Ratio (new series) XXII 1 March 2009 0034–0006

CLIMB EVERY MOUNTAIN?

Michael Ridge

Abstract
The central thesis of Derek Parfit's *On What Matters* is that three of the most important secular moral traditions – Kantianism, contractualism, and consequentialism – all actually converge in a way onto the same view. It is in this sense that he suggests that we may all be 'climbing the same mountain, but from different sides'. In this paper, I argue that Parfit's argument that we are all metaphorically climbing the same mountain is unsound. One reason his argument does not work is that he has misunderstood the way in which a plausible rule-consequentialism should understand the supervenience of rightness on all possible acceptance levels of the ideal moral code. In place of Parfit's own understanding of this, I develop a view I call 'variable-rate rule-utilitarianism', which I argue shares the key insight of Parfit's view but avoids a fatal objection to his own articulation of that insight. Finally, I explore how this modification might allow us to still make a case that we are all 'climbing the same mountain', albeit in a very different way and for very different reasons than the ones Parfit had in mind.

1. Introduction

Derek Parfit ended his classic *Reasons and Persons* on a hopeful note, suggesting that,

> Disbelief in God, openly admitted by a majority, is a recent event, not yet completed. Because this event is so recent, Non-Religious Ethics is at a very early stage. We cannot yet predict whether, as in Mathematics, we will all reach agreement. Since we cannot know how Ethics will develop, it is not irrational to have high hopes.[1]

In his long-awaited *On What Matters*, Parfit in effect goes a long way toward vindicating these high hopes. For in this impressive book,

[1] Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), p. 454.

he argues that Kantian, contractualist, and consequentialist traditions are in some sense all ultimately converging on the same view. Metaphorically, they are all climbing the same mountain, but from different sides. If this claim could be adequately defended, then it would indeed go a long way toward vindicating the high hopes mentioned at the end of *Reasons and Persons*. For these three traditions (Kantian, contractualist, and consequentialist) would surely be on any reasonable person's short list of the most promising moral theories yet developed.

In this paper, I argue that Parfit has not adequately established his conclusion. In particular, I argue that the most plausible version of rule-consequentialism does not, so far as Parfit's arguments go, anyway, converge with the deliverances of the best versions of Kantianism and contractualism. However, there may still be a way to vindicate Parfit's ambitious convergence thesis, albeit with a very different argument from the one he offers. I gesture in the direction of such an argument at the end of this paper. Since I am not sure the argument can be developed in a way that is ultimately convincing, though, I shall not attempt that task here. My positive conclusion is therefore somewhat more tentative and suggestive than my negative conclusion about Parfit's own argument.

2. The Ideal World Objection

Parfit begins with a lengthy discussion of Kantian ethics, the details of which are not germane here. The conclusion of that argument, which I am willing to grant for the sake of argument, is that the most plausible development of Kantian ethics leads to a form of contractualism. At a first pass, Kantian contractualism is the following doctrine:

> KC 1: Everyone ought to follow those principles whose universal acceptance everyone could rationally will.

Parfit argues that what I am calling KC 1 falls prey to the 'Ideal World' objection. The basic idea behind this objection is not a new one. The basic idea is that Kantian moral principles are designed for an ideal world of perfect moral virtue, but that to follow such principles in our very non-ideal world could, and

often would, have horrible consequences. To follow such principles in the ideal world would, the objection goes, be to cause great harm for no good reason.

Parfit's development of this objection is, however, original. He points out that the usual form of the ideal world objection is that Kantian ethics requires us to act in ways which would have horrible outcomes. For example, one might hold that because universal non-violence would be ideal, Kantian ethics requires pacifism. Since pacifism in the face of great evil (just think of World War II) could have horrible consequences, this would be enough for a reductio of the Kantian view.

However, Parfit points out that this objection is too quick. For Kantian ethics as he understands it to require pacifism, there must be no better maxim than the pacifist one which requires non-pacifism in some circumstances. Parfit suggests that there is a better maxim, namely one which requires us to never use violence, except when others have used aggressive violence, in which case we may use restrained violence insofar as this is necessary to protect ourselves or others.

