JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXVIII · NUMBER 2
August 2024

ARTICLES

DISCUSSION

JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY
http://www.jesp.org

The *Journal of Ethics and Social Philosophy* (ISSN 1559-3061) is a peer-reviewed online journal in moral, social, political, and legal philosophy. The journal is founded on the principle of publisher-funded open access. There are no publication fees for authors, and public access to articles is free of charge and is available to all readers under the CREATIVE COMMONS ATTRIBUTION-NONCOMMERCIAL-NODERIVATIVES 4.0 license. Funding for the journal has been made possible through the generous commitment of the Division of Arts and Humanities at New York University Abu Dhabi.

The *Journal of Ethics and Social Philosophy* aspires to be the leading venue for the best new work in the fields that it covers, and it is governed by a correspondingly high editorial standard. The journal welcomes submissions of articles in any of these and related fields of research. The journal is interested in work in the history of ethics that bears directly on topics of contemporary interest, but does not consider articles of purely historical interest. It is the view of the editors that the journal's high standard does not preclude publishing work that is critical in nature, provided that it is constructive, well argued, current, and of sufficiently general interest.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3130

MORAL DEMANDINGNESS AND MODAL DEMANDINGNESS

Kyle York

Moral demandingness is best understood as a modal feature of moral theories. In other words, we usually ought to evaluate the demandingness of moral theories in a way that includes how demanding those theories are in other (perhaps all) possible worlds. Once we come to see that this is the right way of evaluating a theory's demandingness, however, we will come to find that evaluating the overall demandingness of moral theories is very difficult. In particular, we will lose our confidence that even commonsense morality can avoid charges of being extremely demanding. Rather than undermining commonsense morality, these considerations will undermine the tenability of a certain sort of demandingness objection, which I will call *cost based* (as opposed to *reasons based*). Without being able to justifiably appeal to costs per se, I will suggest that we can object to theories only in terms of the reasons they give us for incurring those costs. I will start by explaining more clearly what I have in mind when I use these terms. For simplicity, I will focus on demandingness objections made against maximizing act consequentialism, though my same points will apply to cost-based objections made against any moral theory or principle, including deontological and virtue-ethical ones.[1]

By cost-based demandingness objections, I am not talking about demandingness objections that are based upon the costs as opposed to, say, the difficulty or restrictiveness of following some moral theory.[2] I am happy to stick with the tradition of understanding demandingness in terms of the costs of following some theory. But demandingness objections can be lodged on the basis of the fact that following some theory is either (1) too costly relative to the reasons that an agent has for incurring those costs or (2) too costly in terms of crossing a fixed threshold of acceptable costliness. Demandingness objections of type 2 are what I call "cost-based" demandingness objections. They are the sort that I am targeting.

1 See, for example, Fragnière, "Climate Change and Individual Duties," 805–6; and Sin, "The Demandingness of Confucianism in the Case of Long-Term Caregiving," 166–79.

2 For discussions of such accounts of demandingness, see Chappell, "Willpower Satisficing," 251–65; and van Ackeren, "How Morality Becomes Demanding," 315–34.

The other kind of demandingness objections—type 1—I refer to as *reasons-based* demandingness objections. Unlike cost-based demandingness objections, reasons-based demandingness objections have flexible thresholds of acceptable costs that shift for various reasons. I wish to define this type only in broad terms. It likely admits of different subtypes or specifications, depending on how one views the proper proportionality between costs and reasons, the proper means of aggregating reasons, and so on. For example, a theory requiring me to donate my eyes to someone who would get more utility out of them might be called "too demanding" in the reasons-based sense. This is apparently because I have some legitimate reason to favor myself in such a case. On the other hand, as one reviewer has pointed out, a theory may also be too demanding by asking an agent to sacrifice his life in exchange for everyone in the world getting a free latte, even if the exchange would result in a substantial overall gain in utility. This would also count as a "reasons-based" demandingness objection, though it might appeal to, say, principles of aggregating costs.[3] As an aside, while these terms are my own, the distinction is not. Sarah Stroud, for example, differentiates the view that "whether a given moral demand is acceptable varies with the *reason* offered in support of the demand" from the view that there are certain costs that agents simply cannot be morally required to bear, also finding the latter view to be unviable.[4]

