A Gödelesque argument against being a pure consequentialist

By Magnus Vinding

First published: September 19, 2023. Last updated: October 2025.

My primary aim in this essay is to present an argument against being a pure consequentialist. Somewhat surprisingly, this argument appears to find added support from consequentialist considerations.

By a "pure consequentialist", I mean someone who only assigns normative validity (or normative weight, plausibility, etc.) to **pure consequentialism**: the view that what is morally right in any domain is whatever has the best consequences (cf. Pettit & Smith, 2000, p. 121). This view is often referred to as "global consequentialism", or simply "consequentialism" (see e.g. Parfit, 1984, pp. 24-25).[1]

Note that my definition of pure consequentialism is compatible with indirect efforts to bring about the best consequences, such as via prudent decision procedures that are indirectly optimized for generating the best consequences. In other words, the argument I will present is not merely an argument against naive consequentialist decision procedures, but rather an argument against being a pure consequentialist per se.

---

[1] It is worth distinguishing normative validity (or correctness, plausibility, etc.) from moral rightness criteria (e.g. a criterion like "what is morally right in any domain is whatever has the best consequences"). After all, a given moral rightness criterion might not be normatively valid (or correct, plausible, etc.). For example, the rightness criterion "what is morally right is whatever has the *worst* consequences" is obviously not normatively plausible.
On a related note, one may draw a distinction between the normative validity of epistemic views (or criteria, principles, etc.) versus the normative validity of moral views (or criteria, principles, etc.), and one may assign normative validity to principles in each domain separately. But the key point here is that pure consequentialists (as defined here) ultimately only assign normative validity to their consequentialist moral view, with everything else being subservient to and derived from that as far as normative validity is concerned.

The basic argument: Pure consequentialism is internally consistent or inconsistent, but implausible to endorse in either case

There are various ways to formulate the basic argument that I will present against being a pure consequentialist, but what follows is a simple version.

Pure consequentialism is either internally inconsistent or internally consistent. That is, assuming the normative validity or correctness of pure consequentialism will either imply that pure consequentialism is not normatively valid or correct (in which case it is internally inconsistent), or assuming its normative validity will imply no such contradiction (in which case it is internally consistent). I will focus on each of these options in turn.[2]

If it is internally inconsistent:

If pure consequentialism is internally inconsistent, then we have plausible reason not to adopt this criterion, based on the premise that the internal inconsistency of a given criterion is a plausible reason not to adopt it. In other words, if assuming the validity of a given normative criterion itself implies that this criterion is not valid, this is a reason not to consider it valid.

If it is internally consistent:

Conversely, let us grant that pure consequentialism is internally consistent, meaning that assuming its normative validity does not contradict its normative validity. A relevant question is then: does a pure consequentialist have even a logical possibility of questioning pure consequentialism in a way that might refute it (provided that this person reasons coherently)?

The answer is "no". That is, by the assumed consistency of pure consequentialism, someone who only assigns normative validity to pure consequentialism cannot derive any normative implications that contradict its validity, and hence a pure consequentialist (who reasons

[2] The argument below can be phrased in terms of normative "validity", "weight", "plausibility", "legitimacy", "reasonableness", "correctness", "soundness", or other similar terms that may reflect different views of the nature of normative plausibility and soundness (e.g. objectivist vs. subjectivist views). I will generally phrase the argument in terms of "validity" without loss of generality, though I will occasionally use other terms to vary the language. Sometimes I will simply write "adopt", "embrace", or "endorse" (e.g. endorse a given value) rather than "assign normative validity to".

coherently) cannot possibly refute the normative validity of their own view. Such a refutation can only be made from a position that gives normative weight to independent values, such as values of openly questioning and exploring ideas — values that are not tied to, and not ultimately forced to come into agreement with, the normative validity of pure consequentialism.

