# Should Kantians Be Consequentialists?

Jacob Ross

**Abstract.** Parfit argues that a form of rule consequentialism can be derived from the most plausible formulation of the fundamental principle of Kantian ethics. And so he concludes that Kantians should be consequentialists. I argue that we have good reason to reject two of the auxiliary premises that figure in Parfit's derivation of rule consequentialism from Kantianism.[1]

In chapter 16 of *On What Matters*, Parfit argues that the most plausible form of consequentialism, namely

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

follows from the most plausible form of Kantianism, namely

> **Kantian Contractualism** (KC): Everyone ought to follow the principles whose universal acceptance everyone could rationally will.

Parfit summarizes his argument as follows:

> (KC) Everyone ought to follow the principles whose universal acceptance everyone could rationally will.
>
> (C) There are some principles whose universal acceptance would make things go best.
>
> (F) Everyone could rationally will that everyone accepts these principles.
>
> (H) These are the only principles whose universal acceptance everyone could rationally will.

Therefore

> UARC: These are the principles that everyone ought to follow.

In what follows, I will argue that we should not accept either F or H, and hence that Parfit's argument from Kantian contractualism to rule consequentialism is not compelling in its current form. Parfit himself acknowledges that there are some cases to which the argument he has presented does not apply, and some complications that his argument fails to address. His stated aim is only to provide the main outlines of an argument whose details can be filled in later. Hence, in discussing the problem with F and H, I shall be indicating some of the difficulties that Parfit would face in trying to complete his argument.

But first, some preliminary matters. When we ask whether an agent could rationally will the universal acceptance of a principle, we must assume, first, that the agent has the power to determine what principles of action everyone will accept henceforth, and second, that she is fully informed. Because we assume full information, what an agent could rationally choose coincides with what an agent has sufficient reason to choose. Concerning the options an agent might choose, let us say that an option is *optimific* just in case no available alternative would have better consequences, as evaluated from an impartial point of view. Similarly, let us say that a *principle of action* is optimific just in case there is no alternative principle whose universal acceptance would have better consequences from an impartial point of view. And let us say that a principle is *non-optimific* to the extent that its universal acceptance would be worse, from an impartial point of view, than the universal acceptance of some alternative principle.

Let us now consider F. Clearly, this proposition will be accepted by anyone who accepts the following *strong consequentialist thesis*:

> SC: An option is rational whenever it is optimific.

For if we can rationally choose any optimific option, i.e., any option that would make things go best, then we can rationally choose the universal acceptance of any principle whose universal acceptance would make things go best, and so we can rationally choose the universal acceptance of any optimific principle. However, most people would reject SC, and Parfit himself explicitly rejects it. For most people hold that personal reasons, such as our reasons of prudence, and the reasons that derive from our close personal relations, should be given some weight in decision making. And so most people would accept the following *weak non-consequentialist thesis*:

> WN: For any agent, *s*, and any two options, *A* and *B*, if *A* and *B* are both optimific, and if *A* would be better than *B* from *s*'s personal point of view, then *s* could not rationally choose *B* over A.

If one accepts WN, then one should not accept F. For suppose there were two agents, *i* and *j*, with opposing interests, so that outcomes that are good from *i*'s personal point of view are bad from *j*'s point of view, and vice versa. Suppose there is a plurality of alternative optimific principles, and that there is no overlap between the optimific principles that are best from *i*'s personal point of view and those that are best from *j*'s personal point of view. In this case, if WN is true, then there will be no optimific principle whose universal acceptance everyone could rationally will, since, for every such principle, there will be some alternative optimific principle whose universal acceptance would be better from the personal point of view of either *i* or *j*. Therefore, if we accept WN, then we should not accept F.