For my prototype of the cost-based demandingness objection, I borrow from Brian McElwee's "pure-demandingness objection," which he contrasts with objections based upon the "insufficient importance," "wrong moral ranking," or "wrong overall ranking" of moral demands relative to other relevant considerations.[5] McElwee has formulated the "pure demandingness objection" (which, for consistency, I will refer to as the "cost-based demandingness objection") as follows:

> The purported moral considerations in favour of doing *A*, which genuinely are moral considerations and potentially of sufficient importance to generate a moral obligation, are not outweighed by any moral or nonmoral considerations speaking against doing *A*, but nonetheless are insufficient in context to generate a moral obligation, because the cost to the agent is too great.[6]

3 Compare Temkin, "A 'New' Principle of Aggregation," 218–34.

4 Stroud, "They Can't Take That Away from Me," 228. See also Hooker, "The Demandingness Objection."

5 McElwee, "Demandingness Objections in Ethics," 88–92.

6 McElwee, "Demandingness Objections in Ethics," 90–91. This might be slightly confusing, since presumably many of the nonmoral considerations that may render an act

We will take the cost to the agent being too great here in terms of that cost having crossed a fixed threshold of acceptability. I focus on this formulation not because I think that this is the best version of the demandingness objection—quite the opposite. As my argument will show, it seems like only the reasons-based version of the demandingness objection is viable.[7] In other words, I will argue against objections to moral theories that appeal merely to those theories' costliness per se (defined in terms of crossing some threshold of acceptable costs).

My argument should also have some additional, further-reaching implications for the demandingness debate. Traditionally, the sorts of costs taken to be relevant to moral demandingness were those costs related to moral requirements. That is, demandingness objections usually take aim at what some moral theory requires, particularly concerning our duties of beneficence, as opposed to what it permits or forbids. For example, if a theory implies that you must donate one of your kidneys to a needy stranger, you might object that it is for that reason too demanding. It is not traditional, however, to object to moral theories on the grounds of the opportunity costs incurred because of those theories' restrictions. For example, it would be highly unusual to think that a demandingness objection could succeed against a moral theory because that theory restricts me from murdering someone for their liver, even if I need a new liver to survive.

Shelly Kagan and David Sobel have attacked demandingness objections on this very point. They have asked why, if moral theories (like consequentialism) can be criticized over the costs associated with their requirements, should not those critics' own (say, commonsense) moral theories be similarly objected to on grounds of the costs associated with their permissions or restrictions.[8] This would turn the potential success of the (cost-based) demandingness objection into a pyrrhic victory for most of those who would use it. The commonsense moral theorist, however, still has a significant opportunity to push back by pointing to significant differences between restrictions, permissions, and requirements.[9] A modal view of demandingness, however, might sidestep this issue

---

supererogatory just are costs to the agent. But as long as we take "nonmoral considerations" to denote everything except costs, McElwee articulates an important formulation of the demandingness objection that avoids concerns about the justifications of demands.

7 Some might think that the cost-based formulation of the demandingness objection represents only a fringe view, which would admittedly make my arguments considerably less interesting. I respond to this concern in the first part of the "Objections" section below.

8 Kagan, *The Limits of Morality*, 19–24; and Sobel, "The Impotence of the Demandingness Objection," 238–60.

9 See, for example, Chappell, "Willpower Satisficing," 251–65. Chappell defends satisficing consequentialism in this article, but his argument about the ostensive differences between restrictions and requirements, regarding demandingness, could easily be taken up by any commonsense theorist.

by suggesting that even the requirements of commonsense morality are very demanding. This will not be to suggest that they are as demanding as, say, maximizing act utilitarianism, but they nonetheless might be demanding enough to undermine the commonsense theorist's capacity to raise the objection. To see how, let us first consider why a general understanding of demandingness makes the most sense. This will set the stage for introducing a modal understanding later.

1. General Demands and Contingent Demandingness

Imagine that I live in a village and that everyone in the village has agreed to take turns keeping watch at night. Whoever is keeping watch, it is agreed, must alert the other villagers of attackers by blowing a horn, even though blowing this horn will draw the attention of the attackers and put the guard's life at the greatest risk. Attackers are very rare. The following moral principle seems very plausible: any villager who has entered into this agreement incurs a moral obligation to blow the horn if she is on guard duty and sees attackers. But we run into a problem here. In any particular instance when a guard actually sees attackers, there are *prima facie* grounds for that guard to make a demandingness objection to this moral principle. After all, the principle seems to require her to risk her very life.

