Parfit's Argument for Universal Acceptance of Rule Consequentialism

Adam Berman

Dickinson College

Can He Climb the Mountain?

An Analysis of Parfit's Argument for Universal Acceptance of Rule Consequentialism
Through the Lens of Michael Ridge's "Climb Every Mountain?"

In *On What Matters*, Derek Parfit proposes a convergence theory for the three main contemporary moral theories. He claims that Kantianism, Contractualism, and Consequentialism can be seen as different versions of the same view. He considers his convergence theory one of the attempts at "climbing the same mountain on different sides" (Parfit 419). However, Parfit's convergence theory is flawed. Specifically, his attempt to combine Kantianism and Consequentialism comes up against the *ideal world objection.* Furthermore, attempts to reconstruct the argument in order to save the argument from the *ideal world objection* lead to further insurmountable objections. Parfit's argument for the universal acceptance of *rule consequentialism* fails to show that there are optimific principles that everyone ought to follow. In this paper I will briefly reconstruct Parfit's argument for the universal acceptance of rule consequentialism which I will call the argument for UARC. I will then explain Parfit's attempt at overcoming *ideal world objection*. Michael Ridge in his paper, "Climb Every Mountain," offers a way of defeating the ideal world argument but show that this version of the argument for UARC leads to very undesirable conclusions about the overall status of morality and thus fails. Finally, I will address a Parfitian revision of Ridge's proposal and raise an epistemic objection to this revised argument.

Parfit begins by proposing, "(A) Everyone ought to follow the principles whose universal acceptance everyone could rationally choose, or will" (Parfit 400). This Contractualist premise does not require everyone to actually agree on a set moral code, instead (A) only requires that everyone act on the principles that everyone would be fine with if everyone hypothetically agreed to follow them. Parfit continues, "(B) Everyone could rationally choose whatever they would have sufficient reason to choose" (Parfit 378). We should choose principles, Parfit says, based on the sufficiency of their reasons. So while (A) tell us that everyone ought to follow a soon to be determined set of principles, (B) tells us how we will determine which principles will make up that set.

Parfit then proposes that we must follow *UA-optimific* principles, "(C) There are some UA-optimific principles. (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” (Parfit 378). Compliance with a UA-optimific principle would make things go best. This does not necessarily mean that it would make things go best at a specific instance for a specific person. Instead, following a UA-optimific principle causes things to go best overall for all people in the sense that the world is better off if people follow these UA-optimific principles. As a result, Parfit claims that the principles that make things go best for everyone give us decisively strong impartial reasons to choose those principles, principles that cannot be outweighed by impartial reasons arising from other conflicting principles. Thus UA-optimific principles are the only principles that give us sufficient reasons to choose them over other conflicting principles. Since everyone could will the universal acceptability of UA-optimific principles, Parfit concludes that everyone must follow these UA-optimific principles. For clarity I will now restate Parfit’s argument as it appears in its final form:

> (A) Everyone ought to follow the principles whose universal acceptance everyone could rationally choose, or will.
> (B) There are some principles whose universal acceptance would make things go best.
> (C) Everyone could rationally will that everyone accepts these principles.
> (D) These are the only principles whose universal acceptance everyone could rationally will.
> Therefore
> UARC: These are the principles that everyone ought to follow. (Parfit 400)

Premise (A) opens Parfit’s argument up to the ideal world objection. It is probably true that there are some principles that would make the world best if everyone accepted them. For example, the world would probably be best if everyone accepted pacifism. However, the type of principle that would make things go best if everyone accepted it could also make things go terribly if any number of people <u>fewer</u> than everyone accepted it. As such, everyone ought not <u>always</u> follow these principles. Parfit writes, “If everyone outside of Germany had been pacifists, that would have allowed Hitler to dominate the world, with effects that would have been likely to be even worse than this terrible war” (Parfit 312). Principles that make things best upon universal acceptance can lead to terrible consequences when faced with malicious individuals. In addition, accidents and basic human instinct can cause people to fail to accept a principle. Those who accept the sort of principle that requires universal acceptance may end up responsible for accidental wrongdoing.