The problem here, I submit, is that this internally consistent stance seems implausibly dogmatic and preconceived, as it effectively rules out that any argument could ever refute the normative validity of pure consequentialism. (To clarify the structure of my argument: I here take it as a sound premise that it is implausibly dogmatic to adopt a position that does not entail even the logical possibility of refuting the normative validity of pure consequentialism. In other words, it is implausible to categorically rule out the possibility that pure consequentialism might not be normatively correct or valid.)

To summarize: being a pure consequentialist is implausible because it is either an internally inconsistent position, or it implies an implausibly dogmatic stance toward pure consequentialism itself. We thus have reason not to be pure consequentialists.

Objection: A pure consequentialist could potentially refute their own view based on rule consequentialism

One might object that a pure consequentialist could have the logical possibility of refuting their own view if they endorse a form of rule consequentialism that says that we should accept and adhere to rules and values that have the best consequences. Specifically, if it has optimal consequences to endorse independent values that enable us to refute the normative validity of pure consequentialism, then this kind of rule consequentialism would prescribe that we endorse such independent values.

The problem with this objection, however, is that the (initial) pure consequentialist thereby ends up no longer being a pure consequentialist, as they ultimately come to give normative weight to values that are distinct from pure consequentialism.

Consequentialist reasons not to be a pure consequentialist

In this section, I will briefly outline some tentative reasons as to why being a pure consequentialist might overall lead to worse outcomes in consequentialist terms. Note that I am not saying that these reasons clearly establish that being a pure consequentialist would lead to worse outcomes, but merely that the following are plausible reasons in favor of that claim.[3]

One consequentialist reason not to be a pure consequentialist is that it can give a bad impression in various ways. For example, if pure consequentialists cannot coherently question their own moral view, this might make them seem alarmingly dogmatic, which may in turn reflect badly on them in wide-ranging ways. For instance, people might reasonably wonder: "If they cannot genuinely question their own moral view, how else is their thinking overly rigid and flawed?"

A related reason not to be a pure consequentialist is that people would be justified in not trusting pure consequentialists in some significant regards, especially when it comes to engaging in truly open-ended scrutiny of their own moral view. In other words, it is not only that a pure consequentialist *seemingly* cannot be trusted to engage in such open-ended scrutiny, and hence that they merely give a bad *impression* in this regard, but rather that, as we have seen, a pure consequentialist indeed has no logical possibility of refuting their own moral view (provided that it is consistent). The problem that other people observe and take issue with goes deeper than mere appearances, and it plausibly has suboptimal consequences to adopt a position that gives other people justified reason to conclude that one is too dogmatic to be able to engage in open-ended scrutiny of one's own moral view.

A third consequentialist reason against being a pure consequentialist is that the concern expressed in the question above — "how else is their thinking overly rigid and flawed?" — might likewise go deeper than mere appearances and imagined concerns. That is, giving exclusive normative weight to a single rightness criterion that applies exhaustively to everything

[3] To be clear, saying that we could bring about better outcomes by not being pure consequentialists is not the same as saying that the rightness criterion of pure consequentialism is internally inconsistent or normatively invalid. After all, it could in principle be the case that pure consequentialism is internally consistent and normatively valid *and* that we can create better outcomes by not being pure consequentialists. So these issues should not be conflated. This section is not about the normative validity or plausibility of pure consequentialism, but rather about whether pure consequentialism itself recommends that we be pure consequentialists. The reasons reviewed here are thus a separate class of reasons that may speak against being a pure consequentialist.

may affect various aspects of a pure consequentialist's thinking, such that their thinking becomes more rigid and narrow than would be ideal for bringing about the best consequences. For example, it is conceivable that embracing independent values of open and honest exploration would ultimately lead to better and more flexible thinking than would values of exploration that are wholly tied to pure consequentialism.