Parfit argues that the best version of the ideal world objection maintains not that Kantian ethics requires too much, but that it permits too much. In this context, he asks the reader to consider the following maxim:

> Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can. (*On What Matters*, §38)[2]

Parfit's point is that this maxim could rationally be willed by everyone. For if everyone follows the maxim then nobody will ever engage in acts of aggressive violence, in which case the outcome is just the same as with the universal adoption of a strictly pacifist maxim. Since this maxim can rationally be willed as universal by everyone, it seems that the Kantian contractualist theory as Parfit first develops it would permit acting in accordance with this maxim. This, however, is absurd.

Parfit makes essentially the same point with a maxim which has the agent keep his promises and help those in need, unless some

[2] All references to *On What Matters* are to the numbered sections of the November 2007 draft of Parfit's manuscript. The original title of this manuscript was *Climbing the Mountain*.

people have not acted in this way, in which case the agent is to copy them. Once again, universal compliance of this maxim would be for the best, but following it in the real world would be absurd. Because of its focus on the ideal world of universal acceptance and compliance, the Kantian theory has absurd consequences in the real world. Call this the Ideal World Objection.

Parfit argues that there is a plausible way of revising the Kantian theory to avoid the Ideal World Objection. Instead of defining right action in terms of acceptance of a moral code by everyone, the Kantian should define right action in terms of acceptance of a moral code by everyone *and* by any other number of people as well. In particular, the Kantian should on Parfit's view define right action as follows:

> LN4: It is wrong for us to act on some maxim unless we could rationally will it to be the case that this maxim be acted on by everyone, *and any other number of people*, rather than by no one. (*On What Matters*, §38)

It should be easy enough to see why this would deal with the Ideal World Objection. For consider Parfit's examples. The maxim of not using violence unless others have, in which case, kill as many as I can, would not pass the new test. For while that maxim does very well if everyone accepts it, it does very poorly if not everyone accepts it, and for obvious reasons. A similar point applies to Parfit's maxim of keeping promises and helping others unless some others have not done so.

Parfit argues that rule-consequentialism is subject to a similar objection. For rules which do very well if everyone accepts them may be such that following them in the real world of non-universal acceptance would have horrible consequences. Here again, Parfit argues that the solution is to redefine rule-consequentialism so that it is defined not in terms of universal acceptance, but in terms of universal acceptance *and* any level of acceptance short of universal acceptance apart from no acceptance whatsoever. So on his view, the most plausible version of rule-consequentialism will be roughly the following:

> RC2: Everyone ought to follow those rules whose being followed by any number of people rather than by no one would make things go best. (*On What Matters*, §38)

One reason this is only a rough statement is that rule-consequentialist might well do better to define right action in

terms of *acceptance* of the ideal code rather than in terms of *following* that code. Indeed, Parfit himself goes on to consider such a version of rule-consequentialism later in *On What Matters*. This, however, is orthogonal to the issue under consideration here. For whether we focus on acceptance or compliance, Parfit's view is that in order to avoid the Ideal World Objection, we must revise these theories so that they are defined in terms of not only universal compliance/acceptance, but also in terms of compliance/acceptance by any other number of people rather than by no one. With these crucial revisions to Kantianism and rule-consequentialism in place, Parfit has set the stage for his argument that Kantianism and rule-consequentialism converge.

3. Climbing the Mountain: Parfit's Master Argument

Parfit argues in the concluding chapter of his book that there is only one set of principles everyone could rationally will to be universal laws, and that those principles are also the only principles which nobody could reasonably reject. I shall not rehearse this argument here, but if it is sound then it shows that T.M. Scanlon's contractualism, which holds that an action is wrong if it would be forbidden by principles nobody could reasonably reject, converges with what Parfit takes to be the most plausible form of Kantianism. Since Scanlon's contractualism is perhaps the most plausible version of contractualism, this would be enough to show a convergence between two of the three traditions that Parfit wants to argue converge. The crucial further dialectical burden is to show that Kantianism and contractualism both converge on the most plausible form of consequentialism. For these purposes, Parfit takes the most plausible form of consequentialism to be a form of rule-consequentialism. He argues as follows:

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

Therefore

(F) Everyone would have sufficient reasons to choose that everyone accepts these UA optimific principles.

