should act only on a maxim that one could will to be a universal law. I believe that the best interpretation of what Kant means by a maxim's being a universal law is for everyone to believe it to be permissible to act on that maxim, and to act on it when they are so inclined. The crucial questions in determining what this formula requires are thus: (1) what, in Kant's view, would prevent a maxim from even being a universal law in this sense, and (2) what would make it the case that a maxim could not be willed to be such a law.[2]

Kant's idea seems to be that a maxim "cannot be a universal law" in the sense he has in mind if the plan of action it describes would be incoherent in the event that people's attitudes were of the kind that this universal law describes. The "contradiction" that he is appealing to is thus between the presuppositions of the plan of action that the maxim describes and the conditions that would obtain if this maxim were a universal law. The most plausible example of this is Kant's case of the lying promise: making a promise would not be an effective way of

---

[2] Parfit discusses these questions in sections 25 and 26 respectively. My interpretations of these Kantian ideas differ slightly from his. The claim that it is wrong to act on a maxim that one could not rationally will to be a universal law in the sense I have just described is similar to what Parfit calls the *Law of Nature Formula* except that it substitutes for the phrase "and acts on it when they can" the phrase "and acts on it when they are so inclined." My version of the claim differs from what Parfit calls the *Moral Belief Formula* because it requires one to be able to will not only that everyone believes it to be permissible to act on the maxim in question, but that they also act on it when they are so inclined.

getting the money one desires if everyone believed that having made such a promise was no constraint on anyone's future conduct. Parfit may be right that the terms "contradiction" and "cannot be a universal law" are not the best way to put this point. But I think it is reasonably clear what Kant has in mind.

Parfit's understanding of the idea of something's being rationally willed to be a universal law is different from Kant's as I interpret him. When Parfit asks, in interpreting the various formulae he discusses, whether an action or principle is one that someone could rationally will, he understands this as a question about the reasons that person has, and their relative strengths. One can rationally will something, on his view, if one has sufficient reason to do so; one cannot rationally will it if one's reasons not to will it are stronger than one's reasons to will it (p. 000). Kant's idea of what one can will is different. When he considers the question of whether a given maxim could or could not be willed to be a universal law Kant seems not to appeal at all, or at least not in a fundamental way, to reasons or their relative strength.[3] Indeed, the idea of a reason and of the strength of a reason have at most a derivative role in Kant's account of rational action and morality.[4]

When Kant says that a maxim could not be willed to be a universal law, what he means is that willing such a law (willing that everyone act on the maxim should he or she be so inclined and believe that others will do this as well) would be incompatible with viewing oneself as a rational agent. For example, Kant claims that a maxim of developing one's talents only insofar as one finds this pleasant or attractive, or a maxim of helping others only if it happens to please one, could not be willed to be universal laws, because in willing these laws one would be willing that one give, and that others give, no intrinsic weight to the existence of general conditions that are necessary to the pursuit of our ends. To be a rational agent, however, is to have ends, and one cannot (without being irrational) have ends yet be indifferent to the conditions necessary for their pursuit. The "contradiction" that Kant has in mind is thus grounded in the same thing that (as I maintained earlier) Kant believes grounds the authority of the Categorical Imperative itself, namely the views one must take insofar as one sees oneself as a rational agent.

Kant's claims about what the Formula of Universal law requires are thus not based on claims about what reasons individuals have, or about the relative

---

[3] To act on a maxim is to act for a certain reason. So in asking whether one could will that people act on, or be permitted to act on a maxim, the idea of a reason for action figures in what one is asking *about*. What I am saying is that for Kant such questions are not to be *answered* by appeal to the reasons an agent has.

[4] In an earlier version of the manuscript that became this book, Parfit expressed surprise that Kant seemed not to employ the idea of a reason in the normative sense in which Parfit understands it. My point here is that this observation was correct in a way, but less surprising than it might at first appear.

strength of these reasons. When his claim is that a certain maxim could not *be* a universal law (as in the case of the lying promise), the question of what one can will does not even arise. When his claim is that we cannot *will* a maxim to be a universal law (such as a maxim of indifference to the development of our talents, or to the needs of others), his claim is not that the reasons we have not to will such laws are stronger than those in favor of doing so. What Kant says is rather that insofar as we see ourselves as rational agents we cannot see the development of our talents or the needs of others as considerations that in themselves count for nothing. The claims that provide the basis for Kant's arguments are claims about rationality—about the attitudes we must hold insofar as we are not irrational—not claims about the reasons we have.[5] Accordingly, the *conclusions* of these arguments are also claims that we must, insofar as we are not irrational, *see* these things—the development of our talents and the needs of others—as providing reasons for action rather than substantive claims about the reasons we have.

