The Abusability Objection

Table of Contents

Introduction

As a consequentialist theory, utilitarianism directs us to promote good outcomes. When we can't be certain of the consequences of our actions, it tells us to promote *expected value*. Because it gives no intrinsic weight to commonsense constraints or rights, some worry that utilitarian ethics is too easily abusable, allowing people to construct false justifications for horrifically harmful actions. Blindly following the results of their expected value (mis-)calculations might lead even well-meaning individuals into disaster. As a result, many have claimed that utilitarianism is *self-effacing*, or recommends against its own acceptance.

To evaluate this objection, we must clarify two things. First: what practical guidance does utilitarianism actually offer? Expected value provides a *criterion* against which actions can be evaluated, rather than a *decision procedure* to use in all circumstances. This distinction is crucial for understanding the relation between utilitarian theory and practice, as it turns out that utilitarians should still give significant weight to commonsense constraints on instrumental grounds.

Second: what (if anything) is objectionable about self-effacing moral theories? As we'll see, there are strong reasons to think that all reasonable moral views are at least sometimes self-effacing. So a view's self-effacement is not evidence that it is false.

How Utilitarianism Could Be Misused

It's a common trope that only villains endorse the consequentialist principle that "the ends justify the means". The idea that it's okay to trample human rights for the "greater good" is something we

hear from the likes of Thanos, not from the good guys.[1] And there are reasons why we tell this kind of morality tale: though none of them were plausibly utilitarians,[2] the real-world examples of Hitler, Stalin, and Mao demonstrate the danger of imposing a totalizing ideology in a way that's completely unhinged from ordinary moral constraints.

This is all to say that ordinary moral constraints have immense instrumental value, and we generally expect wholesale disregard of them to result in disaster. It's plainly contrary to utilitarian principles to disregard immense instrumental value. To do great harm while falsely claiming the mantle of the "greater good" would be a clear misuse of utilitarian theory, and one that it's worth guarding against. Utilitarians thus have strong reason to agree that we should regard a person's villainous-seeming claims about the "greater good" with sharp suspicion.

Utilitarianism implies that if an act *really were* to produce the best consequences for overall well-being, then it would be worth it. But we should be suspicious of the further claim that villainous means actually serve this end in practice. Historically, such claims have most often proven to be disastrously false.

Is Utilitarianism Self-Effacing?

As explained in Chapter 6: Utilitarianism and Practical Ethics, a plausible utilitarian decision procedure might direct us to:

1. Pursue any "low-hanging fruit" for effectively helping others while avoiding harm,
2. Inculcate virtues for real-world utilitarians (including respect for commonsense moral norms), and
3. In a calm moment, reflect on how we could better prioritize and allocate our moral efforts, including by seeking out expert cost-benefit analyses and other evidence to better inform our overall judgments of expected value.

Notably, whatever decision procedure utilitarianism *actually* recommends can't predictably yield worse outcomes than an available alternative. For if it did, utilitarianism would instead recommend that better alternative. Agents who genuinely do as utilitarianism recommends will, by definition, do better (in expectation) than if they did otherwise. The same cannot be said of non-consequentialist theories, which risk sometimes *actually justifying* doing (or allowing) more harm than good.[3]

But a residual objection remains, for two reasons. First, sincerely *trying* to follow a moral theory doesn't mean that you'll *succeed* in doing as it recommends; inept agents, inspired by utilitarianism, could still do great harm. Second, not all agents are morally sincere. Some may intentionally do harm while invoking the "greater good" to rationalize their actions. Accordingly,

critics may worry that widespread acceptance of utilitarian justifications would make it easier for bad actors to get away with committing atrocities.[4]

Neither of these residual objections speaks to the *truth* of utilitarianism. Sometimes true claims can be misunderstood or misused in harmful ways.[5] The question is what should be done about this risk.

One possibility would be to embrace some non-utilitarian moral theory as a "noble lie".[6] Many philosophers have speculated that consequentialist ethics may be *self-effacing*, and direct us to believe some other theory instead.[7] For example, one might speculate that people have a psychological tendency to underweight "merely" instrumental considerations, and so we would be better protected against atrocities if people generally believed human rights to have *non-instrumental* moral significance. But in light of the general instrumental value of truth-seeking, it's worth first checking whether the risks can be mitigated without resorting to deception.