But clearly, someone on guard duty who sees attackers coming cannot base a convincing demandingness objection on the costs that she now faces. After all, this is the exact sort of cost that the guard agreed to take on if the occasion arose, and everyone is depending on her for their lives. Since she enjoys the benefits of the agreement when off duty and at the risk of others, she'd certainly be acting wrongly by blowing off her own duties when it is up to her to sound the horn. One might want to say now that, after more reflection, none of the villagers actually incur a duty when they enter into the agreement, because it is too demanding an agreement. This seems plausible when it comes to, for example, a frivolous and swaggering agreement to play Russian roulette. For one thing, nobody has a sufficiently good reason to enter into such an agreement. In the villagers' case, however, they are taking on the risk of large costs in order to avoid even larger costs. If no one entered into an agreement to keep watch, then the whole village would be defenseless against attackers. A few might finally be tempted to say that the guard's blowing of the horn would be supererogatory simply because her nonmoral, prudential reasons for not blowing it outweigh her moral reasons for blowing it. While this does not seem true, even if it were, it is the sort of move that is available only to reasons-based demandingness objections. Cost-based demandingness objections are supposed to point to flaws in moral theories in accordance with their production of overly costly requirements. The notion that these ostensive features of moral theories are

flaws, however, seems to lose its force if moral reasons are not supposed to be overriding in such cases.

This should lead us to conclude that even from a commonsensical perspective, there are some cases in which requirements hold despite extreme costs. Does this spell doom for the cost-based demandingness objection? It probably does not. It would seem plausible enough for the commonsense theorist to respond that theories and principles cannot be characterized as overly demanding in some particular instance but only in light of their demands to agents across the whole spectrum of the situations that they apply to. We can tell the objecting guard that, although she faces large costs, the situation that she's in so seldom arises that the moral principle in question is still *generally undemanding*. This of course will not make her feel any better, but it may take away her grounds to complain.[10] Let us therefore reformulate the cost-based demandingness objection as the objection for which:

1. The purported considerations in favor of moral theory *X* are not outweighed by any considerations against *X* but nonetheless are insufficient to compel us to accept *X* because *X* would be generally too costly for agents to obey.[11]

This reformulation also seems to account for the well-known fact that the burden some moral theory places on me will vary with the situation I am in. Someone isolated on the moon will face little to no moral demands because the moral demands that one faces depend on features of one's circumstances.[12] We might call this the "contingency" of morality's demandingness.

Just as the demandingness of some moral theory or principle will vary according to who in the world you happen to be, so will it vary according to which world you happen to be in. Consider, for example, Sidgwick's suggestion that we need not worry about the demandingness of utilitarianism since the extent of what we can do to help others is fairly minimal.[13] Liam Murphy has objected to this suggestion on the ground that, even assuming this was true during Sidgwick's time, it fails to carry into the twentieth century.[14] But more importantly, its failure to carry over shows us that Sidgwick's reply "is in any case an entirely contingent response to the claim of extreme demands, since it

10 Compare Ashford, "The Demandingness of Scanlon's Contractualism," 295.

11 I do not mean "compel" in a psychological sense. I roughly just mean "would compel a sufficiently rational and moral agent."

12 See Carbonell, "Differential Demands," 427–38; and Scheffler, *Human Morality*, 98.

13 Sidgwick, *The Methods of Ethics*, 428.

14 Murphy, *Moral Demands in Nonideal Theory*, 13.

obviously cannot be ruled out that changing circumstances will make it possible for individuals to do great good for others at great (though lesser) cost to themselves."[15] Those who endorse something like the cost-based demandingness objection could draw from Murphy's point against Sidgwick an even stronger version of the cost-based demandingness objection. The stronger version provides a modal view of demandingness:

2. The purported considerations in favor of moral theory *X* are not outweighed by any considerations against *X* but nonetheless are insufficient to compel us to accept *X* because *X could be* generally too costly for agents to obey.