I should note, however, that as I have interpreted Kant's arguments about what one can will to be a universal law, their conclusions make only the most minimal claim about the strength we must see certain considerations as having. The claim is just that we cannot take these considerations—the development of our own talents and the needs of others—as counting for nothing (apart from their appeal to our inclinations.) If this interpretation is correct, and this minimal conclusion is all that Kant's argument yields, then it is left up to each person to determine (depending, I suppose, on his or her inclinations) how much weight to give to these considerations. But perhaps Kant's argument actually yields a stronger conclusion. Perhaps Kant could establish that a person who sees him or herself as a rational agent cannot consistently will a maxim of not helping others or doing what is required to develop his or her talents when these aims come into conflict with certain considerations of convenience or comfort.

It might seem that in order to establish such a conclusion Kant would have to appeal to premises about the relative strength of reasons: that is, it would have to rest on a claim that the possibility of enjoying the forms of convenience or comfort in question is not a sufficient reason for failing to develop one's talents in certain ways, or for failing to aid someone else in a certain way. But from the Kantian point of view as I am interpreting it this would be to get things backwards. Claims about reasons (more exactly, about what a person must see as reasons) must be grounded in claims about rational agency, claims about what attitudes a person can take, consistent with seeing herself as a rational agent.

---

[5]I discuss this distinction further in "Reasons: A Puzzling Duality?" in R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith, eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (New York: Oxford University Press, 2004), pp. 231-246, and in "Structural Irrationality," in Geoffrey Brennan, Robert Goodin, Frank Jackson, and Michael Smith, eds., *Common Minds: Essays in Honor of Philip Pettit*, (Oxford: Oxford University Press, 2007).

Justification never runs in the other direction, from claims about reasons to claims about what rationality requires.

This view, which I will call Kantian constructivism about reasons, seems to me to be a fundamental feature of Kantian ethical theories, distinguishing them from other views that resemble Kant's in some ways. In particular, as I have said, it distinguishes Kant's view from all of the moral views that Parfit discusses in Part Three of *On What Matters*. All of these views, including those described as Kantian, appeal to an idea of "what one can rationally will" that presupposes an independently understandable notion of the reasons that a person has and their relative strength. So there is one sense in which none of these views is Kantian: none of them accepts Kantian constructivism about reasons. This divergence raises questions facing in two directions. Negatively, why *not* accept Kantian constructivism about reasons? Positively, what can be said in defense of the alternative conception of reasons that Parfit employs, and that I myself would also favor?

On the negative side, Parfit raises objections to what he calls Kant's Impossibility Formula, according to which it is wrong to act on maxims that could not even *be* universal laws.[6] These objections mainly take the form of arguments that Kant's remarks about what could not be a universal law cannot be interpreted in a way that avoids intuitively implausible implications about moral right and wrong. I agree with many of the points Parfit makes here, although I would put them in a somewhat different way.

The "contradiction in conception" test[7] is intuitively appealing because it seems to capture the idea that it is wrong to exempt oneself from the moral requirements that apply to everyone else. Many wrongs do fit this pattern: if certain constraints are needed to provide some essential public good (or to prevent some serious "public bad"), and people are generally complying with them, then it is wrong to free ride on their compliance by exempting oneself from these constraints. But Kant's test does not track this idea in a reliable way.

The class of actions that Kant's test captures are ones in which an agent's plan of action presupposes that others believe that everyone is bound by constraints that rule out action of the kind that the agent is going to perform. The problem is that by focusing on the relation between an agent's action and what that action presupposes about the beliefs and intentions of others this test bypasses the question of whether the constraints in question are indeed justified. (This may be part of the appeal of Kant's test: it seems to provide a criterion of wrongness that can be applied without asking messy questions about the relative strength of reasons.) But the question of justification is essential. If the constraint that others take to be binding is in fact groundless (a mere taboo, for example) then it

---

[6] See, for example, p. 228.

[7] Parfit refers to this test as "Kant's actual version of his Impossibility Formula" (p. 161.)

may not be wrong to violate this constraint, even if the success of one's action depends on the fact that most others take that constraint seriously. On the other hand, when constraints are necessary and justified, then it is wrong to violate them whether or not the success of *this very action* depends on the fact that others take these constraints to be binding and generally observe them. Everything depends on the need for the constraints in question, not merely on whether the success of one's action depends on their being generally observed.

What is commonly called Kant's "contradiction in the will" test might be called upon to answer this question of justification. The idea would be that to determine whether a constraint is justified we should ask whether one could will that it be generally believed to be permissible to violate this constraint when this suits one's purposes. As Parfit says, this criterion of justifiability is similar to the version of contractualism that I myself have proposed.