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

Therefore

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

Therefore

These are the principles that everyone ought to follow. (*On What Matters*, §49)

There is something peculiar about this argument in the context of the rest of Parfit's book. For here Parfit's (A) is simply the version of Kantian contractualism which he earlier argued was vulnerable to the Ideal World Objection. This is peculiar because Parfit argued that this objection could be avoided by making a friendly amendment to Kantian contractualism – by moving from 'universal' to 'by everyone and by any other number of people, rather than by no one'. Furthermore, Parfit argued that the same sort of problem plagued rule-consequentialism, and that the same sort of fix was available. Yet the preceding argument purports to establish a version of rule-consequentialism couched in terms of universal acceptance, rather than in terms of acceptance 'by everyone, and by any other number of people, rather than by no one'.

Perhaps Parfit's dialectical aim is much more modest than it seems. Perhaps he intends to show only that a defective version of rule-consequentialism can be derived from a defective version of Kantian contractualism. This would not be entirely without interest, but it would also be much less interesting than the conclusion that the most plausible form of rule-consequentialism could be derived from the most plausible version of Kantian contractualism.

Furthermore, this more ambitious derivation seems essential to what Parfit takes to be one of the most important metaethical implications of *On What Matters*. For at the end of the concluding chapter of that book (*On What Matters*, §55), he suggests that arguments for anti-realism in meta-ethics often depend on the assumption of rather deep disagreement. If, however, the three major traditions he discusses are all really 'climbing the same

mountain,' the extent of genuinely deep disagreement is very limited indeed. To that extent, Parfit suggests, his argument casts doubt on anti-realist views in meta-ethics – or, rather, on at least one important argument for such views. This coda to *On What Matters* is of course not entirely unrelated to the coda of *Reasons and Persons*, in which Parfit suggested that it is early days for secular ethics, and the extent of deep disagreement between reasonable persons may be easily overstated.

However, this line of thought is much less convincing if all Parfit has shown is that a defective version of each of these views entails a defective version of the other. For recall that the views which figure in Parfit's official derivation are, by his own lights, open to a fatal objection – the Ideal World Objection. So the anti-realist would be within his rights to reply that for all that has so far been said, the most plausible and interesting exemplars of each of the major traditions Parfit considers actually are inconsistent, and that might be enough to mount a powerful anti-realist argument. This suggests that dialectically Parfit needs to see whether his derivation would go through from what by his own lights is the most plausible version of Kantian contractualism to what is by his own lights the most plausible version of rule-consequentialism. That, in turn, would mean that we would need to define these views not in terms of 'universal acceptance', but instead in terms of 'acceptance by everyone, and by every other number of people, rather than by nobody'. Indeed, in an earlier draft of the book, Parfit himself included a footnote to what I am calling his master argument in which he said, '"universal acceptance" here could mean "acceptance by everyone and by every other number of people" '.[3]

In any event, regardless of Parfit's own dialectical aims, it seems to me a very interesting question whether his derivation would go through if we substituted 'acceptance by everyone and by every other number of people, rather than by nobody' for 'universal acceptance' throughout. For we would then see whether what might well be the most plausible version of Kantian contractualism really did entail what might well be the most plausible form of rule-consequentialism – forms of each doctrine which are tailor made to avoid the Ideal World Objection. In the remainder of this section, I am going to argue that this derivation is, unfortunately

[3] See, for instance, the endnote 359 of the March 2006 draft of the manuscript.

unsound. In which case, we do not as yet have any reason to think *the most plausible version of Kantianism* and *the most plausible version of rule-consequentialism* converge.

