# Don't Go Chasing Waterfalls: Against Hayward's "Utility Cascades"

Ryan Doody

*Brown University*

Forthcoming in *Utilitas*

**Abstract**

In "Utility Cascades" (*Analysis*, 2020, 80(3): 433-42), Max Khan Hayward argues that act-utilitarians should sometimes either ignore evidence about the effectiveness of their actions or fail to apportion their support to an action's effectiveness. His conclusions are said to have particular significance for the effective altruism movement, which centers seeking and being guided by evidence. Hayward's argument is that act-utilitarians are vulnerable to succumbing to 'utility cascades', that these cascades function to frustrate the ultimate goals of act-utilitarians, and that one apposite way to avoid them is by 'ostriching': ignoring relevant evidence. If true, this conclusion would have remarkable consequences for act-utilitarianism and the effective altruism movement. However, Hayward is mistaken—albeit in an interesting way and with broader significance for moral philosophy. His argument trades on a subtle mischaracterization of act-utilitarianism. Act-utilitarians are not especially vulnerable to utility cascades (or at least not objectionably so), and they shouldn't ostrich.

## 1 Introduction

In "Utility Cascades", Max Khan Hayward raises a purported problem for act-utilitarians (and their effective altruist brethren): namely, that they are vulnerable to *utility cascades*, which "occur when ongoing rational updating of judgements concerning the effectiveness of an intervention causes a utilitarian to push a situation further and further away from the antecedently optimific outcome," (433). Hayward argues that, when facing a utility cascade, utilitarian agents should either ignore evidence about the effectiveness of their interventions ("to ostrich") or not apportion their support of these interventions to their effectiveness.[1]

[1] The argument is something of a *tu quoque* response to (Doody, ms), which criticizes a class on non-utilitarian views on the grounds that they, in some case, recommend avoiding relevant

If true, this would be a remarkable finding—and one with far-reaching import for those who wish to live by the light of act-utilitarianism, including, as Hayward notes, members of the effective altruism movement. But, in what follows, I will argue that Hayward's paper does not establish either of these claims, because, in principle, the sorts of act-utilitarians with which he is concerned will either (i) avoid the utility cascade or (ii) fail to avoid it but in an unobjectionable fashion. The night may be dark and full of terrors for diligent act-utilitarian effective altruists, but threats of utility cascades need not crowd out nightmares of the coming robot apocalypse.

## 2 Hayward's Target

Hayward's targets are act-utilitarians of a particular stripe: let's call them *expected utility maximizing act-utilitarians*. Traditionally, act-utilitarianism is the view that one ought to take the available action that maximizes utility. Because we are rarely ever in a position to know which action *actually* maximizes utility, act-utilitarianism is rarely action-guiding—which is fine; the view is meant to provide a criterion of rightness, not a decision-procedure (Railton, 1984). That said, act-utilitarianism is typically supplemented with a *subjective* criterion of rightness as well: one that says what one ought to do in light of one's beliefs about the consequences of one's actions.[2] One such proposal involves evaluating actions in terms of their *expected utility*. An action's *expected utility* is the weighted sum of the utilities of its potential outcomes, where the weights correspond to one's rational credence that that outcome will result.[3] On these views, what you *objectively*

information—even when doing so is guaranteed to make everyone worse off. This criticism loses much of its bite if Hayward is correct that, in analogous cases, utilitarians, too, should avoid receiving relevant information.

[2] An early version of the distinction between 'objective rightness' and 'subjective rightness' can be found in Sidgwick (1907: 206-8). The distinction is now widely (although not universally) accepted: e.g., Brandt (1959: 381-5), Gibbard (1990: 42-3), Mulgan (2001: 42), Oddie and Menzies (1992: 512), Smart (1973: 46-7), Timmons (2002: 124), Zimmerman (1996: 10-20). One central motivation for the distinction is to account for cases in which we are torn between two conflicting ways of evaluating an agent's actions, where one of those two ways seems to track praise and blame. For some, though, the distinction is also thought to help address worries concerning a moral theory's action-guidingness. (For more discussion of these issues see Feldman (2006) and Smith (2010).)