One way in which Kant's criterion appears to differ from mine, and Parfit's, is in focusing simply on whether *the agent* could will a principle permitting what he or she proposes to do, rather than on whether there is anyone who could reasonably reject a principle permitting such actions, or whether everyone could will the universal acceptance of such a principle. The question here is how a mode of thinking about right and wrong is to be sensitive to the interests of other people. Different theories solve this problem in different ways.

I believe that on the best interpretation of the way Kant understands his Formula of Universal Law, when we ask whether an agent could will his maxim to be a universal law what we are asking is whether he could will that people be universally permitted to act on such a maxim, where this universality includes situations in which the agent occupies any of the positions involved—for example, situations in which the agent is a person in need of help as well as ones in which he or she is the one called upon to give it. Assuming that this idea is intelligible, and that if the agent were in one of these other positions he or she would have the same reasons as a person who is actually in that position, this test would seem to lead to the same result as asking, as Parfit suggests, whether *everyone* could will this universal permission. Even if this is so, however, I agree with Parfit that it makes things clearer to avoid counterfactuals about the agent's being in different positions and to keep clearly in view the fact that we are dealing with different persons, by asking what everyone in these other positions could will, or could reasonably reject.

Another possible divergence from Kant arises when we consider how the idea of what someone could rationally will is to be understood. One might object to Kant's account of this idea on the ground that its implications about the reasons we have are inadequate or implausible. I have mentioned two objections of this kind. The first is that Kant's account yields only conclusions about what individuals must see as reasons, insofar as they are not irrational. It seems to me, however, that there are true substantive claims about the reasons we have that are different from claims of this kind and cannot be derived from them. Second, leaving aside the difference between these two kinds of claims, I do not

believe that the idea of rational agency is rich enough to yield all the claims about reasons that seem evidently correct.

Going beyond objections of this kind, however, if we are going to reject Kant's account we need to consider the deeper question of where his argument for the Categorical Imperative as the limiting ground of the reasons we have goes wrong, if it does go wrong. Here I would cite Kant's claim that accepting the Categorical Imperative as one's highest level principle of practical reasoning is the only way in which one can see oneself as acting independent of inclination. This claim strikes me as untenable. I do not see why an agent cannot see him or herself as "active" in making judgments about which considerations constitute reasons.[8]

Kant offers a top-down conception of reasons (or at least of our states of taking things to be reasons.) In his view, claims about reasons are grounded in the requirements of rational agency. If this account is rejected, the alternative might seem to be a "bottom up" conception, according to which practical reasoning begins with claims about particular reasons and their relative strengths and proceeds "upward" from there to conclusions about what we have most reasons to do or to think, taking all the relevant reasons into account. A desire-based theory of reasons for action would at least appear to be of this form. Such a view holds that if doing X would promote the satisfaction of some desire that an agent has, then that agent has at least a *pro tanto* reason to do X. What an agent has most reason to do all things considered is determined by balancing these various, and possibly conflicting, reasons.

Parfit considers and rejects desire-based theories in his Chapters 3 and 4. What provides us with reasons for action, he says, are not desires but the various facts about certain aims and acts that make them relevantly good, or worth achieving. Reasons are provided by considerations such as the fact that doing X would injure someone, or would save someone's life. This seems right to me. But when we focus simply on such considerations, considered individually, as ultimate reason-providers, a bottom-up view can be made to seem implausible. Do we really want to claim, it might be asked, that such considerations, in addition to their physical and psychological properties can have the additional normative property of providing a reason of a certain strength, and that the basis of practical reasoning lies in detecting these properties? Put in this way, this does seem odd. But the oddness results, I believe, from the fact that this way of putting things ignores several crucial aspects of reasons.

---

[8] It might be suggested that one can avoid these problems, and also provide the basis for a more extensive set of reasons, by appealing to Kant's Formula of Humanity—that is, to the idea that each person must regard his or her own rational nature (and that of others as well) as an end in itself. I do not believe that this line of argument is any more successful than the one I have sketched, but it would take me too far afield to examine it here.

One thing that seems odd about this atomistic formulation is that it leaves out the relational character of reasons, and their dependence on context. A certain consideration does not provide a reason of a certain sort, full stop. It provides a reason for an agent, in a certain situation, to take a certain action, or to have a certain attitude. The same consideration can provide different reasons in this fuller sense depending on the agent, situation, and attitude involved. Similarly, the "strength" of a reason—that is to say, the way in which one consideration can override, undermine, or be overridden or undermined by other considerations—depends on the context within which a decision is being made.