Here is what a suitably modified version of Parfit's derivation would look like:

(A*) Everyone ought to follow the principles whose acceptance by everyone and by every other number of people, rather than by no one, everyone could rationally will, or choose.
(B*) Anyone could rationally choose whatever they would have sufficient reasons to choose.
(C*) There are some principles whose acceptance by everyone, and by every other number of people, rather than by no one, would make things go best.
(D*) These are the principles whose acceptance by everyone, and by every other number of people, rather than by no one, everyone would have the strongest impartial reasons to choose.
(E*) No one's impartial reasons would be decisively outweighed by any set of relevant conflicting reasons.
Therefore
(F*) Everyone would have sufficient reasons to choose that everyone accepts these principles.
(G*) There are no other significantly non-optimific principles whose acceptance by everyone and by every other number of people, rather than by no one, everyone would have sufficient reasons to choose.
Therefore
(H*) It is only these principles whose acceptance by everyone, and by every other number of people, rather than by no one, that everyone would have sufficient reasons to choose, and could rationally choose.
Therefore
These are the principles that everyone ought to follow.

Having motivated this revision of Parfit's master argument, in the next section I critically evaluate the argument.

4. Multiple Moral Codes and Nihilism for the Wrong Reasons

The first important contrast between this argument and Parfit's derivation is the difference between (C) and (C*). Parfit's (C) is

arguably a platitude, putting complications due to ties for equal best to one side.[4] Putting ties to one side is fair enough in this context, as any version of rule-consequentialism will have to have some way of dealing with ties between different codes.

(C*), however, is very far from a platitude. The most natural reading of (C*) is as claiming that there is some moral code M, such that for any non-zero acceptance level n, acceptance n of M would make things go at least as well as acceptance n of any other moral code. For it could well be that there simply is no code M which is best at every level of acceptance. It could instead by that one code is best at 100% acceptance, while a very different code is best for 90% acceptance, and yet another code is best at 65% acceptance, and so on.

Presumably it is an empirical question whether there is a single code M which satisfies the criterion laid down by (C*). Ex ante, this seems very unlikely, for to avoid arbitrariness we should consider every possible acceptance level, and given that the global population is now in the billions, there will be a truly enormous number of possible acceptance levels. The idea that there would be one code which is best for each and every one of these would seem prima facie to be extremely unlikely. In which case, (C*) is most likely false, which in turn means that the revised derivation I have proposed on Parfit's behalf is most likely unsound.

This objection to Parfit's argument, as I have revised it, relies on the possibility of multiple moral codes, each of which is best at some level of acceptance, but none of which is best at every level of acceptance. I therefore call it the 'Multiple Moral Codes Objection'.

In fact, though, this objection suggests a closely related one. For we should now return to Parfit's proposed version of rule-consequentialism. Parfit thinks that in order to avoid the Ideal World Objection, the consequentialist should hold that everyone ought to follow those rules whose being accepted by any number of people rather than by no one would make things go best. Actually, his initial gloss is in terms of following the ideal code, whereas I have put the point in terms of accepting it, but nothing

[4] We must also put to one side complications raised by the possibility that there are infinitely many possible codes. Since this is also a problem for all forms of rule-consequentialism, I shall ignore this complication throughout the rest of this paper. Thanks to Campbell Brown for flagging this issue.

here shall hang on this, and Parfit himself considers both versions as plausible candidates as the best version of rule-consequentialism.

The problem is that this version of rule-consequentialism entails that if there is no single code which is best for *each and every single level of acceptance* then nothing is morally required. We have already seen that it is very likely the case that there is no such code. In which case, Parfit's rule-consequentialism very quickly entails nihilism is true in the actual world.

This is already somewhat objectionable, but there is a deeper objection. Regardless of whether there is no single code which is best at each and every level of acceptance, this will in any event be a contingent empirical matter. There at the very least *could* be a world with moral agents in which there is no such code. Parfit's rule-consequentialism entails that in such a world, nothing is morally required. This, however, seems to make moral nihilism follow from the wrong sorts of reasons. There are various interesting meta-ethical arguments for nihilism, and I do not mean to suggest that we can simply ignore the possibility that nihilism might be true. I do not even want to presume that it might not be the case that nihilism is true in some worlds but not others, though this is already rather odd. Instead, my point is simply that nihilism should not follow from the somewhat eccentric fact that there is no single code which is best at each and every level of acceptance. Call this the 'Nihilism for the Wrong Reasons' objection.