A more honest option would be to make clear the utilitarian case for moral constraints in practice, as we've done throughout this text.[8] If commonsense norms have high instrumental value, and explicit calculations to the contrary are more likely to be mistaken than correct, then real-life violations of commonsense norms *cannot easily be justified on utilitarian grounds*.[9] Crucially, if more people come to appreciate this fact, then it will be harder for bad actors to abuse utilitarian ideas. Interestingly, this suggests that the abusability objection may *itself* be self-effacing, as explained in the following note.[10]

To this end, it's worth noting that utilitarian underpinnings can justify "moral rules" in different senses of the term. Most obviously, utilitarianism can support treating rules as *heuristics*, or "rules of thumb", for more reliably identifying the best option and avoiding harm. Heuristics are typically understood as overridable, allowing for exceptions when one can secure more reliable information without undue cost. Utilitarianism can also justify *policies*, such as committing to follow a simple rule without exceptions, if adopting such a policy would prove better than failing to do so. (Such a policy might sometimes result in one acting suboptimally, but it could still be worth adopting if any alternative policy, including a policy of trying to act upon expected value calculations, would realistically result in even *worse* suboptimality.)[11] An important example might be the exceptionless *enforcement* of (social and legal) sanctions against those who violate human rights or other generally good rules.

Consider a Ticking Time Bomb scenario, where one supposedly can only prevent a nuclear detonation by illegally torturing a suspect. If millions of lives are on the line, the argument goes, we should accept that torture could be justified. But given the risk of abuse, we might also want anyone who commits torture to suffer strict legal sanctions. If millions of lives are really on the line, the agent should be willing to go to jail. If someone wants to torture others, but isn't willing to go to jail for it, this raises serious questions about their moral integrity—and the likely

consequences of letting them run loose. Accordingly, there's no inconsistency in utilitarians holding *both* that (i) violating human rights could be justified in the most extreme circumstances, and yet (ii) anyone who violates human rights should be strictly held to account.

In these ways, utilitarianism can go a fair way towards accommodating commonsense norms, mitigating the risk of abuse, without resorting to full-blown moral deception or self-effacement.

Are Self-Effacing Theories Objectionable?

We should generally be averse to lying, including about the moral truth itself. But it's ultimately an empirical question what the consequences would be of any particular individual coming to believe any given moral theory.[12] In cases where the results of true beliefs would be bad, we may have practical reasons not to draw attention to those truths, or—in extreme cases—even to outright lie.[13] But that doesn't make the truth inherently objectionable; the problem instead lies with those who would misunderstand or otherwise mis-use it.[14]

Every sensible (non-absolutist) moral theory is *possibly* self-effacing: if an evil demon will torture everyone for eternity unless you agree to be brainwashed into having false moral views, you surely ought to agree to the brainwashing. Moreover, ethical theory is generally regarded as non-contingent: whichever moral theory is true, this isn't an accident—the same fundamental moral theory must be true in all possible worlds.[15] That means that the actually-correct moral theory, whichever one it is, remains true in some possible worlds where it's self-effacing. Perhaps our world is one of them, or perhaps not. The truth of the matter does not turn on this, either way. So a theory's being self-effacing is irrelevant to philosophical assessments of its correctness.

Conclusion

To understand utilitarianism, one must understand the distinction between the theory's *criterion* and recommended *decision procedures*. Canonical statements of utilitarianism state its criterion or moral goal: what makes an act worth doing is that it promotes (expected) value or well-being. When some imagine that this entails constantly calculating utilities, they are making a mistake. We cannot immediately "read off" a decision procedure from the theory alone, for how to pursue utilitarian goals in an instrumentally rational way depends on contingent facts about our cognitive capabilities and broader psychology.

Sometimes a little knowledge can be a dangerous thing, and this seems plausibly true of utilitarianism. Someone who endorses the utilitarian criterion without thinking clearly about our epistemic limitations might end up acting in ways that are (predictably) very bad by utilitarian lights. In theory, one might try to avoid this problem either by depriving people of *any* knowledge of utilitarianism, or by striving to convey the *full* picture. In practice, there are obvious reasons to prefer the latter, as true beliefs—especially about morality—can generally be expected to guide

people towards better actions. So we can best protect against the risk of abuse by being clear that utilitarianism does *not* easily justify atrocities.

Still, at the end of the day there's no guarantee that true beliefs will be socially optimal. It's always possible that any reasonable, non-absolutist moral theory may turn out to be self-effacing. This possibility is not an objection to those views.

Other Objections to Utilitarianism | Next Chapter: Agindo de Acordo com o Utilitarismo

---

Como Citar esta Página

Chappell, R.Y. (2023). The Abusability Objection. In R.Y. Chappell, D. Meissner e W. MacAskill (eds.), *An Introduction to Utilitarianism*, <https://www.utilitarisme.net/objections-to-utilitarianism/abusability>, acessado em 23/05/2025.