A desire-based theory gains some of its plausibility from the fact that it has a certain relational structure built in. A desire is a desire *for* a certain content, but it is also the desire *of* a particular agent, a desire of a particular strength, and it provides reason for different actions depending on that agent's situation. One weakness of a desire-based theory is that the relational structure that it provides is too limited. Insofar as a desire is just a desire of a certain strength for a certain outcome, it provides reasons for actions that would promote that outcome. But not all reasons are goal-directed in this way, and we have reasons for things other than actions. An adequate account of reasons needs to accommodate these facts.

The contrast with the atomistic realism I mentioned earlier brings out another feature of desire-based theories that should be noted, which is that their "bottom up" character is more apparent than real. Desires derive their reason giving force because they are the desires of some desiring agent. In this respect a desire-based theory is similar to the Kantian view, but it focuses of a different aspect of agency and, at least as I have formulated it, yields conclusions about the reasons that an agent has, rather than about what an agent must see as a reason insofar as he or she is rational.

But even if a desire-based theory offers a top down account of the source of reasons, its account of the process of practical reasoning remains bottom up: it sees practical reasoning as beginning with our experience of individual desires and their strength. An atomistic realism about reasons that preserved this bottom up character would share this implausibility. We do not experience considerations one by one as reasons with a certain strength. Rather, to regard one consideration is a stronger reason than another is to see it as more important *in regard to a certain type of decision in a certain context.* For example, whether the fact that it would be fun to make a certain remark counts as a strong reason for making it depends on the context, on what my aims and responsibilities are, and on my relation with the others present. Moreover, judgments about reasons and their importance are subject to requirements of consistency: if I judge A to be a reason for some action in one context, and a stronger reason than B, then I must judge this to be so in other contexts and for other agents as well, unless I can cite some relevant difference between these situations.

This discussion suggests several conclusions about what an adequate account of reasons must be like: It must preserve the idea that questions about reasons arise for, and are about, agents facing certain decisions. Second, it must be holistic in

the way just described: judgments about particular reasons and their relative strengths depend on an overall view of the reasons we have. The strength of the Kantian view lies in its recognition of these important points. But an account of reasons must be substantive: it must include claims about the reasons that agents have, rather than merely about what they must see as reasons. And these claims cannot be derived solely from the agents' desires or from the mere fact that they are rational agents. If I am correct about this, then an adequate account of reasons will be a kind of substantive holism.

I turn now to Parfit's striking claim, in his Chapter 16, that Contractualism and Rule Consequentialism converge or, more exactly, that what he calls Kantian Contractualism will coincide with Rule Consequentialism. I hope that an examination of his careful arguments will help to bring out what is distinctive about a Contractualist theory of the kind I have proposed, and how such a theory would differ from Rule Consequentialism even if the two were to support the same principles.

I will begin with what Parfit calls *The Kantian Contractualist Formula*:

> Everyone ought to follow the principles whose universal acceptance everyone could rationally will (p. 000).

As I have said, Parfit understands the question of what someone could rationally will as a question about what is supported by the overall balance of reasons that that person has. In his view, an agent can rationally will that certain principles be universally accepted just in case he or she has sufficient reason to will this. So the interpretation of the Kantian Contractualist Formula depends, as Parfit says, on claims about reasons and rationality. This formula will yield definite answers about what we ought to do in a given case only if there is a single principle (applicable to our situation) which everyone has sufficient reason to will to be universally accepted. Parfit calls this the uniqueness condition (p. 000) Given some views of the reasons a person has, this condition will not be fulfilled because there will be no principles that everyone has sufficient reason to will. Perhaps Rational Egoism is an example of such a view. [9]

Different moral theories deal with this problem in different ways. Rawls assumes that people will lack concern for how others fare (they will be "mutually disinterested"), but requires that they choose principles behind a veil of ignorance. My own version of contractualism deals with the problem by making particular stipulations about the reasons that are relevant to the choice of principles and the ways that these are to be considered.[10] The view that Parfit calls Kantian

---

[9] As Parfit argues (227-228.) David Gauthier might disagree.

[10] Restricting these to what I call "personal reasons." See *What We Owe to Each Other*, pp. xxx

Contractualism makes neither of these moves. On this view, what we ought morally to do depends on what everyone could rationally will, with full information about their situation and taking into account all the reasons they in fact have. Parfit believes that the uniqueness condition is fulfilled "sufficiently often" (p. 000) because the reasons people have include impartial reasons as well as personal and partial ones.