Parfit might try to reply to this line of objection by invoking so-called 'conditional rules'. Conditional rules of the relevant sort have in their antecedent some claim about acceptance levels of the code itself. Parfit could argue that whenever two codes diverge between different levels, there will be a better code which takes both of these codes into account, but in a conditional way. For example, a code with an unconditional rule R might do well at n acceptance level, but not at n+m acceptance level. Not to worry, Parfit, might reply – just move to a code with a rule R* which has as its antecedent that there is n acceptance of the code and the original unconditional R as its consequent. The idea would be that the resulting code would have all the advantages of the original code without its disadvantages – and this would be because the new code is more sensitive to differences in acceptance levels and the implications thereof than the original.

There are at least four problems with this reply. First, it assumes that the consequences of the acceptance of a code are limited to the direct consequences which follow from people following the rules. Without this assumption, it is hard to see why we should think there is any sort of dominance argument for the code with conditional rules over the original unconditional ones. However, we know from existing discussions of rule-consequentialism and the famous 'collapse objection', according to which rule-consequentialism simply collapses into act-consequentialism, that this is not true. Accepting a code can have a variety of indirect consequences, and the reply on offer says nothing about this.

Second, the inclusion of a multiplicity of conditional rules could be costly in a variety of ways. The costs in terms of remembering and being suitably disposed to follow those rules, even in the face of temptation might well be high. These costs might outweigh the benefits of the extra nuance and discrimination between contexts that the more subtle rules would introduce.

Third, there might well be more special pleading in the actual real world application of the principles. Moral principles are always open to rationalization and special pleading in their application, but arguably conditional rules are especially conducive to such special pleading if the antecedents of the principles in question are hard to determine or unclear. Working out the general acceptance level of a moral code might well be very difficult, and this would introduce new scope for special pleading – unconsciously applying the moral principles in a way that is skewed to one's own interests or concerns. A better code might simply drop the nuance and include more unconditional rules to avoid such special pleading and rationalization.

Fourth, and finally, the main point remains that on any plausible view it will be an empirical and contingent matter whether the inclusion of conditional rules can ensure a single code is ideal for all acceptance levels. All the objection on offer requires is the mere possibility of a world in which this is not the case.

So far in this section I have offered an objection both to Parfit's master argument (as I have revised it) for convergence and an objection to his version of rule-consequentialism. It would seem that Parfit's optimism was misplaced – we are not all climbing the same mountain.

However, this conclusion itself may be premature. In the final section of this paper I shall explore another way in which Parfit's idea that we are all climbing the same mountain might be vindi-

cated. This argument is speculative, and I shall not here try to defend its soundness, though I do think it is interesting enough to be worthy of consideration. Before turning to this alternative argument that we are all climbing the same mountain, though, I must first establish that there is another way for the rule-consequentialist to avoid the Ideal World Objection. For if the Nihilism for the Wrong Reasons Objection is sound, Parfit's solution simply moves the bump in the rug, exchanging one fatal objection for another. I have independently argued for an alternative version of rule-consequentialism which avoids both the Ideal World Objection and the Nihlism for the Wrong Reasons Objection – variable-rate rule-utilitarianism. In the following section I briefly sketch this view and some of its main virtues, and in the final section I explain how we might try to find a convergence between this version of rule-consequentialism and a version of Kantian contractualism – albeit not the version Parfit considers to be the most plausible.

5. Variable-Rate Rule-Utilitarianism[5]

It seems that rule utilitarians face a dilemma. Either they characterize general acceptance as 100% acceptance or characterize it as something less than 100% acceptance. On the first horn of the dilemma the rule-utilitarian is open to the Ideal World Objection, while on the second horn of the dilemma the theory is open to the charge of arbitrariness and a lack of philosophical depth. Rule-utilitarians like Brad Hooker take the second horn of the dilemma. Hooker defines general acceptance as 90% acceptance. As I have argued at greater length elsewhere, picking a specific level of acceptance like 90% is not plausible. I briefly review these objections here.

First, any specific level of acceptance will inevitably be somewhat arbitrary. Hooker makes an interesting case that the selection of 90% is not *entirely* arbitrary, and here he may succeed. For 90% does seem more plausible than, say 51%. However, fixing the level at precisely 90%, as opposed to 95% or 85%, say, still seems rather arbitrary. It is hard to believe that our most fundamental moral principle could be arbitrary in this way.