Note that the consequentialist considerations listed above do not suggest that we should simply *claim* or *appear* not to be pure consequentialists while secretly endorsing pure consequentialism. Such an approach might superficially address the first point regarding bad appearances, but only by doing something that involves, or at least risks, even worse appearances and greater justified distrust, namely to lie or to otherwise be deceptive (which there are many consequentialist arguments against, e.g. Carson, 2010, ch. 4; Harris, 2011; Tomasik, 2013). Moreover, adopting such a deceptive stance does nothing to address the point that being a pure consequentialist may inadvertently harm one's thinking and outlook (since one would still secretly be a pure consequentialist). On the contrary, this deceptive stance would likely add greater such harms, since being dishonest and deceptive appears to have detrimental effects on our thinking and our outlook more generally (Engelmann & Fehr, 2016).

Thus, rather than supporting a deceptive approach, the consequentialist considerations listed above seem to speak in favor of *genuinely* and *honestly* not being a pure consequentialist if one is to bring about the best consequences. These considerations give us further reasons not to be pure consequentialists, in addition to the implausible degree of dogmatism it entails.

What about other views that assign exclusive validity to a single normative criterion?

My focus in this essay has been on pure consequentialism, yet I submit that the argument I have presented applies to any stance that assigns exclusive normative validity to a single global normative criterion.

That is, for any such normative criterion, the criterion will either be internally inconsistent, which is a plausible reason not to endorse it, or someone who assigns exclusive normative

validity to that criterion will not have a logical possibility of refuting its normative validity, which seems unduly dogmatic and thus also a plausible reason not to endorse it.

After all, for any criterion we may propose as a single global normative criterion, it seems that we can always legitimately ask and be unsure about whether it captures (all of) what matters. And again, the only way in which we can conduct a truly open-ended exploration of the possibility that a given criterion might not capture all that matters — assuming that it is internally consistent — is from the vantage point of independent values that are not tied to that criterion. We cannot measure a yardstick with itself.

In short, the argument I have presented here can be generalized into an argument against giving exclusive normative validity to any single normative criterion.[4]

Note that the argument above does *not* claim that no single normative criterion could be perfectly correct or valid. Rather, the core of the argument is that *we*, as fallible creatures, should not be confident that we have identified such a perfectly correct or valid criterion. Indeed, even if we could be sure that there is a single normative criterion that is correct, we should still not be confident that our specific conception of that normative criterion is correct. We should have at least some uncertainty about that.

---

[4] But what if the single normative criterion is broad and open-ended, such as: "we ought to do what seems most reasonable, or most justified, all things considered" — couldn't that be a plausible singular criterion? I see two broad replies. First, although a broad criterion like this might indeed be the most plausible single criterion we can come up with, it seems that we still want to be able to question even this broad criterion (and others like it). For example, isn't there a risk that even a broad criterion like this could be missing something important? Isn't it worth being able to at least question the criterion? Second, if the criterion is construed in an open-ended way such that it allows us to question any *specific* conception of what we ought to do, then it is arguably less vulnerable to the charge of dogmatism. Yet, in that case, it may be objected that the criterion is unduly vague and that its status as a singular normative criterion is questionable. For example, if a phrase like "most reasonable all things considered" implicitly covers an iterative process in which we can continually modify our conception of "reasonable" along with any other key term that defines our core normative criterion, then arguably our core normative criterion has a strong pluralism of possible interpretations baked into it — a plurality of specific criteria that can compete with, question, and refine each other.

What about any set of moral values?

One might wonder whether the argument made above couldn't be applied to any set of values, regardless of whether they are based on a single global criterion or not. In other words, couldn't the argument above be used to show that there are no values whatsoever that are plausible to adopt?

The argument would in that case say that for any possible set of values that we consider normatively valid, the values will either be internally inconsistent, which is a strong reason not to adopt them, or someone who endorses the values in question will not have the logical possibility of refuting the normative validity of these values, which seems implausibly dogmatic.