Let us suppose that Sidgwick was right and there really was not much that people could do to help many others in the nineteenth century, and as a result, utilitarianism did not make extreme demands on people. We would not want to say that during Sidgwick's time, utilitarianism was correct, but it no longer is because the demandingness of the theory has changed.[16] Because moral theories are supposed to be the sorts of things that can apply universally, it seems strange to accept the idea that a moral theory was true up until a certain point in history. At most, all we should say is that a moral theory was truth-tracking up until a certain point in history. This consideration makes the stronger version of the cost-based demandingness objection seem like the more plausible choice: over time, we would have discovered that there had *always* been something wrong with consequentialism.[17]

2. A MODAL VIEW OF DEMANDINGNESS

Let us return to the village. In another possible world, attacks on the village are very frequent, occurring once every few nights. Our guard again finds herself unlucky: attackers are coming. All the same considerations apply. She entered

15 Murphy, *Moral Demands in Nonideal Theory*, 13.

16 It may be tempting to say that utilitarianism was always true but is not obligation generating any longer due to its increased demandingness. But if a theory's requirements can be successfully objected to on grounds of demandingness per se, it seems to fail. After all, a theory does not generate obligations; facts do that. Theories state obligations, and if a theory falsely states that I am obliged to ϕ, that theory is presumably flawed.

17 One might object that version 1 should be sufficient for objecting to consequentialism, even (theoretically) in Sidgwick's time, if we were to include past and future agents in our consideration. But even if the relevant features of the world never changed from Sidgwick's time on, Murphy's point should still hold because those features could have changed. Version 2 is clearly preferable, then, because it can account both for different possible worlds and for different times.

the agreement knowing this risk, specifically so she could benefit from others incurring the risk for all the nights that she is off duty. The agreement keeps her safer than she would otherwise be. (Imagine that there is another village nearby where nonagreeing villagers can go to live—one that is, obviously, less safe to live in.) She still seems to have a duty to blow the horn if she sees attackers, even if she will die. The rest of the village depends on her. The villagers, if any survived the attack, would justifiably consider themselves as having been wronged by her if she failed to blow the horn.

So now we seem to have a commonsensical moral principle that turns out to be extremely demanding in a general sense. Version 2 of the cost-based demandingness objection failed to protect it. Does this now spell doom for the cost-based demandingness objection? Not yet, since a similar, generalizing move can be made again. Recall that the main target of those making cost-based demandingness objections tends to be consequentialism. They may here say something like, "Although commonsense principles can be extremely and generally demanding in some possible worlds, consequentialism is still worse!" This response might work. Consequentialism might (a) make extreme general demands in many more possible worlds than commonsense principles and (b) make *more extreme* demands than commonsense morality in even (most of) the possible worlds where commonsense principles do make extreme general demands. Response b does not strike me as particularly interesting, since it seems only to tell us that it is possible to criticize the extreme demands of consequentialism without necessarily implicating commonsense morality. But as it stands, b says nothing on whether such a move would be warranted. After all, my argument is not that consequentialism is no more demanding than commonsense morality. Whether or not this is the case, the commonsense moral theorist appeals to the cost-based demandingness objection at risk of falling on her own sword. For if consequentialism *is* too demanding, this means that consequentialism's demandingness has crossed a threshold of acceptable costs. But if commonsense morality can also make extreme general demands, then the commonsense theorist will be hard pressed to show why commonsense morality does not also cross the line of being overly demanding. She seems to require a well-motivated theory of where the threshold is. If extreme demands can be commonsensical in some circumstances, then it seems more plausible that demandingness acceptability thresholds respond to reasons for incurring certain costs, not just to costs alone. This seems, again, to support the reasons-based view over the cost-based view of demandingness.

A more promising line of response comes from response a. Looking at how demanding commonsense morality and consequentialism would become in other possible worlds seems to tell us something important about the

demandingness of such theories. Perhaps the problem with consequentialism is that it is the sort of theory that can generate extremely demanding requirements at the drop of a hat, while commonsense morality holds out for rare and unusual conditions. If we want to assess demandingness in a modal way, then we should probably look for more than just whether a particular moral theory can become extremely and generally demanding in certain possible worlds. Rather, we should consider how many worlds have provoked such extreme demandingness. That is, our comparisons of different moral theories' demandingness should zoom out to regard their performance over the range of relevant possible worlds. We might think that the problem with Sidgwick's argument is not that consequentialism's demands have the capacity to be extreme. Rather, the problem is that consequentialism's demandingness is quite likely to become extreme. That is, the very fact that some moral theory (in our case, commonsense morality) may become extremely demanding in certain possible worlds may not suffice to establish that it is too demanding a theory in general. A theory would have to be extremely demanding in *too many* possible worlds. With these considerations in mind, the defender of commonsense morality can (in my view, correctly) suggest the following as the most reasonable version of the cost-based demandingness objection:

3. The purported considerations in favor of moral theory *X* are not outweighed by any considerations against *X* but nonetheless are insufficient to compel us to accept *X* because *X* is generally too costly for agents to obey in *too many* possible worlds.