The *Mistake* thought-experiment illustrates one of the problems with universal acceptance raised by the ideal world objection: everyone is saved if two individuals both do action A, some people are saved if both people do action B, and nobody is saved if one person does A and the other person does B (Parfit 313). Both people accept the maxim that everyone should do A. A problem arises if one person mistakenly does B; the second person is caught between breaking a maxim he believes should always be followed and following a maxim that will cause a greater wrong than not breaking the maxim.

This predicament opens up a significant epistemic problem. Often it is very difficult to tell whether a person is acting in accordance with a certain principle. Certainly, it is tough to determine beforehand whether or not the person you interact with accepts your principle. In *Mistake*, it seems like the circumstances would be rare in which the first person would have access to which choice the second person will make before the choice is made. For example, in the event of a fire emergency responders must execute their evacuation protocol regardless of whether they are related to anyone in the building. Let's say that the emergency protocol requires everyone to check their designated areas as thoroughly as possible; there is no overlap so that the check can be accomplished as thoroughly and quickly as possible. All emergency responders accept this principle before an emergency however if emergency responders are called to a school attended by a child of one of the responders, the responder might abandon the principle in favor of searching for his child. If the other emergency responders are unaware that this responder has a child that attends this school, they are not unlikely to be able to predict this. As a result, an area of the building may go unchecked by emergency responders who are executing the regular protocol. The other emergency responders never had a truly live option of avoiding the harm that can result from an unchecked room during an emergency. So not only is it unrealistic to expect that people will always accept optimific principles, but it is also tough to determine when people will fail to accept them.

For now, let's set aside the epistemic objection and instead focus on the ideal world objection. Parfit points out that those principles that require a person to assume universal acceptance are not truly optimific since these principles do not make things go best in any world that is not ideal. Instead, he writes, "Here is a better maxim: M2: Do whatever I could rationally will everyone to do, unless some other people haven't acted in this way, in which case do whatever I could rationally will that, in these circumstances, other people do" (Parfit 314). The first part of principles like these specifies the desired universally accepted behavior while the second part specifies conduct in case compliance with the first part is not complete. Parfit argues that the optimific principle calling for pacifism should be revised to look somewhat like: "Never use violence, unless some other people have used aggressive violence, in which case use restrained violence when that is my only possible way to defend myself or others" (Parfit

315). This proposal makes sense on the surface. The first part of the principle requires that people accept pacifism and the second part gives a provision that deals with those who fail to accept pacifism.

The M2 revision seems to permit too much. It might be best to require as serious retribution as possible to those who fail to meet the condition of the first part so as to force everyone to accept the first part of the principle. Parfit suggests a principle like: “Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can” (Parfit 315). Because of the seriousness of the consequences dictated by the second part, people may be more likely to accept the first part of the principle. Ridge writes, “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” (Ridge 61). This stance seems absurd however since it places those who accept the principle in a terrible position. A person who accepts this maxim must be willing to commit extreme violence if people fail to accept pacifism! M2 places us in a position similar to *Mutually Assured Destruction*: at some point someone will fail to accept pacifism forcing everyone else who accept this principle to engage in mass murder, a practice they may be unwilling to undertake. People who accept this principle may have to choose between abandoning the principle or engaging in actions that are clearly morally wrong.

The solution to this problem is to revise Kant’s *Law of Nature Formula*. The original *Law of Nature Formula* from which premise (A) derives, states that “It is wrong to act on some maxim unless we could rationally will it to be true that everyone acts upon it” (Parfit 308). This formulation is problematic since it implies that we may always act on a principle if we will that everyone act on it. It is puzzling that Parfit includes premise (A), which is so clearly problematic, in his final argument for UARC. He acknowledges in §45 that it is objectionable to require a person to continue to act on a principle that requires universal acceptance to be optimific even when it is not universally accepted. Ridge points out that Parfit corrects this problem by making a “friendly amendment to Kantian contractualism” (Ridge 64). Let's see how the argument, modified by this “friendly amendment,” deals with the ideal world objection.