Impartial reasons, he says, are reasons we see that we have when we consider matters from an impartial point of view—that is to say, without considering our own place in a situation. We take such a view when, for example, we are, or suppose ourselves to be, merely an outside observer of what happens rather than one of the people whose well-being, or that of others to whom they have close ties, will be affected by it. Central among these impartial reasons are reasons to care about the well-being of others, but our impartial reasons may also include reasons to care about things other than individuals' welfare. Parfit argues that we have these same impartial reasons when we consider matters from our own personal perspective. (68) What the shift to the personal perspective does is merely to add personal and partial reasons to the impartial ones.[11]

A decision about what someone can rationally will must take all of these reasons into account. In some cases, the impartial reasons may predominate: one would not have sufficient reason to do something that would lead to the death of many people just to avoid scratching one's finger. In other cases the opposite will be true: one would not have sufficient reason to sacrifice one's life to prevent the scratching of one other person's finger (or, I would say, any number of persons' fingers.) But Parfit believes that there are many cases in which neither kind of reasons predominate in this way. In such cases, he writes,

> When one possible act would be impartially best, but some other act would be best either for ourselves or for those to whom we have close ties, we often have sufficient reasons to act in either way (p.000).

Parfit believes that the uniqueness condition is fulfilled "sufficiently often" because there are certain principles that everyone has sufficient impartial reason to will to be universally accepted, even though they may have personal and partial reasons to prefer other principles.

Parfit defines the idea of "best outcome" in terms of the idea of impartial reason. We should call an outcome "best," he writes, just in case it is "the outcome that, from an

---

[11] This brings out the fact that the idea of a "point of view" is merely an expository device, a way of focusing our attention. Impartial reasons are not the reasons we *have* from a certain point of view. They are reasons we have *independent of* our particular relation to their objects, in contrast to personal reasons (to care about ourselves) or partial reasons (to care about others to whom we stand in certain special relations.) When we "take up the impartial point of view" we ignore these relations, and thus are aware only of reasons that do not depend on them.

impartial point of view, everyone would have most reason to want" (p. 000). He does not say very much about which outcomes will be best in the sense he defines. In particular, he leaves it open to what degree this idea of bestness will be aggregative: will an outcome containing a greater sum of well-being be better than one which contains less aggregate well-being no matter how well-being is distributed in the two situations? For example, will a situation in which greater total well-being count as better if this total is produced by significant costs to a few people which however bring small benefits to a very great number? As Parfit sets things up, this will depend on whether people have impartial reasons for favoring one of these states over the other. This leaves open the possibility that conception of best outcome he is defining is in important respects non-aggregative.

Using the notion of best outcome, Parfit defines universal acceptance rule consequentialism as the view that

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

He argues that this view is a direct consequence of

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

His argument for this proceeds as follows:[12]

> Kantians could argue:
>
> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally will, or choose.
>
> (B) Anyone could rationally choose whatever they would have sufficient reasons to choose.
>
> (C) There are some optimific principles whose universal acceptance would make things go best.
>
> (D) These are the principles that everyone would have the strongest impartial reasons to choose.
>
> (E) No one's impartial reasons to choose these principles would be decisively outweighed by any relevant conflicting reasons.
>
> Therefore
>
> (F) Everyone would have sufficient reasons to choose these optimific principles.

---

[12] On pp. 246-247.

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

Therefore

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

Therefore

These are the principles that everyone ought to follow.

I do not dispute Parfit's conclusion about the relation between his Kantian Contractualism and Rule Consequentialism. What I want to concentrate on here is what this connection shows about the ways in which the structure of his Kantian Contractualism differs from the version of contractualism presented in my book.

Parfit says that according to Kantian Contractualism, in order to decide whether an action is permissible we must assess a principle that would permit it by conducting number of thought experiments, one for each person. In each of these we ask whether one of these persons could rationally will a principle that would permit such an action. This question is to be answered by considering both the person's personal and partial reasons and his or her impartial reasons. Suppose that the person's impartial reasons support accepting the principle. If the person has personal or partial reasons for not accepting the principle, the question we are to ask is whether, despite these reasons, the person nonetheless has sufficient reason to choose that everyone accept the principles that impartial reasons favor. As we have seen, Parfit holds that this might be true even if the person has sufficient reason to choose the principle that his or her personal and partial reasons favor.