[5] The following section draws heavily on my 'Introducing Variable-Rate Rule-Utilitarianism,' *Philosophical Quarterly*, 56 (2006), pp. 242–253.

Second, this sort of view faces a utopianism objection very similar to the one facing full acceptance versions of the theory. The problem with theories framed in terms of 100% acceptance is that they were simply not designed to deal with situations in which there is less than 100% acceptance of the ideal rules and, after all, in the real world we may well find ourselves in such circumstances. An isomorphic objection can be pressed against a view like Hooker's. For a theory framed in terms of (e.g.) 90% acceptance is simply not designed to deal with situations in which there is less than 90% acceptance of the ideal rules and, after all, in the real world we may find ourselves in just such circumstances.

Third, such a theory lacks explanatory depth. Rule-utilitarianism traditionally aspires to provide the ultimate principle of morality; it should be an (the) axiom and not a theorem. However, it seems clear on reflection that 90% seems plausible to us insofar as it does because 90% seems like a realistic ideal *for us*. By contrast, 99% acceptance seems like a very unrealistic ideal for creatures like us and suitable instead for angels or Vulcans (like Mr. Spock from *Star Trek*). At the other extreme, levels of social acceptance much lower than 90% are realistic enough for creatures like us, all right, but is not sufficiently ambitious. We can do better than *that*. The upshot of all this is that what level of social acceptance is germane for a given group of people will vary from one group to the other in a systematic way depending on the psychologies of the members of the groups in question. There ought to be some principled explanation of this striking covariation. However, if there is a deeper moral principle which provides a function from facts about the psychologies of a group of creatures (and perhaps other related facts) to a level of social acceptance which is suitable for a moral principle governing such creatures then a rule-utilitarian theory couched in terms of the level of acceptance which just happens to be suitable for human beings (90% e.g.) will turn out to be a mere theorem. The ultimate moral principle will not be this locally correct (we are now assuming) principle but instead a deeper principle which provides a function from psychological facts about the group under consideration and perhaps other facts to a version of rule-utilitarianism specified in terms of a particular level of social acceptance.

The preceding dilemma for rule-utilitarianism depended crucially on the rule-utilitarian's apparent need to specify a particular level of social acceptance for the ideal code. Perhaps we should

re-examine the assumption that rule-utilitarians really are committed to providing a specific level of acceptance. A plausible form of rule-utilitarianism does need to avoid the charge of utopianism, but the simplest solution to this problem is not to move from a theory couched in terms of 100% acceptance to a theory couched in terms of some specific level of acceptance which is less than 100%. For we can instead reject the more basic idea that rule-utilitarianism needs to be formulated in terms of any specific level of social acceptance without making our theory hopelessly vague and indeterminate. Perhaps, in other words, we should reject what we might call the 'fixed-rate' interpretation of rule-utilitarianism which insists on privileging some specific level of acceptance for a given society. Instead of privileging one specific level of social acceptance we could in effect include all possible levels of social acceptance in our account of right action.

In particular, we could hold that an action is right just in case it would be required by rules which have the following property: when you take the expected utility of every level of social acceptance between (and including) 0% and 100% for those rules and compute the *average* expected utility for all of those different levels of acceptance, the average for these rules is at least as high as the corresponding average for any alternative set of rules. Call this account of right action 'variable-rate rule-utilitarianism'. The variable-rate approach has a number of important advantages over more traditional 'fixed-rate' approaches. Since I have discussed these advantages, as well as the replies to some of the more obvious objections to the view elsewhere, though, I shall not go through these points again here. Suffice it to say that variable-rate rule-utilitarianism can avoid the Ideal World Objection as well as the three objections just discussed against a fixed-rate view like Hooker's which fixes the rate at something less than 100%.

Here instead, I want to compare and contrast variable-rate rule-utilitarianism with Parfit's preferred version of rule-consequentialism. The views are very similar in that, unlike fixed-rate views like Hooker's, both variable-rate rule-utilitarianism and Parfit's rule-consequentialism define moral rightness in such a way that it supervenes on the consequences of the ideal code for all possible non-zero levels of acceptance. It is this similarity that explains how each of them can avoid the Ideal World Objection.