On its face, this can seem like a sound argument. Yet where it fails, I will argue, is in the assumption that internal inconsistency is necessarily a strong reason not to adopt a given set of values. To be more precise, the internal inconsistency of a set of values is indeed a strong reason not to *fully* and *absolutely* endorse each of its constituent values; yet it is arguably not a strong reason against endorsing the values in a way that assigns them *partial* plausibility and applicability. In this way, the potential for tensions and even contradictions to emerge between different values remains open, but it is a kind of tension that may allow us to continually negotiate and refine the appropriate range of plausibility and applicability for different values within a broader set of values.

Note how this reply does not work for views that assign exclusive validity to a single global normative criterion: that a global normative criterion contradicts itself is (likewise) a strong reason not to fully endorse this criterion, and to instead (at most) assign it less than full plausibility and applicability. However, if we only give the criterion partial plausibility and applicability, we no longer have a single fully valid criterion that applies exhaustively to everything.

The argument presented here effectively reverses a common intuition in favor of adopting a single global and internally consistent normative criterion, namely that it is ideal to have a view that is simple and which allows for no contradictions. Yet the possibility of genuine (partial) tensions and contradictions in our normative views is arguably desirable, as that is what affords us the possibility to refute and fundamentally refine our values. Again, we cannot measure a yardstick with itself, and hence we need other yardsticks to independently assess the validity and

soundness of any given yardstick. Similarly, we need a multitude of yardsticks to continually hone and develop our yardsticks: a single tool cannot sharpen itself, but a collection of tools can mutually sharpen and refine each other.

Thus, fully adopting a moral view of perfect simplicity and consistency does not seem preferable overall, especially considering our lack of omniscience and the potential fallibility of our views.[5]

The practical relevance of openness about our current moral values

A key premise in my argument against being a pure consequentialist is that it is implausible not to maintain a genuine openness with respect to our current moral values, in the sense of allowing ourselves the possibility of refuting and updating these values. In this section, I will briefly say a bit about the practical relevance of maintaining such openness.

It is worth noting that the plausibility of doubting our current conception of moral values generalizes beyond values with a single normative criterion. That is, it seems that we can always legitimately ask and be unsure about whether our current conception of moral values captures (all of) what matters morally, regardless of whether our values are based on a single normative criterion or not.

How is such openness about our current moral values practically relevant? One way it is relevant is that there is a tension in terms of how we steer by our moral values. On the one hand, we have our current moral values, and we could go ahead and act on those values. On the other hand, there is the realistic possibility that we may refine our moral values, and we can hardly say in advance how different such more refined values might be in terms of their practical implications. (Related ideas are explored in Callard, 2018.)

There is thus a real and practically relevant tradeoff in terms of how strongly we should bet on our current conception of moral values versus the potentially more refined conception that we

[5] The themes discussed here bear similarities to discussions about epistemological holism, the underdetermination of scientific theories, and the Duhem–Quine problem concerning the ambiguity of evidence. For example, when calibrating scientific instruments, how can we determine which ones are accurate given that such instruments are all we have to make measurements with? Arguably, we have no better method than holistic triangulation using a multitude of tools. My claim is that the same applies to normative views and principles: it is implausible to endorse a (supposedly) perfect yardstick that is beyond all doubt.

might endorse in light of further reflection. And there is arguably no straightforward way to resolve this tradeoff based on our current values, as that would seem to beg the question in their favor.

To give an example of "questioning our conception of moral values" that is relevant to consequentialism: imagine that we endorse moral values that give great importance to bringing about the best consequences. A key question is then what we understand by "consequences". The meaning of this term might seem obvious, but the truth is that it is at the same time both rather underspecified and quite loaded with implicit ontology. For example, are consequences restricted to effects in our future light cone? Or might consequences also include events that take place elsewhere, such as in parts of our local universe that exist outside of our future light cone, or even in hypothetical parallel universes (cf. evidential cooperation in large worlds)?[6]

These questions hint at the practical importance of not dogmatically locking ourselves into our current conception of values such that we evaluate everything entirely by their lights, and in effect become unable to think outside of them. After all, our current values may be underspecified, and perhaps they will even turn out to be deeply implausible in light of a fuller understanding of the world, or in light of deeper reflections on ethics that are not narrowly tied to our current outlook.[7]

A key practical recommendation may thus be that we generally try to steer by our current values — the conception of values that we currently find most plausible — while also being open to, and even seeking out, further refinements of our values.

