DISCUSSION NOTE

A COUNTEREXAMPLE TO PARFIT'S RULE CONSEQUENTIALISM

BY JACOB NEBEL

JOURNAL OF ETHICS & SOCIAL PHILOSOPHY

DISCUSSION NOTE | JULY 2012
URL: WWW.JESP.ORG

A Counterexample to Parfit's Rule Consequentialism

Jacob Nebel[1]

IN *ON WHAT MATTERS*, DEREK PARFIT ARGUES that the most plausible versions of Kantianism and Contractualism coincide with a form of Consequentialism, and the resultant principle might be "the supreme principle of morality" (342).[2] Parfit revises Kant's formulas to arrive at the following principle:

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

According to Parfit, the principles whose universal acceptance everyone could rationally will are just the principles whose universal acceptance would make things go best. If that is true, then the single true morality requires the following principle:

> (UARC) Everyone ought to follow the principles whose universal acceptance would make things go best (377).

We accept a principle just when we believe that it is true (341).[3] We follow a principle when we succeed in doing what it requires (405). If a principle's universal acceptance in a world would make things go best, then that principle is *UA-optimific* in that world (425). UARC requires that everyone follow the UA-optimific principles.

Parfit argues that KC may be the supreme principle of morality, and that KC implies UARC. A supreme principle of morality would tell us what we ought to do in all possible worlds. The modal status of this principle is a supposition of Parfit's metaethical views: If our fundamental normative principles were only contingent truths, then Parfit believes we would have to know them through empirical discovery (128).[4] But if moral properties are non-natural properties, as Parfit argues, then empirical discovery cannot reveal them. So, if the supreme principle of morality were not true in all possible worlds, then it would be a genuine mystery how we could know it to be true, even in our world.

---

[1] I am most indebted to Peter Singer for his endless patience, support and guidance on several drafts of this paper. I am extremely grateful to Derek Parfit and Larry Temkin for their helpful criticism and generous suggestions. I also thank Richard Yetter Chappell, Ben Cogan, Neil Conrad, Ryan Davis, Alex Gregory, Jussi Suikkanen, Gideon Rosen, Matt Wage and anonymous reviewers for their feedback.

[2] All references are to Parfit (2011) *On What Matters: Volume I*, Oxford: Oxford University Press, unless otherwise noted.

[3] Parfit might appeal to another conception of what it would be to accept a principle. I discuss this possibility in the last section of this paper.

[4] See *On What Matters: Volume II*, p. 489.

I argue that UARC is false in at least one world, so it is not the supreme principle of morality.

Consider a world in which no one's moral beliefs have any motivating force at all. In this *Indifference World*, no one cares about the moral facts, even those of which they are aware. Indifference World might contain people who act (by our standards) morally, but not *because* they believe their acts to be right: Perhaps they fear retribution or believe that kindness is in their own interest. In Indifference World, the consequences of accepting one set of principles (by any number of people) would be the same as the consequences of accepting any other set of principles, because no one's motivations would change as a result of changed moral beliefs. My argument runs as follows:

(A1) UARC is false in Indifference World.

(A2) Indifference World is a possible world.

Therefore,

UARC is false in at least one possible world.

I proceed with a defense of (A1). I then argue that the Rule Consequentialist objection to (A2) is not available to Parfit. I conclude by considering two of Parfit's objections to (A1).

1. UARC in Indifference World

When applying UARC, we could say one of two things about Indifference World. We might first say,

(B1) There are no UA-optimific principles in Indifference World,

since the universal acceptance of any one set of principles in this world would have no better outcome than that of any other set. On the other hand, we might say,

(B2) Every principle is UA-optimific in Indifference World,

since no principle is worse than any other. Neither implication fares well for UARC, but I will first assess (B1).

If we take this first route, then UARC implies that there are no principles that everyone ought to follow in Indifference World. This means there are no principles of which it is true that everyone, in all possible worlds, ought to follow. If there are no principles that everyone ought to follow in all possible worlds, then UARC is not a necessary moral truth, because no ought-style principles (including UARC) are true in all possible worlds. Therefore, someone who accepts (B1) must conclude that UARC is not the

supreme principle of morality.