In order to eliminate problems about the optimificity of certain principles that may not be universally accepted, Parfit proposes a change to Kant’s *Law of Nature Formula*. He writes, “LN4: It is wrong for us to act on some maxim unless we could rationally will it to be true that this maxim be acted on by everyone, and by any other number of people, rather than by no one” (Parfit 317). Unlike the previous formulation of the *Law of Nature Formula*, this version forces us to find principles relevant to a less than ideal world. Since it is unrealistic to expect all people to accept a formula, this revision narrows the scope to principles that are optimific as long as at least one person accepts them.

Ridge provides this version of Parfit’s revised argument for UARC:

> (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 could 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
> [UARC*:] These are the principles that everyone ought to follow. (Ridge 66)

This revision of Parfit’s UARC argument allows the consideration of principles that fit a more reasonable and practical definition of optimific. According to this account, principles can be optimific as long as they make things go best if at least one person accepts them; we have sufficient reason to accept a principle even if it is not universally accepted. The UARC* argument fends off the ideal world objection since now principles can be optimific even if fewer than everyone accepts them.

With this revision UARC is now more applicable to the real world. However, this revision opens the argument up to further objections. Premise (C) is clearly defensible; it is likely that if everyone truly accepted pacifism, it would make things go best. Unfortunately this is almost certainly irrelevant since it will likely never be the case that pacifism will be universally accepted. Premise (C*) on the other hand is a much tougher claim to defend. Ridge explains that according to (C*) “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” (Ridge 67). It seems possible, or even likely, that as the level of acceptance changes, the principle that is optimific will change. At sixty percent acceptance, a different principle may be optimific than at ninety percent acceptance. For example, a principle like the one mentioned above calling for pacifism and only as much violence as is necessary to prevent aggressive violence is probably optimific at a high level of acceptance. However, a low level of acceptance may cause greater suffering, since the group that would administer violence against the initially violent group may not be strong enough to defeat the aggressors. The principle would then be non-optimific at a low level of acceptance.

It seems unlikely that there will be a principle that is optimific at all levels of acceptance. Since the conditions for (C*) are not met, the conditions for (D*) are not met either. Thus there are no principles of this sort that provide us with sufficient reason to follow them. Ridge argues that this is a major problem, “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” (Ridge 68). As we have discussed earlier, it seems likely that there is no single principle that is optimific for each and every level of acceptance. Thus the UARC* argument seems to result in nihilism!

UARC* fails because it is unlikely that a principle will be optimific for all levels of acceptance. To avoid this problem, Parfit might look to change the requirement that a principle be optimific across all levels of acceptance. He would need to propose a version of UARC relating to an individual level of acceptance. He might propose something like LN’: It is wrong for us to act on some maxim unless we could rationally will that this maxim be acted on by any one number of people, rather than by no one. This proposal would allow a version of UARC in which a principle can be said to be optimific for at least one acceptance level. Thus we could revise the argument for UARC to the following form:

(A') Everyone ought to follow the principles whose acceptance by any one 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 any one number of people, rather than by no one, would make things go best.
(D') These are the principles whose acceptance by any one number of people, rather than by no one, everyone could 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
UARC’: These are the principles that everyone ought to follow.

Using this strategy, Parfit would be able to do away with the otherwise insurmountable objection that since we cannot have optimific principles for every level of acceptance, nihilism is inevitable. Instead, this new version of the argument for UARC only requires principles that are optimific for, at the very least, one specific level of acceptance. This strategy has the advantage of allowing a variety of optimific principles based on differing levels of acceptance and in so doing the UARC' argument leads to principles that imply decisively impartial reasons. Since there are sufficient reason to choose these principles over all other principles, this modification reopens the door to the possibility of moral action and thus avoids the objection that leads to nihilism.

While the UARC’ argument answers the objection that there are no optimific principles, it opens the door to a significant epistemic problem similar to the epistemic problem mentioned earlier. According to the UARC’ argument, there are optimific principles that are optimific only for specific acceptance levels. While the UARC’ argument opens the possibility that an optimific principle might be optimific for multiple, many, or even all levels of acceptance, it is also possible there is a different optimific principle for each and every level of acceptance. Thus in order to determine which principle is optimific, a person would first need to determine the current level of acceptance.