Recall that earlier, the defender of commonsense moral theory suggested that the fact that commonsense morality may ask me to give up my life is not a sufficient basis to call the theory too demanding. This is why McElwee's version of the cost-based demandingness objection did not seem sufficient. We have to explicitly consider the *general* requirements of a moral theory, not simply how it would apply in one particular circumstance. Likewise, the defender of commonsense moral theory is now saying that we also cannot judge a theory by the general requirements it produces in one particular world. If this is true, then I cannot, as in version 1, merely talk of a theory being "generally too costly for agents to obey" in some world. While being costly or demanding can be an "intermodal" or "intramodal" property, being *too demanding* can only be an intermodal property (i.e., one that can only be correctly attributed to a theory from a modal perspective). One can think of it this way: while the move from McElwee's version of the cost-based demandingness objection to version 1 involved the idea that we should not apply the charge of over-demandingness from the perspective of a theory's requirements in some particular

circumstance but only in general, the move from version 2 to version 3 involves the idea that we should not apply the charge of over-demandingness because of the theory's requirements in some particular *world* but because of its modally general applications. Just as commonsense morality cannot be too demanding just by virtue of having extremely costly requirements for some agent *A* in some circumstance *C*, it also cannot be too demanding just by virtue of having extremely costly general requirements in some possible world *W*.

3. A Problem for the Commonsense Theorist's Demandingness Objections?

There are two further responses available to counter the commonsense theorist. The first response is that perhaps this is one of the rare worlds where consequentialism turns out to be demanding. This sort of point has been made by Katarzyna de Lazari-Radek and Peter Singer:

> In Victorian times it would have taken weeks for news of a distant famine to reach London, and months for any substantial amounts of grain to be gathered and transported to those in need. Now we can receive news instantly, and transport food and medical supplies within days. . . . Also highly significant, however, is the fact that the gap between rich and poor—and thus the power of the rich to help the poor—has greatly increased since the mid-nineteenth century.[18]

The consequentialist might hope to point towards circumstances in our world that have rendered consequentialism particularly demanding at the present moment and to argue that these sorts of circumstances are unusual. Perhaps a sizable portion of other possible worlds features conditions under which altruistic sacrifice is less costly or frequently needed. But if one must always maximize the good, and only a small range of actions will satisfy this requirement, it seems likely that what we will have to do almost all of the time will not be what we want to do. There will of course be occasional ties and situations where deliberation between options is more costly than just picking one's preferred option, but it seems plausible enough that such situations will be the exception rather than the rule.[19]

18 De Lazari-Radek and Singer, "How Much More Demanding Is Utilitarianism Than Common Sense Morality?" 433.

19 It might be argued here that although maximizing act consequentialism will always prescribe only a small range of actions as the morally required options, the wrongness of not following these requirements will vary depending on the world. In possible worlds where world poverty has ended, maximizing consequentialism may then require people

Even if this is so, however, the commonsense theorist runs into another problem. As discussed earlier, if consequentialism turns out to be extremely demanding in more possible worlds than commonsense morality does, this does little to show that commonsense morality will not nonetheless turn out to be too demanding in a modal assessment. We still need to see whether commonsense morality crosses some threshold of acceptable demandingness. This is a plausible objection, and it points to an even larger problem: there seems to be no way to evaluate in how many possible worlds a moral theory would be extremely demanding. After all, how many possible worlds are there? There are seemingly infinite possible worlds. As a result, how are we to assess the general modal demandingness of commonsense morality?[20] Even if we can confidently assert that maximizing act consequentialism is more demanding, demandingness objections still risk backfiring against the commonsense moral theorist if she cannot account for how demanding her theory is. It does not seem like the mere fact that we do not know how demanding commonsense morality really is excuses us from these worries. After all, we either are or are not required to follow commonsense moral requirements, and these requirements supposedly depend on the demandingness of commonsense moral theory. So the fact of how demanding commonsense morality is will be very important. It may be true that as a polemic move, accusations of demandingness made towards maximizing consequentialism are more easily accomplished than accusations against commonsense morality. Nevertheless, it seems plausible enough that commonsense morality could be extremely demanding and therefore, according to the cost-based demandingness objection, false. Therefore, the commonsense theorist makes her objection, at the very least, at the risk of undermining her own claim to knowledge regarding the truth of commonsensical moral principles.[21]

to donate their time and money to a less urgent cause. And ignoring this requirement will be, as one reviewer has suggested, less wrong than ignoring the requirement to help end world poverty. However, the demandingness of a moral theory has traditionally been thought of as corresponding to that theory's requirements, which will still certainly be considered impermissible to violate. And even in a world without poverty, the requirements themselves of maximizing act consequentialism do not seem to become any less time and energy consuming.