Different accounts of 'subjective rightness' can be given: in terms of, e.g., whether an act is *likely* to be objectively right (Hospers, 1961; Russell, 1910; Smart, 1973), or whether an act *would* be objectively right if the world were as the agent believes it to be (Brandt, 1959; Broad, 1985), or whether an act has highest *expected* value (Parfit, 1984; Timmons, 2002). Furthermore, some define 'subjective rightness' in terms of what the agent *should* believe given their evidence (Brandt, 1959; Gibbard, 1990; Hospers, 1961) rather than in terms of what the agent actually believes.

[3] Hayward (2020) calls an initiative's expected utility its "*effectiveness score*", but I prefer to stick with the usual terminology because I worry that "effectiveness score" invites confusion: e.g., the

ought to do is perform the action, out of those available, that maximizes utility; and, what you *subjectively* ought to do is perform the action that maximizes *expected utility*.[4] Hayward's true target, I take it, is this latter view: namely, that one ought to maximize expected value.[5]

The targeted view is embodied by Hayward's well-intentioned character, Bill, who undertakes a variety of doomed philanthropic spending initiatives. Bill is an adherent of the effective altruism movement, which (according to Hayward (2020: 440-1)) "*combines* the act-utilitarian doctrine that we should apportion our efforts according to efficiency with a doctrine concerning the *acquisition* of evidence about efficiency: that 'We should employ the best empirical research methods available in order to determine, as best we can, which efforts promote those values most efficiently' (Berkey, 2018: 147)".[6] Following Hayward, I'll continue to refer to this species of view as "act-utilitarianism", although there are distinct positions that share the name.

As previously mentioned, Hayward's ultimate conclusion is that utilitarian agents like Bill should sometimes either ignore evidence about effectiveness, or

---

expected utility of donating $x to a vaccine initiative needn't straightforwardly track the effectiveness of that vaccine. Because the effectiveness of an initiative can significantly deviate from the "effectiveness score" assigned to *supporting* that initiative, I think it's best to not confuse the two by using such similar terms.

[4] See, for example, Parfit (1984: 25), Gibbard (1990: 42-3), and Timmons (2002: 124). See also Jackson (1991: 462-3) for an influential example motivating the *expected value* account of 'subjective rightness'.

[5] Hayward characterizes the decision-rule differently: his rule says to "distribute [one's] resources in proportion to effectiveness scores," (435). Although there might be some cases in which the action that maximizes expected utility happens to involve distributing one's resources in proportion to "effectiveness scores", this won't typically be the case. In fact, the available actions are, traditionally, taken to be jointly exhaustive and *mutually exclusive*. But if your actions are mutually exclusive, it's impossible to distribute one's resources among them in proportion to their expected utilities. As far as I know, no act-utilitarians endorse the decision-rule as Hayward characterizes it. It's possible that some effective altruists do, but, if so, this would be a heterodox (and implausible) position. For these reasons, I'll interpret Hayward as speaking loosely here: one should distribute one's resources in the way that maximizes expected utility.

[6] I'm not sure how accurate a characterization of effective altruism this is—to my knowledge, the movement comprises a diverse range of normative views, most likely including act-utilitarians, but also other forms of consequentialists, as well as deontologists, etc. (For further discussion, see, e.g., Berkey, 2021; MacAskill, 2019)—but I'll grant that there is something of a *spiritual* kinship between effective altruism and act-utilitarianism. Furthermore, I think that Hayward's argument has much less to do with the *utilitarian* part of act-utilitarianism—it's not the axiological claim about ranking world-histories in terms of their sum of utilities that leads to utility cascades—than it does with *maximizing expected value* (whatever that value happens to be). If Bill were an egoist rather than an altruist, analogous examples could be constructed. Because many effective altruists do seem sympathetic to expected utility theory, then, they are fair targets. (That said, the fact that Hayward's argument really targets expected utility theory should make us suspicious that it succeeds—not the least of which because there are a number of formal results that strongly suggest that agents who maximize expected utility aren't vulnerable to these kinds of objections (e.g., Good, 1967).)

fail to apportion support to effectiveness. How are we to understand this 'should'? If it's the *objective* 'should', then Hayward's conclusion is undoubtedly true—but uninteresting. When evidence is misleading, it would be better *objectively* to ignore it; the initiative that is *actually* best is what should *objectively* be supported, not the initiative (if it is different) that is *expectedly* best. One doesn't need a utility cascade to show this, though; all one needs is an unpurchased winning lottery ticket. So, instead, I presume that the 'should' in the conclusion is meant to be *subjective*: that is, in some cases, act-utilitarians *subjectively* ought to ignore evidence, or *subjectively* ought to fail to apportion support in accordance with an initiative's expected utility. This is, I take it, what Hayward's utility cascades are meant to show. I agree that, if they succeed in showing this, that's an interesting and troubling problem for those, like members of the effective altruism movement, who are broadly sympathetic to act-utilitarianism. But, we have good reason, or so I'll argue, to be skeptical that Hayward's argument succeeds.

