REVIEW OF *UTILITARIANISM: FOR & AGAINST*

1. Anti-realism

The utilitarian contributor writes from a non-cognitivist perspective and thus "renounce[s] the attempt to *prove* the act-utilitarian system."[1] I've spent a lot more time reading moral realists, so it was interesting to see this perspective. His goal is mainly to try to remove obstacles to belief in utilitarianism so that people's natural attitudes can lead them there:

> It is possible … that many sympathetic and benevolent people depart from or fail to attain a utilitarian ethical principle only under the stress of tradition, of superstition, or of unsound philosophical reasoning. If this hypothesis should turn out to be correct, at least as far as these readers are concerned, then the utilitarian may contend that there is no need for him to defend his position directly, save by stating it in a consistent manner, and by showing that common objections to it are unsound.[2]

I think that's a sensible approach; my experience is that people's problem with utilitarianism is not usually that they don't grasp the motivation for it all, but rather that they think it has some problematic implications. Still, the anti-realist position remains a bit mystifying to me. If I thought my own ethical views were ultimately just arbitrary preferences, I think that belief would make me feel those preferences less strongly, and care less about whether other people shared them.

2. Deluded Sadist thought experiment

This thought experiment Smart mentions does a good job of drawing out one of the contentious implications of at least the hedonistic variety of utilitarianism:

> Are there pleasurable states of mind towards which we have an unfavourable attitude, even though we disregard their consequences? In order to decide this question let us imagine a universe consisting of one sentient being only, who falsely believes that there are other sentient beings and that they are undergoing exquisite torment. So far from being distressed by the thought, he takes a great delight in these imagined sufferings. Is this better or worse than a universe containing no sentient being at all? Is it worse,

[1] J. J. C. Smart, "An Outline of a System of Utilitarian Ethics," in *Utilitarianism: For and Against* (Cambridge [Eng.]: University Press, 1973), 4–5.

[2] Ibid., 31.

> again, than a universe containing only one sentient being with the same beliefs as before but who sorrows at the imagined tortures of his fellow creatures? I suggest, as against Moore, that the universe containing the deluded sadist is the preferable one.[3]

I think you can even make this thought experiment a little less abstract: many ordinary real people do take pleasure in suffering that they inaccurately believe others are undergoing. Some major religious traditions teach that most humans go to hell forever, and some adherents seem to (or claim to? or try to?) see this as a beautiful fact. If (unrealistically) this delusion had no harmful side effects, would the pleasure it gives those who believe it make it a desirable belief? I'm inclined, uncomfortably, to say yes.

3. Ordering states

The non-utilitarian contributor points out something I hadn't really thought about before:

> Although the non-consequentialist is concerned with right actions - such as the carrying out of promises - he may have no general way of comparing states of affairs from a moral point of view at all. Indeed … the emphasis on the necessary comparability of situations is a peculiar feature of consequentialism in general, and of utilitarianism in particular.[4]

Williams uses this as the basis for some unexpected psychosocial speculation. He thinks consequentialism starts to look attractive when societal changes force people to confront situations that were unthinkable in their traditional worldview—situations where they must choose among unacceptable choices.

> It could be a feature of a man's moral outlook that he regarded certain courses of action as unthinkable, in the sense that he would not entertain the idea of doing them … But, further, he might equally find it unacceptable to consider what to do in certain conceivable situations. Logically, or indeed empirically conceivable they may be, but they are not to him morally conceivable, meaning by that that their occurrence as situations presenting him with a choice would represent not a special problem in his moral world, but something that lay beyond its limits. For him, there are certain situations so monstrous that the idea that the processes of moral rationality could yield an answer in them is insane…

[3] Ibid., 25.

[4] Bernard Williams, "A Critique of Utilitarianism," in *Utilitarianism: For and Against* (Cambridge [Eng.]: University Press, 1973), 88.

> …Rationality he sees as a demand not merely on him, but on the situations in, and about, which he has to think; **unless the environment reveals minimum sanity, it is insanity to carry the decorum of sanity into it**. Consequentialist rationality, however … has no such limitations: making the best of a bad job is one of its maxims, and it will have something to say even on the difference between massacring seven million, and massacring seven million and one.[5]

Mostly I just think the bolded portion is a cute little aphorism. But Williams is making an interesting point that I missed at first. If I understand correctly, his concern is that confrontation with such insane situations in the modern world has reasonably pushed people to give up the notion "that there was nothing which was right whatever the consequences"[6], but that the leap to believing "the different idea that everything depends on consequences"[7] is a rash overreaction.

My reply would be simply that this is not why I'm a consequentialist. I didn't think, *oh no, I need a way to compare all situations*, and latch on to consequentialism because it was the first suggestion I came across. Rather, I thought, *what reasons are there for preferring one action to another*, and consequentialist reasons seemed the most persuasive.