One might object that although there may be no particular, first-order, optimific principle whose universal acceptance everyone could rationally will, there will still be a second-order optimific principle whose universal acceptance everyone could rationally will, namely *follow some optimific first-order principle or other*. But there is no guarantee that this second-order principle will itself be optimific. For from the fact that two first-order principles, *X* and *Y*, are each optimific, it does not follow that the universal acceptance of a principle permitting one to follow either *X* or *Y* would itself be optimific. For a situation in which some agents follow *X* while others follow *Y* might have disastrous consequences.[2]

Even apart from the possibility of alternative optimific principles, there is reason to reject F. We have seen that, if one holds that personal reasons should be given any weight whatsoever in decision making, then one should accept the weak non-consequentialist thesis. Further, if one holds that personal reasons have more than just tie-breaking significance, or in other words, if one holds that impersonal reasons do not have lexical priority over personal reasons, then one should accept the following *moderate non-consequentialist thesis*:

> MN: For any agent, *s*, and any two options, *A* and *B*, if *A* would be much better than *B* from *s*'s personal point of view, and if *A* would be only slightly worse than *B* from an impartial point of view, then s could not rationally choose *B* over *A*.

Now it might well be true that for every optimific principle, *O*, there exists an agent, *s*, and an alternative to *O*, $N_S$, such that the universal acceptance of $N_S$ would be almost as good as the universal acceptance of *O* from an impartial point of view, and much better than the universal acceptance of *O* from the personal point of view of *s*. (For a given agent, *s*, $N_S$ might be a principle which agrees with *O* except in cases where a violation of *O* would greatly benefit *s*, and where this violation would not cause other harms that significantly outweigh the benefits to *s* as evaluated from an impartial point of view.) And in this case, it will follow from WN that for every optimific principle, there is someone who could not rationally will its universal acceptance. Hence, contrary to F, there will be no optimific principle whose universal acceptance everyone could rationally will.

The problem just raised involves principles that are only slightly non-optimific. Parfit explicitly sets aside the consideration of principle of this kind, saying that such details can wait. So let us do likewise. And let us turn our attention to another claim that figures in Parfit's argument, namely

> (H) [Optimific] principles are the only principles whose universal acceptance everyone could rationally will.

Parfit argues for H as follows:

> If everyone accepted any [non-optimific] principle, that would make things go in ways that would be much worse in the impartial-reason-implying sense than some way in which things could have gone. In nearly all such cases, things would also go very badly for some unfortunate people. These people could not rationally choose that everyone accepts this non-optimific principle, since they would have both strong impartial reasons and strong personal reasons not to make this choice.[3]

Parfit's reference to "nearly all such cases" is ambiguous. He might mean

> (P1) Nearly all non-optimific principles are such that their universal acceptance would have very bad consequences for some people.

But P1 would not get Parfit what he needs. For in order to show that we should follow optimific principles, he must show that there are no permissible alternatives to following such principles; it does not suffice to show that there are *few* such alternatives. More plausibly, Parfit might instead mean

> (P2) For nearly every type of circumstance, *C*, every non-optimific principle that is applicable in *C* is such that its universal acceptance would have very bad consequences for some people.

On the basis of P2, Parfit might argue that, in nearly all circumstances, there are no applicable non-optimific principles whose universal acceptance everyone could rationally will. And thence he might argue that, in nearly all circumstances, one ought to follow optimific principles. And this would be a very interesting result, in spite of being less than fully universal. I will now argue, however, that P2 is false, as there many types of circumstance in which no one would be harmed by the universal acceptance of a non-optimific principle. And I will further argue that there are many types of circumstance in which everyone would significantly benefit from the universal acceptance of non-optimific principles, and hence in which it is plausible that everyone would have sufficient reason to will the universal acceptance of such principles. Hence, I will argue that we should not accept H.