20 This is not, by the way, to say that the commonsense theorist cannot appeal to intuitions about how demanding a moral theory or principle is in a given case, world, or sufficiently small set of worlds. My point is just that none of us has sufficient information to have reliable intuitions (if any) regarding how demanding a moral theory is from the perspective of all possible worlds.

21 This of course also applies to any other moral theorist, whether a virtue ethicist, Kantian, or a satisficing consequentialist, who accepts the cost-based demandingness objection but argues that her theory's requirements are undemanding enough to avoid the objection.

Before moving on to objections, I should acknowledge that the difficulty involved in providing a modal assessment of demandingness does provide a slight reason against using such an assessment. However, this reason could be decisive only if we were comparing it to another (at least roughly) equally good method of assessment. But as I have been arguing, we have independent reasons to think that the modal method of assessment makes the most sense—reasons that the difficulty involved in using it cannot override. By analogy, any reliable way of determining the number of other planets in the universe with intelligent life is probably impossible to carry out, but this would not mean that some other, easier method (like guessing) would be better.

4. OBJECTIONS

While these arguments are far from conclusive, I would like to consider a few possible objections. First, I will consider whether my argument properly understands demandingness objections. Next, I will consider the objection that we should not take all possible worlds into account but rather only close possible worlds. Finally, I will consider an objection that commonsense morality's requirements have demandingness-related release conditions, and therefore the theory cannot become overly demanding.

Objection 1: This Argument Relies on a Distortion of Demandingness Objections

Since demandingness objections are often not very specific, one might worry about whether I have correctly understood what demandingness objections are doing in the first place. Maybe what I have been calling the "reasons-based" demandingness objection just is the demandingness objection properly understood. But Brian McElwee is not the only one to distinguish between importance and ranking-based demandingness objections on the one hand and cost-based demandingness objections on the other.[22] As mentioned in the introduction, Sarah Stroud has described and argued against the viability of arguments that attempt to "set a definite limit to morality's demands, by establishing a 'hands-off' or 'no-fly' zone which moral requirements could not penetrate."[23] Brad Hooker has also distinguished between (a) objections that some theory "*unreasonably* requires you to sacrifice your good" and (b) objections that "a plausible principle of beneficence toward strangers cannot require too big a reduction in the agent's good."[24] Hooker, moreover, has noted that "the more familiar form of

22 McElwee, "Demandingness Objections in Ethics," 88–92.

23 Stroud, "They Can't Take That Away from Me," 204.

24 Hooker, "The Demandingness Objection," 158.

the demandingness objection definitely is [b]," i.e., demandingness objections based upon the extremity of some moral theory's costs per se.[25]

Hooker's observation continues to be well supported by the literature. Matthew Braddock has characterized the demandingness objection as the view that "If a moral view demands too much of us, then it is mistaken."[26] David Sobel also seemed to have in mind the view that extreme demands per se undermine a theory's plausibility when he wrote about claims that consequentialism "asks too much of us to be a plausible ethical theory."[27] The predominance of the cost-based view might come as a surprise to some. After all, as proponents of the reasons-based view may ask, how could a moral theory be too demanding if its demands were justified? The unpopularity of the reasons-based view might seem less surprising, however, when we keep in mind that this is also the view of the maximizing act utilitarian and Shelly Kagan's "extremist," who promotes maximizing the good within deontological constraints.[28] The reasons-based view of demandingness, which does not in principle take any *a priori* position regarding what amount of costs per se would be excessive, consequently shares common ground with those who wish to disregard demandingness objections altogether.