Moreover, if there are no principles that everyone ought to follow in Indifference World, then there are no moral obligations or prohibitions, making all acts and omissions in that world morally permissible. If there is a supreme principle of morality, however, it is unlikely that it permits acts like rape, murder and torture. Parfit might object that we should not expect the supreme principle of morality to apply to people who are completely indifferent to morality. But that expectation, I believe, is entirely legitimate: Even if amoral agents cannot be morally blameworthy, it seems clear that they can act wrongly. Moreover, we could imagine the people in Indifference World being sensitive to the *non-moral* features that make acts right or wrong, so they seem like (and perhaps are) moral agents even though they are not disposed to follow their moral beliefs. I find it hard to believe that a supreme principle of morality might not apply to people who care about the morally relevant features of acts for their own sakes, just because they do not care about rightness or wrongness as such.[5] One might suggest that, even if UARC does not apply to Indifference World, some other moral principles might. But it seems that the wrong-making features of people's acts in Indifference World are not fundamentally distinct from the wrong-making features of acts by people who are disposed to follow their moral beliefs, so it seems arbitrary to introduce some other moral theory to cover Indifference World.

There is another reason that (B1) is problematic for Parfit. According to Parfit's Formula of Universally Willable Principles,

> (FUWP) An act is wrong unless such acts are permitted by some principle whose universal acceptance everyone could rationally will (341).

Parfit claims that Kantian Contractualism is just a simplified version of FUWP, and Parfit's argument for convergence requires that UARC yield the same results as FUWP. But if there are no UA-optimific principles in Indifference World, then there are no principles in Indifference World whose universal acceptance everyone could rationally will. And if there are no such principles, then there are no such principles that permit any act. Thus, by FUWP, all acts in Indifference World are wrong, because we cannot satisfy the "unless" condition in FUWP. But we just found that UARC makes all acts *permissible* in Indifference World. So, UARC yields a very different result than FUWP. This conclusion undermines Parfit's argument for convergence.

Now turn to (B2). When Parfit considers cases where two or more outcomes are not worse than any other outcome, he uses the word "best" to describe those outcomes, in addition to the simple cases where one outcome is better than every other outcome (373). Parfit does not consider cases where *every* outcome is equally good (or bad), but suppose that we take his

[5] See Gideon Rosen (2009) "Might Kantian Contractualism Be the Supreme Principle of Morality?" *Ratio* XXII: 96.

point here as implying a definition of "best": An outcome is best if it is not worse than any other outcome. On this view, we should accept that every principle is UA-optimific in Indifference World.

If universal acceptance would make things go best for any and every principle in Indifference World, then UARC implies that everyone in Indifference World ought to follow any and every principle. This implication is implausible on its own, since it is unlikely that everyone ought to rape, murder and torture each other, but it is even more implausible because UARC would demand that everyone follow *contradictory* principles. It would be the case that everyone both ought to and ought not to rape, murder and torture each other. Every act, then, would be both morally required and morally wrong according to UARC. This implication counts against the view that UARC is the supreme principle of morality, which should tell us what we ought to do without contradiction.

I have considered (B2) to suggest that it does not matter what precisely we say when applying UARC to Indifference World: The conclusion, in either case, is that UARC is not a necessary moral truth.

2. Parfit's Metaethics and the Possibility of Indifference World

Rule Consequentialists may argue that Indifference World is not metaphysically possible on the following grounds. Accepting a set of principles, they might argue, is not merely the act of believing the principles to be true propositions, absent some corresponding reflection in motivation to act or to be disposed to react in appropriate ways. The main defenders of Rule Consequentialism – including Richard Brandt, Brad Hooker and Tim Mulgan – consider the expected consequences of rule acceptance to be largely a function of the causally efficacious dispositions to act and react in certain ways.[6] In other words, acceptance is a matter of internalization, which includes but is (crucially) not limited to compliance. On this view, the concept of rule acceptance precludes the possibility of a world in which rule acceptance has no effect on people's motivations.