4. Integrity

Williams famously (or at least, famously enough for me to have heard of it before) complains that utilitarianism requires people to compromise their integrity. He gives two thought experiments to support this[8]; the gist is:

1. George must decide whether to take a job assisting a company whose business he believes to be immoral. Utilitarianism requires him to do it, because it would help his family, and because if he refuses then the job will go to someone who does it even more efficiently.
2. Pedro threatens to kill 20 people unless Jim kills one of them. Utilitarianism requires Jim to do it, since one death is clearly better than the same death plus 19 other deaths.

Williams says:

[5]Ibid., 92–93, emphasis added.
[6]Ibid., 93.
[7]Ibid.
[8]Ibid., 98.

> A feature of utilitarianism is that it cuts out a kind of consideration which for some … makes a difference to what they feel about such cases: a consideration involving the idea … that each of us is specially responsible for what *he* does, rather than for what other people do. This is an idea closely connected with the value of integrity.[9]

Williams considers a couple ways a utilitarian might try to argue that George shouldn't take the job and Jim shouldn't kill anyone: "the psychological effect on the agent"[10] and "the precedent effect"[11]. The former, as Williams notes, is most compelling in the George scenario, whereas:

> In Jim's case … his feelings might seem to be of very little weight compared with other things that are at stake. There is a powerful and recognizable appeal that can be made on this point: as that a refusal by Jim to do what he has been invited to do would be a kind of self-indulgent squeamishness.[12]

This gets near the heart of why I don't find the Jim scenario to be a problem for utilitarianism. If I really knew that my choice would make the difference in whether 1 person or 20 people died, and I knowingly chose the route that would lead to 20 deaths in order to avoid myself having a particular causal relation to any of the deaths, it seems like I'm making a self-centered choice. And if caring about other people is one of my core commitments, then making such a self-centered choice feels like a betrayal of my own principles and a violation of my integrity!

Williams thinks this "squeamishness" line of argument only explains why Jim choosing to kill would make sense within a utilitarian worldview; he does not think it is an adequate response to a non-utilitarian who believes that choosing to kill would violate their personal integrity:

> The 'squeamishness' appeal is not an argument which adds in a hitherto neglected consideration. Rather, it is an invitation to consider the situation, and one's own feelings, from a utilitarian point of view.
>
> The reason … one can be unnerved by the suggestion of self-indulgence in going against utilitarian considerations, is not that we are utilitarians who are uncertain what utilitarian value to attach to our moral feelings, but that we are partially at least not utilitarians, and cannot regard out moral feelings merely as objects of utilitarian value.[13]

[9] Ibid., 99.
[10] Ibid., 101.
[11] Ibid., 106.
[12] Ibid., 102.
[13] Ibid., 103.

First of all, yes we *can*, it's just a question of whether we *should*. Part of the point of thought experiments is precisely to invite us to consider things from multiple viewpoints in order to evaluate which viewpoint is most compelling. Secondly, our moral *feelings* are highly affected by our moral *beliefs*, and both may change when we try considering the situation from an impartial perspective.

One last thing I want to say about these thought experiments, which applies widely to ethical thought experiments in general: the discussion don't usually devote much attention to **uncertainty**. But in the real world, the confidence Jim is justified in having that *Pedro will both try and succeed in killing these 20 people if I refuse his request* will be much lower than the confidence he's justified in having that *I will succeed in killing one person if I shoot*. The greater confidence we can have in the immediate effects of our own actions, as opposed to the indirect effects of our choices, may often be a reason to act differently than an omniscient utilitarian would.

5. The best objection

I think the following thought experiment, which I'll refer to as the fragile racists scenario, is Williams's best argument against utilitarianism:

> Suppose that there is in a certain society a racial minority. Considering merely the ordinary interests of the other citizens, as opposed to their sentiments, this minority does no particular harm; we may suppose that it does not confer any very great benefits either. Its presence is in those terms neutral or mildly beneficial. However, the other citizens have such prejudices that they find the sight of this group, even the knowledge of its presence, very disagreeable. Proposals are made for removing in some way this minority. If we assume various quite plausible things (as that programmes to change the majority sentiment are likely to be protracted and ineffective) then even if the removal would be unpleasant for the minority, a utilitarian calculation might well end up favouring this step, especially if the minority were a rather small minority and the majority were very severely prejudiced, that is to say, were made very severely uncomfortable by the presence of the minority.[14]

I think it's worth comparing this to a couple other common thought experiments. Like "Torture vs. Dust Specks",[15] it highlights how utilitarianism can require you to impose a

[14] Ibid., 105.