As a first kind of case in which everyone could will the universal acceptance of non-optimific principles, consider cases in which impersonal goods are at stake. By an impersonal good, I mean a good that is valuable not because it good *for* anyone, or because it promotes anyone's interests, but rather because it improves an outcome in a manner that can be appreciated from an impartial point of view. There is some disagreement concerning which goods, if any, are genuinely impersonal in this sense. Some have argued that biodiversity is an impersonal good, since, though people may in fact benefit from it, its value does not depend on

anyone's benefiting from it. Others have argued that other goods, such as equality, desert, perfection, virtue, and respect for rights are impersonal goods in this sense. For the purpose of the present argument, what we must assume is that there is at least one impersonal good.[4] We must also assume that, from an impartial point of view, personal goods don't always have priority over impersonal goods. We must assume, in other words, that, from an impartial point of view, if a first outcome is better than a second outcome with respect to personal goods, the second outcome may still be better all things considered, so long as the second outcome is better by a sufficient margin with respect to impersonal goods.

Now consider the following principle:

> (PG) Act in such a way as to promote personal goods whenever doing so would be best from everyone's personal point of view, even when, from an impartial point of view, the benefits of so acting are outweighed by harms in relation to impersonal goods.

It is very unlikely that PG is an optimific principle. And yet it may well be that no one would be harmed by the universal acceptance of PG. Indeed, it may well be that everyone would be better off given the universal acceptance of PG than given the universal acceptance of any optimific principle. Hence, assuming that it is rationally permissible to give significant weight to personal reasons, it may well be that everyone could rationally choose the universal acceptance of PG over the universal acceptance of any optimific principle.

As a second kind of case in which everyone might rationally will the universal acceptance of a non-optimific principle, consider cases in which we must choose between the interests of those who will live in the nearer future and those who will live in the more distant future. Generally, our personal reasons for caring about other individuals are based on our personal ties with these individuals. These ties can be direct or indirect: we have direct personal ties with our friends and with our children, while we have indirect personal ties with our friend's friends and with our great great grandchildren. It is plausible that, other things being equal, the more indirect our personal ties are with an individual, the weaker is our *personal* reason for concern about this individual's welfare. And as we consider consecutive future generations, our strongest personal ties with individuals in these generations become more and more indirect, and so our strongest personal reasons for concern become correspondingly weaker. Hence, everyone may have personal reason to will the universal acceptance of principles that would favor people who will live in the nearer future over people who will live in the further future.

Consider, for example, the following *future discount principle*:

FD: In deciding the rate at which to consume resources, give significantly greater weight to the welfare of those who will live in the nearer future over the welfare of those who will live in the more distant future, discounting the future relative to the present at a constant discount rate.[5]

It is unlikely that a principle of this kind would be optimific. And yet it may well be that each individual (past, present, and future) has strong personal reason to will that FD be accepted by all her contemporaries, and by everyone who lives after her. Hence, assuming that it is rationally permissible to give significant weight to personal reasons, it may well be that everyone could rationally choose the universal acceptance of FD over the universal acceptance of any optimific principle.

One might object that we could not rationally will that *everyone* accept FD, since no individual could will that those who lived long before her accept FD. But this is not a relevant objection. For, as Parfit states, "when we apply the Kantian Formula, we ask which principles each person could rationally choose, if this person supposed that she had the power to choose which principles would be accepted by everyone, *both now and throughout the future*"—we do not ask what this person could rationally choose if she had the power to rewrite the past.[6] And there is good reason for Parfit to restrict the relevant choice in this manner. For if all those who lived before us accepted a certain principle of action, one likely result is that we would not exist: the past would have gone very differently, and hence we would never have been born. Thus, we may be unable to rationally will that everyone accepted this principle in the past, on the grounds that we may have overriding personal reason not to will our own non-existence. But this consideration is not relevant to whether we should now act on the principle in question. And so, to exclude this consideration, we must restrict the relevant choice to the present and future. Moreover, we will now see that further problems arise for H regardless of whether the relevant choice is restricted to the present and future or concerns the past as well.