While this response is available to acceptance-based Rule Consequentialists like Brandt, Hooker and Mulgan, it is not available to Parfit, and for very important reasons. In defending his convergence thesis, Parfit revises Kant's Moral Belief Formula into the Formula of Universally Willable Principles, which becomes the Kantian Contractualist formula and then UARC. In introducing FUWP, Parfit claims that belief implies acceptance. Parfit writes, "When people believe that some kind of act is morally permitted, they accept some principle that permits such acts" (341). If belief were not a sufficient condition for acceptance, Parfit would not have this crucial link between Kant's Moral Belief Formula and UARC. Parfit's argument for convergence therefore requires that belief is a sufficient condition for acceptance. (For

[6] See, e.g., Hooker (2000) *Ideal Code, Real World*, Oxford: Oxford University Press, p. 75.

readers who doubt this claim, I discuss it further in the next section.) The remaining question, then, is whether moral beliefs are intrinsically motivating.

Parfit argues, in Part 6 of *On What Matters*, that moral beliefs are not intrinsically motivating: We can have a moral belief without having the slightest motivation to act accordingly. Judgment internalists disagree, and the impossibility of Indifference World is sometimes used as an argument for at least some modest form of internalism.[7] But Parfit's externalism is a crucial component of his metaethical picture. Parfit considers the following Humean Argument for noncognitivism:[8]

> (C1) It is inconceivable that we might be sincerely convinced that some act was our duty, but not be in the slightest motivated to act in this way.
>
> (C2) If moral convictions were beliefs, such a case would be conceivable.

Therefore,

> Moral convictions cannot be beliefs, but must be some kind of desire, conative attitude, or other motivating state.

Parfit grants that a weak version of the Humean Theory of Motivation is undeniable:

> (HTM) No belief could motivate us *all by itself*, since no belief could motivate us unless it is also true that we are *disposed* to be motivated by this belief.

And he grants that HTM is enough to support (C2). So, he has to reject (C1), which is a version of judgment internalism. If he does not reject (C1), Parfit must accept the noncognitivist conclusion.

Parfit claims that (C1) seems plausible only because it refers to sincere convictions or beliefs. We only call a moral belief "sincere" or a "conviction" if the believer is at least somewhat motivated to act accordingly. But that does not entail judgment internalism, which requires the following revision to (C1):

> (C1*) It is inconceivable that we might *believe* that some act was our duty, but not be in the slightest motivated to act in this way.

---

[7] See James Lenman (1999) "The Externalist and the Amoralist," *Philosophia* 27; Jon Tresan (2009) "The Challenge of Communal Internalism," *The Journal of Value Inquiry* 43. For a response that defends the possibility of Indifference World, see Joshua Gert and Alfred Mele (2005) "Lenman on Externalism and Amoralism: Interplanetary Exploration," *Philosophia* 32. Furthermore, the intuition that communal amoralism is impossible may really be tracking the connection between moral utterances and motivation, rather than judgment internalism, according to Caj Strandberg (2011) "The Pragmatics of Moral Motivation," *The Journal of Ethics*.

[8] *On What Matters: Volume II*, pp. 382-83.

The difference between this premise and (C1) is that (C1*) replaces the phrase "be sincerely convinced" with "believe."

Parfit offers two counterexamples to (C1*). The first is a case of moral knowledge: Perhaps the amoralist does not have a sincere conviction, but she might *know* that some act is her duty, and knowledge implies belief. The second is a case of deep depression: The depressed agent may lose only her motivation to do what she thinks she has most reason to do, not her normative beliefs. These cases are meant to show that (C1*) is false, and that claims like (C1) only seem true because they involve sincerity and conviction over and above belief.

Parfit's rejection of judgment internalism is key to his defense of cognitivism in Part 6 of *On What Matters*. Parfit's account of moral motivation implies that we are motivated to follow our true normative beliefs insofar as we are fully substantively rational, because we would then have the disposition required by the weak, undeniable version of the Humean Theory of Motivation. But that does not rule out the possibility of Indifference World; it just means that the agents in Indifference World are not fully substantively rational, and Parfit gives no argument for why a world in which no agents are fully substantively rational is impossible. I do not take a stance here on whether this metaethical picture is preferable to an internalist one.[9] But, if I am right that belief is sufficient for acceptance, then by Parfit's own lights, people can accept a principle that permits, requires or forbids some act without being even partly disposed to act accordingly – which is exactly what happens in Indifference World.