---

[6] One could define consequentialism in terms of optimizing "consequences" in the sense that will ultimately prove most plausible. Yet a problem with this stance is that it relies on an extremely vague notion of "consequences" — a notion so vague that the consequentialism in question could potentially end up being a non-consequentialist view by current standard definitions. Hence, this kind of vaguely defined consequentialism does not give us much of a yardstick to navigate by; it gives us an extremely fuzzy yardstick. It seems worth being aware of this fuzziness of consequentialism, lest we mistake it for a clear and crisp view. (These points echo those of Hempel's dilemma for physicalism, which roughly says that physics-based definitions of physicalism are either false or unduly vague.)

[7] To elaborate further, one of the complications we face is that our moral values are inevitably based on our ontological views, at least in part. And since our ontological views are uncertain — for example, we cannot be completely certain about what kinds of value entities exist or might exist in the world — our ontological uncertainty ultimately shades into moral uncertainty as well. In other words, our ontological uncertainty implies at least some degree of moral uncertainty, and perhaps even a very large degree.

My own view

My own moral view may roughly be described as consequentialist, so my argument against being a pure consequentialist might leave it unclear what exactly my view is. This section is an attempt to briefly clarify that.

I think that consequences are generally the most important and most plausible moral focus, with the reduction of extreme suffering as the overriding priority. And I have a high level of confidence in this view.[8] Yet I do not endorse the claim that whether something brings about the best consequences (according to my current conception of "best consequences") is the sole normative criterion for evaluating absolutely everything, with zero room for doubt in this criterion. One reason for this is that I also, given my limited and fallible perspective, endorse independent values of open and honest exploration (i.e. values that are not purely grounded in a consequentialist framework). As I have tried to argue, it seems implausible to endorse pure consequentialism at the expense of open exploration to such an extent that we cannot coherently question pure consequentialism.

Another reason I am not a pure consequentialist is that my current conception of "consequences" seems too underspecified and too tied to doubtable ontological assumptions for me to put complete stock in this provisional conception (as argued in the previous section).

Thus, I am not a pure consequentialist, but I think it would be fair to call my view quasi-consequentialist.[9]

---

[8] An analogy might be how we can have a high level of confidence in our best scientific theories and observations, while still in principle being open to questioning even the best-supported theories. For example, saying that we are in principle open to questioning that the Earth is roughly spherical and revolves around the sun need not imply a low level of confidence in these statements. An illustrative metaphor for such robust yet revisable confidence may be a solid diamond structure: it is not unbreakable or immune to revision in principle, yet its core is nonetheless extremely robust. That is roughly how I think about my own fundamental values.

[9] It is worth clarifying that, contrary to what one might hope, this quasi-consequentialist view is not a license for lax ethical standards, nor is it an argument for evaluative optimism over evaluative pessimism. After all, independent values of free and open exploration could well, on reflection, lead us to endorse *higher* ethical standards and *greater* evaluative pessimism compared to what we currently endorse. In short, values of open exploration do not push us in any particular direction on these issues a priori.

Conclusion

I have argued that we have plausible reasons not to be pure consequentialists, and I have suggested that even consequentialist considerations seem to support this claim.

More broadly, I have argued that we have reason to maintain a certain level of openness toward our current moral values, whatever those values may be, such that we avoid locking ourselves into a suboptimal conception of moral values.[10]

[10] Thanks to Teo Ajantaival, Michael St. Jules, and Simon Knutsson for helpful comments.