# 3 Unpacking Hayward's Argument

The argument is developed narratively—Hayward engagingly sets forth two examples of a utility cascade: the first, an *intra*personal case; the second, an *inter*personal case. In what follows, I focus on the former, developing lines of critique that have application in the interpersonal case as well.

## 3.1 The Intrapersonal Case: Bill and the Vaccine Initiative

Bill, Hayward's protagonist, is a wealthy philanthropist who is considering backing the rollout of a vaccine, effectanol. In June, the evidence available to Bill suggests that effectanol is 80% effective, which, given the good it could yield, means that the option *donate $10,000 per month to effectanol* maximizes expected utility. In July, Bill learns that the vaccine is only 70% effective, which changes his assignment of expected utilities so that, now, the option *donate $8,000 per month to effectanol and $2,000 to mosquito nets* maximizes expected utility.[7] But, because

[7] As an aside, it's worth re-emphasizing the difference between the *effectiveness* of the vaccine initiative (which, according to the story, has dropped) and the *expected utility* of supporting that initiative by allocating a particular sum of money to it. The fact that the former has dropped doesn't entail anything in particular about the latter. In fact, a drop in the effectiveness of the vaccine might rationalize allocating *more*, not less, money to the initiative. Suppose, for example, that Bill initially regards the vaccine as 100% effective. Among other things, Bill hopes the vaccine initiative will help the community achieve herd immunity. Furthermore, suppose that herd immunity can only be achieved if 70% or more of the population have immunity to the disease. If Bill were to learn that the vaccine is actually only 70% effective, it might make sense for him to allocate more money to the program because, now, it appears that more dosages will be needed—we'd need to inoculate everyone rather than only 70%—in order for herd immunity to

of Bill's reduced support, in August, Bill learns that the effectanol project is now even less likely to succeed, which makes it the case that, now, the option *donate $4,000 per month to effectanol and $6,000 to mosquito nets* is what maximizes expected utility. Bill's reduced support makes it such that, in September, he is forced to conclude that the project is no longer worth investing in at all. So Bill reduces his support to $0 per month. Once all is said and done, Bill has wasted thousands of dollars on a vaccine program that is ultimately unsuccessful. He has, Hayward tells us, fallen victim to the eponymous 'utility cascade'.

But what's the real problem here? Here's one way we could read the case. Bill, we might say, has behaved in a way that has brought about a sub-optimal outcome; as it turns out, it would've been better for him to have spent the entirety of his $10,000 per month on the mosquito nets all along. But, under conditions of uncertainty, bringing about an outcome that turns out to be sub-optimal isn't particularly objectionable, even for the act-utilitarian. We do it every time we buy fire insurance and our house doesn't burn.

What does this have to do with Hayward's claim that Bill would have done better to have ignored evidence? Hayward suggests that it would've been better for Bill to have never learned that effectanol was only 70%, rather than 80%, effective. This, in turn, suggests that the option *donate $10,000 per month to effectanol* is the one that maximizes actual utility—it was the optimal choice. But if that's right, then the information Bill received in July—the information that led him to conclude that effectanol was no longer worth donating the $10,000 per month to—was misleading.[8] Of course, there is a sense in which it is better to disregard *misleading* information (it always is), but, given that Bill had no reason to suspect that his evidence was misleading, there's nothing particularly objectionable about his having failed to do so.

On the other hand, it would be objectionable for Bill to *predictably* bring about a sub-optimal outcome (or to *knowingly* fail to disregard misleading information). But, given that Bill is a rational act-utilitarian who maximizes expected utility, he won't do these things. If he knows that some action will result in a sub-optimal outcome relative to some other, the expected utility of the former will be lower than that of the latter; and, because Bill maximizes expected utility, he will avoid performing it.