According to my version of contractualism, deciding whether an action is right or wrong also involves a series of thought experiments. These consist in asking, in the case of each person considered, whether that person could reasonably reject a principle that would permit the action in question.[13] As in the previous case, suppose that one such person, call her *Green*, has personal reasons for rejecting the principle in question because of the burdens it would require her to

---

[13] Parfit and I may take different views about the correct characterization of the "individuals" whose reasons are to be considered. Although he does not say so explicitly, some of what he does say suggests that he has in mind actual persons affected by the action, or by the acceptance of the principle. In my case what we consider are not the reasons of actual persons but the "generic" reasons that someone would have in virtue of occupying a certain role in regard to the principle in question, such as being the person who has relied on the assurance of others, or a person in need of help, or a person called upon to give it. I discuss this issue in *What We Owe to Each Other*, pp. 202-206.

bear. According to my version of contractualism, to decide whether Green could reasonably reject the principle we need to consider the opposing reasons that others, considered individually, have for wanting the principle to be accepted. This involves a further series of thought experiments, corresponding to the various ways that people might be affected by the principle in question. In each case we are to ask whether, given the reasons that a person in the position in question would have for wanting the principle to be accepted, it would still be reasonable for J to reject it. The reasons that we consider here, in opposition to Green's personal reasons for rejecting the principle, *correspond* to reasons that Green would have if she took an impartial view of the situation, but there is a significant difference. In the form of contractualism that I have proposed, what we are to consider are not two kinds of reasons that Green might have (such as personal reasons and impartial ones) but, rather, the reasons that individuals in two different positions have: Green's reasons and those that a person would have who would be affected by the principle in a different way than Green would be.

The difference between these two ways of interpreting the reasons that someone might have for accepting a principle, or not rejecting it, can be illustrated by considering the way in which Parfit deals with a potential objection to his argument that Kantian Contractualism leads to Rule Consequentialism. Imagine a lifeboat case in which one is faced with the choice between saving five strangers and saving one's own child. Parfit believes that in such a case one would have decisive reason to save one's child. It may appear that optimific principles would require one to save the five strangers. If this were so then one might have decisive reason to reject these optimific principles, despite the impartial reasons in favor of willing their universal acceptance, contrary to premise (E) of Parfit's argument in the passage I have quoted above. Parfit responds as follows:

The optimific principles would *not*, however, require you to save the strangers rather than your child. If everyone accepted and many people followed such a requirement, things would go in one way better, since more people's lives would be saved. But these good effects would be massively outweighed by the ways in which it would be worse if we all had the motives that such acts would need. For it to be true that we would save several strangers rather than one of our own children, our love for our children would have to be much weaker. The weakening of such love would both be in itself bad, and have many bad effects. Given these and some other similar facts, the optimific principles would often permit us, and often require us, to give some kinds of strong priority to our own children's well-being (p. 000).

This line of argument is familiar from the literature on consequentialism.[14] It has a distinctively consequentialist flavor because it appeals to what would be best overall—the kind of outcome that everyone has most impartial reason to prefer. I make a similar point within my version of contractualism, but with an important

---

[14] See, for example, Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," *Philosophy & Public Affairs*13 (1984), pp. 134-171.

difference.[15] Rather than appealing to the idea of the best outcome—what everyone has impartial reason to prefer—my argument was based on what each individual has reason to want for him or herself. A principle requiring us always to give the needs of strangers the same weight as those of friends and family members would be one that each of us could reasonably reject, because it would make impossible special relationships that we have strong reasons to want to have. Even if these two arguments lead to the same conclusion, and assign normative significance to the same facts about human life, they take these facts into account in different ways.

As I said above, according to my version of contractualism the considerations that we need to consider in order to decide whether it would be reasonable for J to reject a principle take the form of reasons that others would have to want that principle to be accepted. In Parfit's Kantian Contractualism these considerations enter in the form of impartial reasons that J has to want the principle to be accepted. But these are only some of the impartial reasons that could count in favor of Green's accepting the principle according to Parfit's Kantian Contractualism. Two differences are particularly significant. First, in addition to reasons corresponding to the reasons that other individuals have to want things to go better for them, Green's impartial reasons as Parfit would describe them can include impartial reasons that Green has for wanting more people to be benefited rather than fewer, or for the aggregate benefit to be as great as possible. According to the version of contractualism described in my book, however, what is to be taken into account in assessing the reasonableness of a person's rejecting a principle are only the reasons that *each* affected person has for wanting that principle to be accepted. Aggregative considerations are not directly relevant. Second, my view excluded impersonal reasons such as those associated with the value of natural objects or works of art, considered apart from the benefits to individuals of being able to experience these things. But impartial reasons as Parfit describes them could include reasons of this kind.

These two differences may be seen as improvements over the view stated in my book, which seemed implausible to many because it excluded aggregative arguments and because it gave no weight to impersonal values in determining what is right or wrong. These objections could be dealt with by allowing reasons of these two kinds to be considered in determining whether a principle could be reasonably rejected.[16]

It is worth saying a little more here about the way in which the problem of aggregation is dealt with in Parfit's Kantian Contractualism, and therefore would be dealt with on this revised version of my view. The problem of aggregation is this. There are many cases in which what we should do, and even what it is

---

[15] See *What We Owe to Each Other*, pp. 160-161.