A question, one that is closely related to the epistemic problem raised earlier, can now be raised: how would you go about determining the current level of acceptance? It seems likely that such a determination, at least to an acceptable degree of accuracy, would be impossible. One might require people to determine the exact acceptance level of a principle in order to make an optimificity judgment. You would need to somehow tap into the mind of every person on earth at the same time and be confident that nobody would change his mind. This strategy is impossible, and any attempt to ascertain the exact acceptance level of a principle will ultimately fail. But, without knowledge of the exact acceptance level, it is impossible to determine which principles are currently optimific.

A more reasonable approach would require people to determine the acceptance level of a principle with a high degree of certainty in order to make an optimificity judgment. While exact certainty would not be required, you would still need to be confident that you had a well grounded understanding of all relevant factors required for a determination of an acceptance level. Yet by this standard a determination would still be impossible. At best, a person has access to the norms of his family and friends. Perhaps one could extrapolate from this information to the norms of people ideologically and socially similar to people one knows. However, even in small group of people, individuals often have vastly different moral beliefs. When you expand that group to include a city, a country, or the entire world, it becomes clear that the moral beliefs of certain groups will be epistemically unavailable to anyone attempting to make an optimificity judgment. In addition some groups will be unable to share their moral beliefs with a random person making these sorts of judgments thousands of miles away. You would have to speak many other languages, engage in an immense research project, and perform a complicated statistical analysis.And more than that, some groups may have moral beliefs unintelligible to those with differing beliefs. Because of the immense range of moral beliefs and the logistical problems of communicating them, we must conclude that it is impossible to develop a well grounded understanding of all relevant factors involved in making an accurate judgment of acceptance level. As a result, by this standard it is impossible to judge whether or not a principle is optimific.

There is one further obstacle preventing accurate judgments of principle optimificity: the dynamic nature of acceptance levels. People are influenced by the words and actions of other people. If a religious leader tells his congregation that acting in a certain way is good, it is probably more likely that members of his congregation will act in this way. If a person sees someone they disdain doing certain things, they may be less likely to do similar things in the near future. In this way, the acceptance levels of principles are constantly changing. A person's determination that someone has accepted or rejected a principle can influence their own acceptance or rejection of that principle. Because people’s moral beliefs are so dynamic, it is impossible to ever determine the acceptance level of a principle for long enough to act on it.

Now that we have reached the conclusion that determining the acceptance level of a principle is impossible, we are left with a dilemma. Either we decide to accept principles that we believe to be optimific even though we have no legitimate justification for this belief, or else we must agree that it is impossible to determine optimific principles. Those who choose the first option must face the possibility that, since they have no real basis for their belief that their principles are optimific, they may be wrong about their principles being optimific. The *Mistake* case discussed earlier showed the harm that can come from blindly following a principle that is no longer optimific. Since people who choose the first option are acting on principles they do not know to be optimific, they may always be blindly following principles that may cause harm. On the other hand, those who choose the second option are left blindly grasping at unknowable optimific principles through an endless fog of moral ambiguity. Without the moral guidance that optimific principles provide, it is unclear how we are supposed to make moral choices.

These analyses of Parfit’s argument for the universal acceptance of rule consequentialismmake clear that the argument for UARC is flawed. The initial argument leads to a set of principles that may be optimific in an ideal world but are not applicable in the real world. However, attempts to revise the argument for UARC in order to allow real world applicability result in nihilism or moral ambiguity. Further these revised arguments are vulnerable to empirical evidence. All you need to defeat the requirement of an optimific principle encompassing all levels of acceptance is to show empirically that a single optimific principle is non-optimific at some acceptance level. Similarly, the argument allowing different optimific principles at different levels of acceptance is vulnerable to the inaccuracy of our predictions about acceptance levels. At this point, it appears that the argument for UARC fails to show that there must be optimific principles that people must follow. If optimific principles do exist, it seems unclear how we should know them or when we should follow them.

Works Cited

Parfit, Derek. *On What Matters*. Ed. Samuel Scheffler. Vol. 1. Oxford: Oxford UP, 2011.

Ridge, Michael. "Climb Every Mountain?" *Ratio* 22.1 (2009): 59-77.