[15] Eliezer Yudkowsky, "Torture Vs. Dust Specks," accessed September 30, 2023, https://www.lesswrong.com/posts/3wYTFWY3LKQCnAptN/torture-vs-dust-specks.

huge cost on a small group in order to bestow a small benefit on a large group, if the difference in group sizes is big enough. The key extra element in the fragile racists case is the element of injustice: the connection between the cost and the benefit—the fact that the racists' discomfort can only be relieved by harming the minority—exists because of a character flaw in the very people who would receive the benefit. A variant of the utility monster thought experiment[16] shares that element, if we imagine that the monster receives pleasure specifically because it enjoys the suffering of others; but the fragile racists case is less abstract and feels somewhat more like a plausible real-world situation. Those elements of injustice and relative plausibility make the fragile racists thought experiment especially pointed.

Williams brings up this case in order to pose a dilemma for utilitarians: should pleasure/pain that someone experiences due to their own irrational dispositions count when evaluating a state of affairs? (E.g. pleasure from doing something immoral, or pain from being in the presence of a person you are racist against.) If the utilitarian *does* count it, they're led to the abhorrent conclusion in this thought experiment; if they *don't* count it, they can't use the "psychological effect" defense in the George and Jim cases above. I'm not concerned about the George and Jim cases, but my my meta-ethical view (see Sharon Hewitt Rawlette's *The Feeling of Value*)[17] doesn't leave room for considering the causes of a subjective experience when evaluating its intrinsic (as opposed to instrumental) value. So I can't evade Williams's thought experiment by saying the racists' suffering doesn't matter at all.

One solution might be to invoke value lexicality.[18] On this view it's possible to have two kinds of bad things—say X and Y—such that some finite number of Xs is worse than even an infinite number of Ys. This lets you say the dust specks can never outweigh one person's torture; the utility monster's bliss can never outweigh anyone's suffering; and no amount of "severely uncomfortable" people can outweigh the suffering caused by a purge.

But I'm skeptical of value lexicality, and it's not a full solution anyway. You could adjust the thought experiment to specify that the subjective suffering the racists experience is just as intense as the suffering they're planning to inflict on the minority. This variant lacks one

[16]"Utility Monster," in *Wikipedia*, September 23, 2023, https://en.wikipedia.org/w/index.php?title=Utility_monster&oldid=1176763397.

[17]Sharon Hewitt Rawlette, *The Feeling of Value* (King George, Virginia: Dudley & White, 2016).

[18]Magnus Vinding has written several essays on value lexicality, such as Magnus Vinding, "Clarifying Lexical Thresholds," *Center for Reducing Suffering* (blog), 2020, https://centerforreducingsuffering.org/research/clarifying-lexical-thresholds/.

of the strengths of the original thought experiment, because it's a much more unrealistic or pathological scenario. But I think it highlights that the most potent part of the original scenario is the element of injustice, *not* the element of outweighing large harms with many small benefits.

And it's an injustice with a particular structure:

- there's a win-lose scenario (either the racists suffer or the minority does);
- the only thing preventing a win-win scenario is the attitudes of one group (everyone could live happily if the racists weren't racist);
- the group making it a win-lose scenario is in a sense using an implicit threat of self-harm to tilt the cost/benefit analysis in their own favor (the only reason that placating the racists causes less overall suffering than the alternative is that the racists' own brains are going to torment themselves if they don't get their way)

Like an abusive partner threatening suicide unless their lover complies with their demands, the racists are effectively holding themselves hostage in order to blackmail the utilitarian decision-maker into deciding in their favor. In the real world, I think utilitarianism would generally tell us to reject such threats. The "precedent effect" that Williams dismisses in the George and Jim cases surely does apply in this case: appeasing the racists would incentivize everyone in the future who has the slightest racist inclination (or any other arbitrary and harmful inclination) to work themselves into as much of an emotional frenzy over it as they can. That sort of downward spiral is not at all conducive to the utilitarian goal of maximizing net happiness.

Of course, you could modify the thought experiment again and declare we magically know that this is the last time in history that any form of prejudice will arise, so the precedent becomes irrelevant. You can add as many such qualifications to the scenario as necessary to keep forcing attention back to the core issue: whether there are cases where utilitarianism would use the interests of the oppressor to justify oppression. But all these qualifications make the cases drastically less realistic. Unrealistic cases are still important, but have limitations. Our ethical intuitions are informed by what we know about the real world; if a theory (utilitarianism) gives a different answer than our intuitions on a question that requires us to imagine a radically different reality from the one we live in, it becomes hard to tell whether the issue is with the theory, or if it's just difficult for us to prevent our background assumptions about the real world from leaking into our judgments about the imaginary one.

(There are other unrealistic aspects to the fragile racists scenario. I'm skeptical that a typical racist experiences much suffering when contemplating the targets of their racism. And the idea that whatever suffering they do experience would stop if the hated group were removed seems highly implausible; the hatred would just be transferred to another set of targets. Williams wants us to assume "that programmes to change the majority sentiment are likely to be protracted and ineffective", but history suggests sentiments can and do change. Even if the process is "protracted", the long-term benefits of people learning to get along would be enormous compared to dealing with every conflict by crushing one side.)