Let's apply this last thought to Hayward's utility cascade. Suppose that, in July (after receiving the disappointing news about effectanol), Bill can predict that, if

---

be secured.

[8] It might be that the information was correct about the effectiveness of the effectanol vaccine—it's only 70%, and not 80%, effective—but misleading in the sense that it downgraded the expected utility of the option that, as a matter of fact, maximized utility. To borrow a distinction from (Buchak, 2010), the information might not be *epistemically* misleading, but nevertheless misleading in an *instrumental* sense.

he lowers his monthly contribution from $10,000 to $8,000, there'll be a spike in infections which will make the initiative not worth donating $8,000 to (and likewise for lowering his monthly contribution from $8,000 to $4,000, and lowering it from $4,000 to $2,000, and so on).If Bill can predict this for sure, then the expected utility of lowering his contribution from $10,000 to $8,000 per month should be about roughly the same as the expected utility of pulling *all* his support. After all, if Bill is attentive to all the consequences of his actions, he can predict that that's where he's headed. If that's right, then Bill effectively has two options still in the running:

A Continue donating all $10,000 per month to effectanol.

B Pull all funding from effectanol and transfer it to mosquito nets.

The example doesn't provide us with enough detail to determine which of the two options maximizes expected utility. If it's A, then, given that A is the option that maximizes actual utility, Bill ends up doing exactly what he objectively ought to do—and there's no sense in which it would've been better for him to have avoided July's evidence. On the other hand, if it's B, then Bill fails to do what is *actually* best, but only as the result of doing what made the most sense given the misleading information he'd been handed. Options that look best *ex ante* aren't always best *ex post*. This mundane fact poses no new threat to the act-utilitarian.

Let me counter a potential response here—one that Hayward gestures toward when he says, of suggestions like the one that Bill should assign very low expected utility to the prospects of lowering his contribution from $10,000 to $8,000 per month, that:

> [T]his does not vitiate the problem of utility cascades—it capitulates to them. It illustrates the limitations they place upon the deliberations of utilitarian agents. After all, this lowering of efficiency scores is based not on any innate feature of the situation, but on Bill's *own* vulnerability to cascades. [...] [i]n this case, one of the limitations of the world is that it contains epistemically rational act-utilitarian agents, (Hayward, 2020: 438).

Hayward's idea is that, if Bill's reason for assigning very low expected utility to lowering his contribution from $10,000 to $8,000 is that he anticipates that Bill-in-a-month-from-now will lower it even further (and so on and so on), this is *still* objectionable because this is an example of Bill obstructing himself.

I think this is mistaken. It's true that Bill (correctly) anticipates that, were he, in July, to lower his contribution to $8,000, he'd think, in August, that he was allocating too much to the initiative. But the reason he, in July, should assign very

low expected utility to lowering his contribution to $8,000 isn't *because* future-Bill will lower the contribution even further; instead, it's because, if he lowers the donation, the vaccine initiative won't be worth funding at that level—a fact that he's both in a position to appreciate *now* as well as anticipate appreciating in a month from now. By assigning a low expected utility to reducing his donation, Bill *isn't capitulating* to his future-self; he's not making a concession to a future decision that he, now, doesn't endorse; instead, he's responding to the fact that funding the initiative at $8,000 will result in it no longer being worth funding at that level.

It's not the *cascade itself* that explains why Bill assigns low expected value to the option; rather, it's the facts underlying the cascade—the facts about the effects his choices have on the initiative's success, to which his future-selves would be responding—which justify his evaluations of the options. So, *contra* Hayward, "this lowering of efficiency scores" *is* based on features of the situation, and not on Bill's vulnerability to cascades.

## 3.2 The Interpersonal Case: Bill (& co.) vs. Climate Change

The central lesson of the previous section—that expected utility maximizing act-utilitarians will either avoid the utility cascade or fail to avoid it but in a way that is not objectionable—carries over to Hayward's interpersonal example which deals with the problem of climate change. In this case, Bill is recognized by other philanthropists as an important contributor to charitable projects: his every move is scrutinized and others weigh the likely efficacy of their own contributions in relation to those that they assume Bill will tender. Now, in this case, either Bill knows that, for example, shifting his support from preventative measures to mitigative ones will herald a similar shift on the part of other members of his community (thus rendering the preventative measures even less likely to succeed than before) or he doesn't.