Resources and Further Reading

- Allan Gibbard (1984). Utilitarianism and Human Rights. *Social Philosophy and Policy*, 1(2): 92–102.
- R.M. Hare (1981). *Moral Thinking*. Oxford University Press.
- Katarzyna de Lazari-Radek & Peter Singer (2010). Secrecy in Consequentialism: A defence of esoteric morality. *Ratio*, 23(1): 34–58.
- J.L. Mackie (1985). Rights, Utility, and Universalization. In R.G. Frey (ed.) *Utility and Rights*. Basil Blackwell.
- Derek Parfit (1984). *Reasons and Persons*, Part One: Self-Defeating Theories. Clarendon Press.
- Philip Pettit & Geoffrey Brennan (1986). Restrictive Consequentialism. *Australasian Journal of Philosophy*, 64(4): 438–455.
- Bernard Williams (1973). A Critique of Utilitarianism. In J.J.C Smart & Bernard Williams, *Utilitarianism: For and Against*. Cambridge University Press.

---

1. It's also notable that superheroes are depicted as putting so little effort into cause prioritization, often fighting local crime when they could (more helpfully, but far less

dramatically) use their powers in more scalable ways to do good on a global scale—as this SMBC comic satirically illustrates. ↩

2. In particular, it doesn't seem plausible to suppose that they were primarily driven by impartial beneficence. ↩

3. That is, less demanding views may justify selfish (in)actions, such as neglecting the needs of the global poor, non-human animals, and future generations. So it's worth considering how competing views fare against their own versions of the abusability objection. ↩

4. Though again, it's interesting to consider how competing views fare against this objection. Many are so vague that they leave plenty of room for self-serving interpretations, and so would also seem easily exploitable by bad actors. ↩

5. As John Stuart Mill writes in Chapter 2 of *Utilitarianism*, "There is no difficulty in proving any ethical standard whatever to work ill, if we suppose universal idiocy to be conjoined with it". ↩

6. Or perhaps as a simplified "lie-to-children". ↩

7. Most famously, Bernard Williams wrote that "utilitarianism's fate is to usher itself from the scene." (1973, p.134). The idea of "esoteric morality" is found in Henry Sidgwick's (1874) *The Methods of Ethics*, and was subsequently criticized (for its elitist vibes) as "government house utilitarianism". But only implausibly absolutist views can strictly rule out the possibility that esotericism may sometimes be justified. For broader discussion, see de Lazari-Radek & Singer (2010) Secrecy in Consequentialism: A defence of esoteric morality. *Ratio*, 23(1): 34–58. ↩

8. For a famous historical example, see John Stuart Mill's (1859) *On Liberty*, which argues for the utilitarian importance of respecting others' freedom. ↩

9. Moral uncertainty is also relevant here, as one needn't have *most* confidence in deontological views for them to still exert an additional tempering effect. ↩

10. In spreading the false idea that utilitarianism easily justifies abuses, proponents of the abusability objection are, ironically enough, contributing to the very problem that they worry about. Given the strong theoretical case for utilitarianism, it's inevitable that many reflective people will be drawn to the view. If you start telling them that their view justifies real-life atrocities, some of them might believe you. That would be bad, because the claim is both harmful and false. As a result, we do better to promote a more sophisticated understanding of the relation between utilitarian theory and practice—emphasizing the value of generally-reliable rules and heuristics, and the unreliability of crude calculations when these conflict with more-reliable heuristics. ↩

11. For discussion of related issues, see Part One of Derek Parfit (1984). *Reasons and Persons*. ↩

12. Whether a certain belief has good or bad effects may vary across different individuals and contexts. There may be good reasons not to teach kindergarteners about the possibility of rare exceptions to moral rules, for example. ↩

13. Compare the case of the Murderer at the Door, inquiring as to the whereabouts of their intended victim. Kant notoriously denied that lying is ever permissible, but few have found his response to this case remotely plausible. ↩

14. That is, if we must withhold the truth—from ourselves or others—that may be a reason to think less of the relevant people, rather than to think poorly of the relevant true claim. ↩

15. When philosophers speak of "possible worlds", they just mean a possible *scenario*, or *way the world could have been*. A proposition *p* is said to be "true in" a possible world *w* if and only if, *were w to be actual, p would be true*. The (non-contingent) fundamental ethical theory combines with (contingent) facts about a world to yield the (contingent) applied moral claims or verdicts that are true in a world. ↩