3. Parfit's Responses

Parfit offers several responses to my claim that UARC is false in Indifference World; I shall discuss only two of them here.[10]

Parfit's first response is that my argument makes two conflicting assumptions about the modal status of moral principles. When arguing that UARC is false in Indifference World, I assume:

> (D1) When applying UARC, we should ask which are the principles whose universal acceptance would be best in some particular world.

I assume (D1) because my objection appeals to the fact that no one's acceptance of any moral principles would have any effects in Indifference World. But Parfit also thinks I assume:

---

[9] In (forthcoming) "Internalists Beware – We Might All Be Amoralists," *Australian Journal of Philosophy*, Gunnar Björnsson and Ragnar Francén Olinder raise the cynical hypothesis that our world is Indifference World. They concede (rightly, I think) that this hypothesis is unlikely but not conceptually impossible.

[10] In correspondence. My replies to Parfit's responses likely face further problems that I have not considered.

> (D2) UARC, and the particular moral principles which would be selected by UARC, are necessary truths that apply to all possible worlds.

Parfit argues that, if we are looking for the true moral principles that apply to all possible worlds, the relevant task is to figure out the moral principles whose acceptance *in all possible worlds* would make things go best. We can, therefore, ignore Indifference World because the acceptance of different principles would not have different effects in this imagined world.

I think we should reject (D2), but let me first explain how my argument does not assume (D2). I do assume that the supreme principle of morality is a necessary truth that applies to all possible worlds. But the principles it selects may be contingent. In my defense of (A1), which claimed that UARC is false in Indifference World, I argued that UARC implies either that there are no principles we ought to follow (including UARC) or that we ought to follow every principle, including repugnant principles and principles that contradict each other. The implausibility of these results does not depend on (D2).

Parfit might respond that when I reject UARC because it selects repugnant principles in Indifference World, I am assuming those principles to be necessarily false and, therefore, assuming (D2). But my argument only requires the weaker claim that those principles are false in Indifference World. My view is that rape, torture and murder cannot be made permissible or obligatory by the consequences of moral beliefs alone, but perhaps they could be made permissible or obligatory by other facts that obtain in some possible worlds – for example, if rape, torture and murder had unusually good consequences instead of their usual, horrible ones. Universal acceptance without compliance, however, is not enough to justify rape, torture and murder. If it *were* true that everyone ought to follow such repugnant principles in any world, it would not be because the consequences of their universal acceptance were no worse than those of their alternatives. So, my argument does not assume (D2).

Moreover, (D2) seems implausible. I do not know how we would realistically go about selecting principles under UARC if we had to assess the consequences of universal acceptance across all possible worlds, nor do I see why we should care about the consequences for merely possible worlds when determining the optimific principles in our world.[11] For example, consider an *Evil World* in which everyone tries to do whatever they believe is wrong simply because they want to act wrongly, or a *Mistakes World* in which everyone's attempts to follow their accepted principles lead to their inadvertent violation. If those worlds and worlds like them are possible, we should not have to evaluate the effects of our moral principles' acceptance in those worlds to

[11] Weighing these consequences across all possible worlds seems even more difficult if there are infinitely many such worlds.

figure out whether we should follow them in our world. If we did evaluate those effects, they would skew the evaluation of principles that are UA-optimific in the actual world. So, for example, a principle forbidding torture would end up with a fair amount of compliance in the actual world, so it would be an improvement over a principle permitting torture, but the opposite would happen in Evil World and Mistakes World. Therefore, I think we should reject (D2).

When Parfit argues, in Part 6, that normative truths apply to all possible worlds, his claims are restricted to the most *fundamental* normative facts.[12] The badness of pain is one such truth, as is the supreme principle of morality. But the principles selected by the supreme principle of morality may be different in worlds with different kinds of agents and laws of nature.

Parfit's second response is that UARC may appeal to a different conception of a moral principle and of what it would be to accept some principle. We accept a principle in this sense when we *decide to try to follow* this principle. Parfit writes, in a note, that Kantian Contractualism focuses on principles that "can be more like the maxims to which Kant appealed"; these maxims are "like rules or policies," not beliefs which can be true or false (471).

It is unclear to me how this response is supposed to cohere with the rejection of Kant's focus on maxims. Parfit argues that, in order to avoid many of his objections to Kant's formulas, we should drop Kant's appeal to maxims in the sense that covers policies (341). And Parift offers convincing, independent reasons why Kant's Moral Belief Formula is more plausible than Kant's formulas that focus on maxims (320, 471). Parfit may be correct that there is a sense in which principles are like maxims or policies, but which does not fall prey to Parfit's own objections to Kant's focus on maxims or policies. But I do not know what that sense is.