Suppose he does know this. Then, given that he's a rational act-utilitarian, he will take that information into account when deciding what to do. So after receiving the first round of bad news about the efficacy of the preventative measures, it's not obvious that Bill is committed to shifting his contributions from prevention to mitigation. That bad news will lower the expected utility he assigns to the preventative strategy, but—given that, as we're supposing, he can predict that other members of his community will follow his lead (plus, that at least some of them are privy to the same information he is)—it's implausible to think that it will have lower expected utility than shifting his contributions from prevention to mitigation will. Why? So long as Bill isn't shortsighted, he'll know that the predictable result of shifting his contributions away from prevention will eventually result in his (and the rest of the community) going *all in* for mitigation. So, un-

less Bill thinks that it's not worth putting *any* resources toward prevention, he'll regard keeping his contributions where they are as the better of the two options. As before, the cascade is avoided.

On the other hand, if Bill has no idea that his decision will influence the behavior of the others (and his future-self), then down the cascade he might go. But why is that a mark against act-utilitarianism? If it turns out that, unbeknownst to me, a murderous villain will destroy the planet if I snap my fingers, there's a sense (the objective one) in which I shouldn't snap my fingers; but, given that I have no inkling of the influence my decision will have, how am I criticizable? Furthermore, just as before, this is a case in which Bill receives misleading evidence about the effectiveness of the potential interventions—the evidence he receives suggests that prevention isn't worth supporting at its current level when, in fact, it is.

Let me address a potential objection—analogous to the one from §3.1—which holds that, while Bill might be doing the best he can do given the situation in which he finds himself, that situation would nevertheless be even better if it didn't contain so may act-utilitarians, but instead contained agents who acted in accordance with different principles.[9] The worry is that, if Bill's reason for assigning very low expected utility to shifting his support from prevention to mitigation is that he anticipates that this will herald a similar shift on the part of the other members of his community, this is an example of act-utilitarianism *collectively* getting in their own way. As before and for analogous reasons, I think this is mistaken. The reason Bill should assign very low expected utility to shifting his resources from prevention to mitigation isn't *because* other members of the community will lower their contributions *per se*; instead, it's because, if *he* shifts his support away from prevention, prevention won't be worth funding by his peers either. His contributions—or lack thereof—alter the effectiveness of the interventions, which in turn affect (in ways he wholly endorses) what other act-utilitarians ought to do.

That said, one potentially important difference between Hayward's interpersonal case and his intrapersonal one is that Bill's peers might act contemporaneously—and, thus, absent knowledge of what he and the others have likewise decided. This added degree of uncertainty might make it difficult—especially if the decisions are made independently and in ignorance of each other—for the community to coordinate on the optimal level of funding. But, as previously argued, it's far from clear that bringing about a sub-optimal outcome in the face of uncertainty is particularly objectionable.[10] However, even if this were objectionable in some sense,

---

[9] Thanks to an anonymous referee for pressing this point.

[10] Spelled out in this way, Hayward's interpersonal case potentially raises some interesting issues about coordination. Suppose, for example (and simplicity), that there are two effective altruists—

it's not clear that it constitutes an objection to act-utilitarianism. Instead, what it would seemingly show is that act-utilitarianism is "indirectly collectively self-defeating" (Parfit, 1984: 27): things would be worse, by act-utilitarian lights, if everyone tried to behave like one. But (as argued in Parfit, 1984: 27-8) this doesn't show that act-utilitarianism is *false*—only that, if it's true, it would be worse if we all attempted to behave like it.

## 4 Conclusion

The versions of the examples in which Bill is vulnerable to sliding his way down a utility cascade involve two things: (i) misleading evidence, which pushes Bill away from the initiative that is, in fact, the one that it would be optimal to support; and, (ii) uncertainty surrounding the effects that Bill's decisions will have on the overall effectiveness of the available interventions. (In the versions without these two features, Bill avoids the cascade.) Bringing about a sub-optimal outcome in these circumstances—that is, in the face of misleading evidence and under a thick fog of uncertainty—isn't particularly objectionable. Furthermore, the remedy for such situations, *contra* Hayward, is *not* to ignore evidence—it's to gather more of it. Subjectively, the best cure for misleading information is more information; and, a promising way for an act-utilitarian to avoid a utility cascade is to become better informed about the effects of their actions on both the effectiveness of the available interventions and the group dynamics of other altruistically-minded agents.