[16] Parfit has previously urged that I should make this change by giving up my "Individualist Restriction" on reasons for rejection. See his article "Justifiability to Each Person", in *On What We Owe To Each Other*, Philip Stratton-Lake, ed., (Blackwell, 2004), pages 67-8.

permissible to do, seems to depend on the number of people who would be affected by the courses of action available to us. It seems that an adequate account of moral argument should make aggregative considerations relevant in these cases but do this in a way that does not support implausible aggregative arguments such as ones what would justify the killing or enslaving of a few people to make a huge number of people better off, each in a very small way.

Parfit's proposal, as I understand it, is to deal with this as a problem about which outcomes are indeed "best" (that is to say, ones that everyone has impartial reasons to prefer.) So he would say that in a case of the kind I have just considered the fact that aggregate well-being would be increased by enslaving a few people in order to benefit a great many people in small ways does not mean that a situation in which this was done would be one that we have impartial reason to prefer: the idea of "best outcome" is sensitive to numbers, but is not strictly aggregative. I leave aside the question of how such an account of impartial reasons and "best outcome" might be spelled out.

I have been discussing different views about the reasons that should be taken into account in deciding whether a principle is one that everyone could will to be universally accepted, or whether it is one that could reasonably be rejected. Let me turn now to the importance of the difference between these two ways of understanding the question we should ask in carrying out the thought experiments on which the rightness or wrongness of an action depends. According to Parfit's Kantian Contractualism one is to ask whether each person could rationally will that a principle permitting that action be universally accepted. On my view one is to ask whether every such principle would be one that someone could reasonably reject. How might the differences between these questions lead to different answers about which actions are right?

As we have seen, Parfit allows that there are many cases in which a person has sufficient impartial reasons to accept a principle but also sufficient self-interested reasons to refuse to do so. It seems possible that in some cases of this kind it would be reasonable for the person to reject the principle in question. It might be that the universal acceptance of the principle would involve a cost that the person would have sufficient reason to accept (it would not be like a case of losing one's life because this would prevent the scratching of someone else's finger.) But this would also be a cost that a person could reasonably refuse to make. If there are cases of this kind, then Kantian Contractualism would involve higher costs than my version of contractualism would.

It will be helpful to divide possible cases into two types. In cases of the first type, although following the optimific principle would involve a major cost to someone, another person would suffer an even graver loss if the optimific principle were not followed. In cases of the second type this is not so: the sacrifice required of one person by the optimific principle is greater than the loss that any other individual would suffer if everyone were to follow some non-optimific principle.

Here is a possible case of the first type. Suppose that, in

> *Case One*, by giving some organ of his for transplant, *Grey* would be shortening his life by a few years. But by doing this he could give White, whom he does not know, many more years of life.

If this is so, then Grey would have sufficient impartial reason to donate the organ, and the outcome, if he were to do so, would be better in Parfit's impartial reason-involving sense. But Grey would also have sufficient self-interested reason not to make this donation. Moreover, it seems plausible to say that it would be reasonable for someone in Grey's position to reject a principle requiring this person to make such a donation.

Cases of the second type would involve two principles, P, which is optimific and imposes a high cost on people in the position of *Blue*, and Q which does not impose that high a cost on anyone (there is no one who would lose as much by a shift from universal acceptance of P to universal acceptance of Q as someone in Blue's position would gain from such a shift.) If P is optimific, and everyone has impartial reasons to prefer its universal acceptance to the universal acceptance of Q, this is most likely because the aggregate benefits to various people in P is accepted outweigh the costs to people in Blue's position. Perhaps Q would permit us to save Blue's life at the cost of failing to prevent a large number of people from being paralyzed, whereas P would require the opposite. Or perhaps P would require us to prevent many people from losing a leg rather than saving Blue's life, as Q would permit. In order to know which of these cases would fit the pattern I have described, one would have to know how Parfit's notions of impartial reasons and "best outcome" deal with aggregation. As I have said, this is not obvious. But presumably there will be some cases that fit the abstract pattern I have described.

These reflections have a bearing on Parfit's argument for the convergence of Rule Consequentialism and the two forms of Contractualism that he discusses. In this argument, he claims that everyone would have strong impartial reasons to choose that optimific principles be universally accepted, and that, because these reasons are not decisively outweighed by any conflicting reasons, everyone could rationally choose these principles. He then argues that, because there are no other significantly non-optimific principles that everyone could rationally choose, these optimific principles are the only ones whose universal acceptance everyone could rationally choose. When Parfit turns to my version of Contractualism, he then says that if certain optimific principles are the only ones whose universal acceptance everyone could rationally choose, this means that there are stronger objections to every other set of principles, and that if this is so then these optimific principles could not reasonably be rejected.