Parfit might add that the supreme principle should not assess moral beliefs, since we would then be ignoring whether these beliefs are true. But, as Parfit argues, Contractualist formulas can include a restriction on deontic beliefs, so our reasons to reject some principle would not include the belief that the principle is wrong (416). Similarly, Consequentialist formulas can use the word "best" in its deontic-value-ignoring sense, so whether some outcome is best is unaffected by whether the principles leading to that outcome are right (474). With these restrictions in place, it is not implausible to ignore whether the principles under consideration are true when applying the Contractualist and Consequentialist thought experiments to moral beliefs.

At this point, Parfit might instead appeal to the acceptance of *sincere* moral beliefs, which (as discussed in the previous section) require some motivation. But the resulting formula seems to have implausible implications in worlds like Mistakes World. In Mistakes World, things would go best if everyone sincerely believed a principle prohibiting the prevention of pointless

[12] *On What Matters: Volume II*, pp. 307, 489.

suffering and a principle requiring torture for fun.[13] In this kind of world, everyone's attempts to follow these principles would prevent pointless suffering and torture, because the disposition to comply with these principles would lead to their violation. Therefore, these repugnant principles would be UA-optimific in Mistakes World. UARC would then imply, in this kind of world, that preventing pointless suffering is wrong and that torturing others for fun is obligatory.[14] But even people in Mistakes World ought to prevent pointless suffering, and even people in Mistakes World ought not to torture others for fun – although they ought to *try* to do the opposite – so this version of UARC seems to me false.

4. Conclusion

UARC is not, I believe, the supreme principle of morality. But that does not make Parfit wrong about convergence among Kantians, Contractualists, and Consequentialists. Perhaps some compliance (as opposed to acceptance) version of Rule Consequentialism is the supreme principle of morality.[15] Parfit considers this possibility, and he notes that we could revise Kantian Contractualism to cover this focus on compliance: "Everyone ought to follow the principles that everyone could rationally will to be universal laws," and a principle could be a universal law by being universally followed (407). While I cannot explore this suggestion here, I think it is a more promising candidate for the supreme principle of morality, at least for judgment externalists like

---

[13] Parfit might object that, if everyone tried to follow principles requiring torture and preventing beneficence, then things would go worse because everyone would have such bad motives. If the badness of having those motives outweighed the badness of increased torture and suffering, then these principles would not be UA-optimific. So, perhaps UARC does not require such repugnant principles in Mistakes World. But if people knew their attempts would fail, then these motives might not be so bad. And, even if they are bad, I doubt that the badness of these motives is worse than the badness of the torture and suffering that would result from the acceptance of more benign principles.

[14] Parfit might revise UARC to appeal to the principles that *would be* optimific if they were efficacious in the normal way, with no abnormal, distorting features (and no factors leading to "conditional fallacy"-type problems that may arise from this stipulation – e.g., the possibility that, without the distorting features, the agents in Mistakes World would love being tortured). Thanks to Richard Yetter Chappell for this suggestion. I am not sure what non-arbitrary sense of abnormal, distorting factors would exclude Mistakes World but include ordinary failure to follow one's principles, and it is unclear to me that we can avoid "conditional fallacy"-type problems in this way without leading to similar problems. Even if we could solve these problems, it may not make sense to care about whether one's acts conform to a gerrymandered principle of this kind for such conformity's sake, and that may cast doubt on the revised UARC as the supreme principle of morality. See Rosen, "Kantian Contractualism," 93-96.

[15] Parfit suggests that these different versions of Consequentialism may cover different parts of our moral theory (407). But the consequences of accepting and trying to follow a principle seem to me more plausibly relevant to the part of our moral theory that says which principles we ought to accept and try to follow – not which principles we ought to follow.

Parfit. And, as Parfit acknowledges, compliance-based Rule Consequentialism is closer to Act Consequentialism than UARC is, so any hope of convergence *within* Consequentialism lies with this strategy.

Jacob Nebel
Princeton University
Department of Philosophy
jnebel@princeton.edu