I have argued that Hayward's utility cascades don't pose a significant worry for act-utilitarians in principle. But what about the members of the effective altruism movement? Are they vulnerable to utility cascades? Perhaps members of the effective altruist community are prone to naively evaluate interventions only in terms of their local effectiveness, ignoring the effects that their very support might have on the intervention's success. If so, I agree that this would be a mistake. But, I imagine, so would they; if this is a mistake effective altruists make, I doubt it's one

---

Bill and Pete—who are both currently funding an initiative to a high degree. Suppose, for simplicity, that there are three levels at which they each can elect to fund the initiative: High, Low, or No. Following the current evidence, they have both elected High. New evidence comes in which suggests that the initiative is only worth funding at a moderate level. The appropriate level can be achieved in various ways: Bill could lower his contribution from High to No, while Pete keeps his contribution High; or Pete could lower his contribution, while Bill keeps his contribution fixed; or both could lower their contribution from High to Low. Because there are several ways to accomplish what they ought to do together, if they must make their choices in ignorance of what the other is doing, they face a coordination problem. What Bill ought to do depends on what Pete will do, which in turn depends on what Bill will do. They very well might fail to coordinate, resulting in a sub-optimal outcome. Whether this is a serious problem for act-utilitarians (and/or effective altruists) is a matter of some dispute (see, for example, Collins, 2019; Dietz, 2019; Gibbard, 1965; Parfit, ms; Regan, 1980) and outside the scope of this paper.

that they'd endorse making. But, in any case, what's called for is *more* information, not less. The lesson I take from Hayward's examples is not that act-utilitarians should avoid evidence, or selectively ignore it, but that, rather, they should seek out as much information as possible. Their view does not license ostriching.

All that said, I think Hayward's utility cascades potentially do raise some very interesting issues—if not for impeccably rational expected utility maximizing act-utilitarians, than at least for the rest of us—about how best to navigate a complex, interconnected world. It might be that, in an environment replete with potential cascades, groups of expected utility maximizers will do worse—in some sense—than groups employing some other decision-rule. When maximizing is difficult, it might be better to not even try—to adopt some simple heuristic instead (e.g., Todd and Gigerenzer, 2012).[11] That might be right, but more would be required to show it.[12]

## References

Berkey, Brian. 2018. The Institutional Critique of Effective Altruism. *Utilitas*, 30: 143–171

——. 2021. The Philosophical Core of Effective Altruism. *Journal of Social Philosophy*, 52(1): 93–115

Brandt, Richard. 1959. *Ethical Theory: The Problems of Normative and Critical Ethics* (Englewood Cliffs: Prentice-Hall)

Broad, C.D. 1985. *Ethics*, chap. 3 (Dordrecht: Martinus Nijhoff)

Broi, Antonin. 2019. Effective Altruism and Systemic Change. *Utilitas*, 31: 262–276

[11] Of course, care is called for in our choice of heuristics. Some members of the effective altruism movement (e.g., MacAskill, 2015; Wiblin, 2016) endorse a heuristic for "cause prioritization"—the Importance-Tractability-Neglectedness (ITN) Framework—that is particularly prone to misfire when applied to Hayward's examples. In particular, if Hayward (2020: 436) is correct that these are cases of *increasing* marginal utility, then Neglectedness (or the lack thereof) would not provide a reliable guide to how one ought to prioritize various causes: e.g., the fact that many others have already invested in prevention—that it is anything but neglected—doesn't mean that it would be less valuable for you to too. (For similar criticisms of the ITN framework, see Broi, 2019; Halstead, 2019). Because all heuristics are just that, the measure of a good one cannot be infallibility. But, if Hayward's cases are prevalent enough, it might be wise for effective altruists to move away from the ITN framework nonetheless. Thanks to an anonymous referee for suggesting this point.