However, my own view defines rightness in terms of the code with the highest *average* score across all acceptance levels. By contrast, Parfit's rule-consequentialism defines rightness in terms

of the single code which is best across all possible acceptance levels. We have seen (in section three) that this latter feature means that Parfit's rule-consequentialism is vulnerable to what I have called the Nihilism for the Wrong Reasons Objection. This is because there may well be no single code which is best for each and every possible acceptance level. My diagnosis is that Parfit was right to define rightness in such a way that it supervened on all of the different acceptance levels, but he was wrong in the specific way in which he suggested it supervened thereon. My own view avoids the Nihilism for the Wrong Reasons Objection because there will always be at least one moral code which is such that the average of its utility across all acceptance levels is at least as high as any other code. There will still be the possibilities of ties, but of course that is an issue all forms of rule-consequentialism must confront, and my only aim here is to show that my own version of rule-consequentialism is the most plausible version of the view – I am not arguing that any such view is correct. Variable-rate rule-utilitarianism seems to be more plausible than Parfit's version of the view. By going with the average across all acceptance levels, instead of holding out for a code which is ideal at every level, we do not give hostages to nihilistic fortune.

Having seen how we can avoid the Ideal World Objection without falling prey to the Nihilism for the Wrong Reason Objection, I now want to return (in section five) to the question of whether we might all be 'climbing the same mountain' after all.

6. Climb Every Mountain?

The arguments I have so far presented suggest that we are not all metaphorically climbing the same mountain that Parfit suggests we are. For that mountain's peak is a form of rule-consequentialism which by Parfit's own lights is defective – it is open to the Ideal World Objection. Furthermore, the base of that mountain is also unstable, as the base of that mountain is a version of Kantianism which is itself vulnerable to the Ideal World Objection. It is somewhat unsatisfying that Parfit's master argument is couched in terms of views which he himself has so powerfully refuted.

This would not be a serious objection to Parfit's main point if we could simply substitute Parfit's more considered views and still

have a sound argument. Unfortunately, this does not work. We cannot all be plausibly taken to be (again, metaphorically) climbing a mountain whose peak is Parfit's own modified version of rule-consequentialism. For the derivation of that view from a suitably modified Kantianism itself rests on what I have called (C*), and (C*) is itself very implausible for reasons discussed above. Not only is the route from the base of this mountain to its peak unsafe (that is, the derivation is unsound), the peak itself is problematic. For the peak of this mountain in Parfit's own version of rule-consequentialism, which I have argued is itself vulnerable to the Nihlism for the Wrong Reasons Objection.

Perhaps, though, there is a third mountain we are all metaphorically climbing? Here I shall conclude very tentatively, for I am not entirely convinced that the idea I am about to explore can really be made to work. Since it still seems to me by far the best chance for Parfit's intriguing mountaineering metaphor to be redeemed, though, I at least wanted to put it on the table for consideration.

The main dialectical task is to find a path from the most plausible form of Kantian contractualism to variable-rate rule-utilitarianism. Suppose instead of Parfit's preferred Scanlonian contractualism, we take a more Rawlsian contractualism as the base of our mountain. Such a view would include a Rawlsian veil of ignorance. Unlike Rawls's own view, the construction would aim at a comprehensive moral theory, and not just a political view.[6]

For our purposes, though, we must depart from the Rawlsian construction in another important way. For Rawls very explicitly was developing a view in what he calls 'ideal theory', and in ideal theory we can simply assume universal compliance with the principles under consideration. Here, by contrast, we are interested precisely in the issues raised in non-ideal theory by less than full compliance/acceptance of the relevant principles.