As a final kind of case in which everyone might rationally will the universal acceptance of a non-optimific principle, consider cases in which who shall exist in the distant future will depend on how people act in the interim. Parfit considers cases of this kind in *Reasons and Persons*, under the heading *the non-identity problem*, and I shall now consider a variant of one of Parfit's examples.

Suppose we are choosing among possible energy sources. For the sake of simplicity, let us suppose there are only two alternative principles of action:

RS: Derive energy from renewable sources.

FF: So long as there are fossil fuels, burn them; switch to renewable sources of

energy when the fossil fuels run out.

Suppose that, in the long run, the universal acceptance of RS would be much better than the universal acceptance of FF, since the latter would lead to severe global warming and other negative environmental consequences. Suppose, therefore, that RS, and not FF, is the optimific principle. Suppose, however, that for the next two hundred years, everyone will be considerably better off given the universal acceptance of FF, since this would lead to much more energy being available at much lower cost, and hence to a much stronger global economy. Suppose, further, that the choice between RS and FF will have wide ranging consequences, so that the course of human history will vary greatly depending on whether we accept RS or FF. Suppose, in particular, that which policy we choose will affect who shall exist more than two hundred years from now, so that there is no one who will exist more than two hundred years from both on the condition that everyone accepts RS and on the condition that everyone accepts FF. Suppose, finally, that if everyone were to accept FF, then those who will live more than two hundred years from now would have lives that are well worth living, even though they would not be as good as the lives of those who would live more than two hundred years from now if everyone were to accept RS.

Given these assumptions, it could well be that everyone who will live during the next two hundred years will have strong personal reason to prefer the universal acceptance (past, present and future) of FF. And it could also be that everyone living more than two hundred years in the future will have strong personal reason to will the universal acceptance (past, present and future) of FF. For if, as a matter of fact, those who live for the next two hundred years accept FF, then those who will live more than two hundred years in the future will owe their existence to the fact that their predecessors accepted FF rather than RS. Hence, those living in the distant future will have strong personal reason to will that everyone, including everyone who lived before them, accept FF rather than RS—since if everyone had accepted RS, they themselves would never have existed, and since everyone whose life is well worth living has strong personal reason to prefer her existence to her non-existence. Hence, assuming that it is rationally permissible to give significant weight to personal reasons, it may well be that everyone who ever lives could rationally will the universal acceptance, at all times, of the non-optimific principle FF.

I conclude that there are a number of apparent counterexamples to two of the claims that underlie Parfit's argument. If Parfit is able to resolve the difficulties that these examples appear to present, his argument for the conclusion that Kantian contractualism supports rule consequentialism will be considerably more forceful.

*School of Philosophy*
*University of Southern California*
*3709 Trousdale Parkway*
*Los Angeles, CA 90089-0451*
*jacobmro@usc.edu*

1. I am indebted to Larry Temkin and Mark Schroeder for very helpful comments on earlier drafts of this paper. I am especially grateful to Derek Parfit, who provided invaluable comments on multiple drafts of this paper, which greatly improved the argument.

2. Parfit acknowledges that complications are raised by the possibility of a plurality of alternative optimific principle, but these, he says, are questions of detail that can be answered later. In my view, this understates the difficulties presented by the possibility of alternative optimific principles. The problem this possibility raises is not, as Parfit suggests, that there might be a plurality of alternative principles, each of which is such that everyone could rationally will its universal acceptance. Rather, this possibility raises the problem that there might be no optimific principle whose universal acceptance everyone could rationally will, since, for every such principle, there may be some agent who has more reason to will the universal acceptance of some alternative principle.

3. See *On What Matters*, chapter 16.

4. See, for example, Larry Temkin, "Harmful Goods, Harmless Bads," in *Value, Welfare, and Morality*, eds. Frey, R. G. and Morris, Christopher, pp. 290-324, Cambridge University Press, 1993.

5. For a discussion of some closely related principles, see George Ainslie, *Breakdown of Will*. Cambridge: Cambridge University Press, 2001.

6. *Op. Cit.*