Suppose that optimific principles would require that we save many other people from smaller burdens rather than saving Blue's life. Though someone in Blue's position may have sufficient reasons to will the universal acceptance of these optimific principles, this person may also have sufficient reasons to will the acceptance of some non-optimific principle which would permit or require us to save Blue's life.

It might be that, taking only impartial reasons into account, everyone has stronger reason to will the acceptance of these optimific principles than to will the acceptance of some non-optimific principle that would require us to save Blue's life. This might also be put by saying that (considering only impartial reasons) there are "stronger objections" to this alternative than to the optimific principle. But taking *all* reasons into account, someone in Blue's position might have a stronger objection to the optimific principle that would impose such a sacrifice on Blue than anyone would have to some non-optimific principles that did not impose such a sacrifice. If this is correct, then the fact that these alternative principles are open to stronger (impartial) objections need not mean that they are open to decisive objections and hence need not entail that the optimific principles could not be reasonably rejected.

If what I have just said is correct, then shifting from the question "could anyone reasonably object to these principles being universally accepted" to the question "could everyone rationally will that they be universally accepted" produces a moral theory that requires us to make significantly greater sacrifices, and permits or requires others to impose such greater sacrifices on us.

This move would, however, also solve a difficulty that arises for a contractualist view like mine in cases of the first type.[17] If someone in Grey's position could reasonably reject a principle requiring him to make the organ donation, why would it not follow that someone in the position of the proposed recipient could reasonably reject a principle permitting Grey not to make the donation. After all, the personal reason that this person has for objecting to such a principle seems at least as strong as Grey's reason for rejecting the more demanding principle, and the cost to Grey is less. This would seem to lead to a moral standoff, in which there is no right answer to the question of what one should do. Shifting to the "what everyone could rationally will" (or concluding, with Parfit, that the reasonable rejection standard in fact collapses into this one) would solve this problem, albeit at a certain cost.[18]

Let me close by expression my agreement with a point that Parfit makes in his conclusion. Given its emphasis on impartial reasons and optimific principles, the Triple Theory that he proposes in his conclusion sounds (at least on first impression) more like consequentialism than my version of contractualism does. So one may question whether his Triple Theory is essentially a contractualist theory or a consequentialist one.

---

[17] Thomas Nagel raises this problem in *The View From Nowhere* (New York: Oxford University Press, 1986), pp. 50-51, 172.

[18] That is to say, it would solve the problem if in such situations there always is some principle that everyone could rationally will to be universally accepted (if the "uniqueness condition" is fulfilled.) This depends on the relative strength of impartial and self-interested reasons.

Parfit is correct, I believe, in saying that this theory is contractualist. Any plausible moral view makes what is right or wrong in many cases depend on the harms and benefits to individuals. A theory is consequentialist only if it takes the value of producing the best consequences to be the foundation of morality. Parfit's combined theory does not do this. According to that theory it matters whether the principles that would permit an action would be optimific. But this matters only because these are the principles that everyone has reason to will, and taking what can be justified to others—what they have reason to will—as the most fundamental moral idea is the essence of contractualism, at least as I have described it.

Recognizing the idea of justifiability to others as basic opens up a possibility that Parfit does not discuss, but which I think should not be neglected. Many people may be drawn to consequentialism because they see that there are some situations in which it the morally correct way to decide what to do is to figure out what would produce the best consequences overall. Decisions by public officials about what kind of hospitals to build may be a good example. Because producing the best consequences seems so obviously to be the right standard in these cases, people then infer that this idea is always morally basic. This seems to me to be a mistake: producing the best consequences might be the correct standard in these cases not because it is the basis of morality but because it is what is owed to people in situations of that kind, by agents who stand in a certain relation to them. Recognizing the contractualist idea of justification to others as morally basic allows us at least to raise the possibility that although what is owed to others in some situations is to follow the principles that would produce best consequences, impartially understood, this need not always be the case. In other cases our responsibilities and obligations may be different.

Of course it needs to be asked why this should be so, if it is so. And it might be responded that the cases in which it appears to be the case are in fact misleading: they are cases in which, because of the burdens of being impartial, *optimific* principles would permit people decide what to do on a basis other than what would be impartially best. But, as I said earlier in discussing Parfit's treatment of partiality toward one's friends and relatives, there are two was of describing such cases. Is partiality morally permitted because permitting it is impartially best? Or is it permitted because principles that demanded a higher level of impartiality would be ones that individuals could reasonably reject (for reasons that are not impartial)? The latter seems to me more plausible. In any event, this is a point where the residual tension between Rule Consequentialism and my version of contractualism seems to show itself.

37,333