Perhaps, though, there is a principled way of dealing with this within the Rawlsian framework. Perhaps the most sensible way to modify the Original Position in light of a wide range of possible acceptance levels would be to include in the veil of ignorance that the parties do not know what level of acceptance they will find

[6] For the relevant contrast, see John Rawls, *Political Liberalism* (Columbia: Columbia University Press, 1993).

when the veil is lifted, though they can assume it is some non-zero level. The motivation for including this in the veil would *not* be the same as the motivation for including race, sex, talents, skills, wealth, etc. – it will not be that the level of acceptance is in Rawls's sense 'morally arbitrary'. Indeed, such levels clearly will be morally relevant, which is why some of Parfit's so-called 'conditional rules' may well form part of an ideal moral code even on a variable-rate view. Instead, the motivation for including this in our modified Original Position will be *versatility* of the resulting code – we want to construct a code which will be suitable for a wide range of contexts and hence a wide range of acceptance levels.

This is already controversial, of course, but in order to derive variable-rate rule-utilitarianism from this modified Original Position, we need another controversial premise. For we must reject Rawls's own view that the correct principle of choice in the Original Position is a maximin principle, but instead a classical principle requiring the agent to maximize her expected utility. Of course, while this is controversial, it is also something many of Rawls's critics have argued is more plausible than Rawls's own highly conservative maximin premise, the special circumstances of the Original Position and Rawls's arguments notwithstanding.

Finally, we need the closely related assumption that it is rational when behind this veil of ignorance to assume that it is equally likely that one will be any of the subjects in the society for whom the theory is being constructed. Without this assumption, or some other way of assigning probabilities, we would not be able to deploy the expected utility principle as a criterion of choice. This is also a controversial assumption, but it is also one that many of Rawls's critics have defended against his own view that it is rational not to assign any probabilities at all when behind such a veil.

These special assumptions in place, though, it does seem that a plausible derivation a special form of variable-rate rule-utilitarianism is not hard to find. Putting the issue of different acceptance levels to one side, the maximizing assumptions discussed above are generally taken to entail some form of average utilitarianism. For the society with the highest average utility would be one in which my expected utility is highest *given* that I can rationally take myself to be equally likely to be any of the subjects of that society.

Adding ignorance of acceptance levels seems to lead to the same sort of result. For now my expected utility will be given by the average of the average utilities across all the different possible acceptance levels. Which is just to say that I should choose a form of variable-rate rule-utilitarianism which focuses on average utility rather than aggregate utility.

My presentation of this argument has, as I noted at the outset, been extremely brief and I do not pretend to have shown that it can really be made to work. It seemed interesting and suggestive enough to me to be worth at least sketching, though. For if the argument could be made to work, then Parfit's intriguing mountaineering metaphor might still be vindicated, albeit in a very different way from the way he himself had in mind.

7. Conclusion

We are not all climbing Parfit's mountain. Parfit's own derivation of rule-consequentialism from Kantian contractualism is deeply problematic, and for reasons which emerge from his own discussion of the Ideal World objection. Nor is there any easy fix to this problem. If we revise what I have called Parfit's 'Master Argument' so that it avoids the Ideal World Objection, then two new problems emerge. For we now face the problem that (C*) is both crucial to the new derivation but highly implausible and the closely related objection that the resulting version of rule-consequentialism is itself vulnerable to the Nihlism for the Wrong Reasons Objection.

There is a form of rule-consequentialism which avoids all of these objections, though – variable-rate rule-utilitarianism. I have argued that variable-rate rule-utilitarianism shares Parfit's insight that rule-utilitarians need to define right action in such a way that rightness supervenes on all of the different (non-zero) acceptance levels, but does so in a way that avoids the fatal Nihilism for the Wrong Reasons Objection to which Parfit's own version of the view is vulnerable. Finally, I have suggested, very tentatively, that there may be a way of deriving variable-rate rule-utilitarianism from a plausible and interesting (if more Rawlsian than Parfit's preferred Scanlonian) form of Kantian contractualism. If such a derivation can be made to work, then perhaps Parfit's optimistic hypothesis about convergence of secular moral theories can be

vindicated after all. We may all be climbing the same mountain, even if it is not the mountain Parfit has suggested we are climbing.[7]

University of Edinburgh
Philosophy Department
Dugald Stewart Building
3 Charles St
Edinburgh EH8 9AD
mridge@staffmail.ed.ac.uk

[7] Thanks to Campbell Brown, Matthew Chrisman and the participants at the conference on Parfit's book which was the catalyst for this paper for helpful comments and suggestions.