[12] For helpful feedback and discussion, I'd like to thank Max Hayward, Simone Gubler, Frances Howard-Snyder, an audience at the 2021 APA Eastern Division Meeting, and an audience at the 2020 Annual PPE Society Meeting. I would also like to thank Geoff Sayre-McCord for starting the cascade of events that occasioned the existence of this paper, and two anonymous referees for their helpful feedback.

Buchak, Lara. 2010. Instrumental Rationality, Epistemic Rationality, and Evidence-gathering. *Philosophical Perspectives*, 24: 85–120

Collins, Stephanie. 2019. Beyond Individualism. In *Effective Altruism: Philosophical Issues*, edited by Hilary Greaves and Theron Pummer (Oxford: Oxford University Press), pp. 202–217

Dietz, Alexander. 2019. Effective Altruism and Collective Obligations. *Utilitas*, 31(1): 106–115

Doody, Ryan. ms. Consider the Ostrich: Non-Utilitarians, Ex Ante Interest, and Burying Your Head in the Sand. Unpublished manuscript

Feldman, Fred. 2006. Actual Utility, the Objection from Impracticality, and the Move to Expected Utility. *Philosophical Studies*, 129(1): 49–79

Gibbard, Allan. 1965. Rule-Utilitarianism: Merely an Illusory Alternative? *Australasian Journal of Philosophy*, 43: 211–220

——. 1990. *Wise Choices, Apt Feelings* (Oxford: Oxford University Press)

Good, I.J. 1967. On the Principle of Total Evidence. *The British Journal for the Philosophy of Science*, 17(4): 129–139

Halstead, John G. 2019. The ITN framework, cost-effectiveness, and cause prioritisation. *Effective Altruism Forum*. URL `https://forum.effectivealtruism.org/posts/Eav7tedvX96Gk2uKE/the-itn-framework-cost-effectiveness-and-cause`

Hayward, Max Khan. 2020. Utility Cascades. *Analysis*, 80(3): 433–442

Hospers, John. 1961. *Human Conduct: An Introduction to the Problems of Ethics* (New York: Harcourt, Brace and World)

Jackson, Frank. 1991. Decision-theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics*, 101: 461–482

MacAskill, William. 2015. *Doing Good Better: How Effective Altruism Can Help You Make a Difference* (New York: Penguin Random House)

——. 2019. The Definition of Effective Altruism. In *Effective Altruism: Philosophical Issues*, edited by H Greaves and T Pummer (Oxford: Oxford University Press)

Mulgan, Tim. 2001. *The Demands of Condequentialism* (Oxford: Clarendon Press)

Oddie, Graham and Peter Menzies. 1992. An Objectivist's Guide to Subjective Value. *Ethics*, 102: 512–533

Parfit, Derek. 1984. *Reasons and Persons* (Oxford: Clarendon Press)

——. ms. What We Together Do. Unpublished manuscript

Railton, Peter. 1984. Alienation, Consequentialism, and the Demands of Morality. *Philosophy and Public Affairs*, 13(2): 134–171

Regan, Donald. 1980. *Utilitarianism and Cooperation* (Oxford: Oxford University Press)

Russell, Bertrand. 1910. The Elements of Ethics. In *Philosophical Essays* (London: Longmans, Green and Co.), pp. 3–51

Sidgwick, Henry. 1907. *The Methods of Ethics* (The University of Chicago Press)

Smart, J.J.C. 1973. An Outline of a System of Utilitarian Ethics. In *Utilitarianism: For and Against*, edited by J.J.C. Smart and Bernard Williams (Cambridge: Cambridge University Press)

Smith, Holly. 2010. Subjective Rightness. *Social Philosophy and Policy*, 27(2): 64–110

Timmons, Mark. 2002. *Moral Theory: An Introduction* (Edinburgh: Edinburgh University Press)

Todd, Peter M. and Gerd Gigerenzer. 2012. Ecological Rationality: The Normative Study of Heuristics. In *Ecological Rationality: Intelligence in the World*, edited by Peter M. Todd, Gerd Gigerenzer, and the ABC Research Group (Oxford: Oxford University Press)

Wiblin, Rob. 2016. A framework for comparing global problems in terms of expected impact. *80,000 Hours*. URL `https://80000hours.org/articles/problem-framework/`

Zimmerman, Michael. 1996. *The Concept of Moral Obligation* (Cambridge: Cambridge University Press)