Objection 2: We Should Care Only about Close Possible Worlds

Maybe I am reaching my conclusion only because I am not restricting the scope of evaluation of demandingness to *close* possible worlds. After all, there are (roughly) two modal scopes we can choose from in evaluating a theory or principle's demandingness. One option is to take *all* possible worlds into consideration. The other option is to take a more limited and proximate collection of possible worlds into consideration.[29] Restricting ourselves to close possible worlds might seem appealing to those motivated to ward off the relevance to demandingness of strange, science-fictional cases. But while there could be epistemic reasons to avoid relying on moral intuitions taken from such cases, our task here is simply to evaluate levels of demandingness of moral theories in other possible worlds with our intuitions about what counts as demanding already in place. Thus, such a motivation seems question begging insofar as it already presumes the irrelevance of such thought experiments.

25 Hooker, "The Demandingness Objection," 157.

26 Braddock, "Defusing the Demandingness Objection," 169.

27 Sobel, "The Impotence of the Demandingness Objection," 1.

28 Kagan, *The Limits of Morality*, 9–10.

29 The boundaries of such a collection can be vague, as we might, for example, give a diminishing amount of weight to facts concerning possible worlds that are farther and farther away.

Considering all possible worlds, as we have been doing, seems appealing insofar as the presumed necessity of moral truths implies that we ought to evaluate their features in the broadest modal sense. This modally broad view seems to help us account for the fact that moral theories do not suddenly become true or false. If a moral theory is true, its truth seems "safe" in certain important ways from changes in circumstances and contingencies about what the world happens to be like.[30] Nonetheless, truths about certain features of moral theories are not necessary truths. One of these features is demandingness. Even maximizing act utilitarianism, for example, is undemanding in worlds where agents have no capacity to help each other or no capacity for happiness or suffering. This is of course why we want to perform a cross-modal analysis in the first place: demandingness varies.

I do not think these considerations actually give us any reason to worry. Although demandingness varies, the principles governing the relationship between demandingness and obligation—if there are such principles—do not. The principles governing the relationship between demandingness and obligation are either going to be moral principles themselves, or they are going to be broader normative principles (such as "all things considered" norms). The purported considerations in favor of a moral theory, in a demandingness objection, either are morally insufficient to compel us to ϕ or are insufficient, all things considered, to compel us to ϕ. If we think that moral norms are always overriding, we must also think that moral considerations to ϕ can fail to compel us only if we are *morally* excused from performing them.[31] Alternatively, if we think that moral norms are not always overriding, we might think that some moral considerations are not sufficient to compel us due to nonmoral reasons.[32] In the first case, the facts that get us off the hook from overly demanding moral theories are themselves moral facts. Since we are assuming that moral facts are necessary, then it would seem natural to consider all possible worlds in evaluating demandingness. Moreover, the principles that govern the relationship between norms in an all-things-considered framework are presumably also necessary. For example, if some reason *R* compels me, all things considered, to ϕ in

30 Even principles that need to take into account what kind of creatures we are nonetheless are themselves (presumably) subsumed under more general principles concerning, for example, the moral appropriateness of taking into account the characteristics of creatures to whom a moral theory applies.

31 Recall that I do not mean "compel" here in a psychological sense. I again just mean "would compel a sufficiently rational and moral agent."

32 Compare Dorsey, "The Supererogatory, and How to Accommodate It," 355–82; Uzunova and Ferguson, "A Dilemma for Permissibility-Based Solutions to the Paradox of Supererogation," 723–31.

world *W*, then, *ceteris paribus*, *R* will also compel me to ϕ in world *W**. This is just to say that the norms of practical rationality, whether or not moral norms are overriding, seem to be the sorts of things that, if true, are necessarily true. This would imply that no matter what view of moral overridingness you take, all possible worlds—not just close ones—will be relevant to our evaluation of a theory's demandingness.

It is also worth noting, as one reviewer has pointed out, that if we may constrain our assessment of demandingness only to close possible worlds, there seems to be no reason why people in other possible worlds could not call for the same constraint. The problem with this would be that in any possible world *W* where an agent faces demanding moral requirements from some theory, and similar demandingness obtains in possible worlds close to *W*, the agent can accuse that theory of being too demanding. This is because that agent is giving more weight to her own world and those close to it. The effect of this would be that one could avoid demanding moral requirements of various moral theories in different parts of modal space just by pointing out that the moral theories are extremely demanding in those modal regions.

It might still be suggested that our evaluation of moral theories should depend upon their application to "normal" worlds, understood according to some intuitive, noncontingent, and modally invariant criteria of normality. It seems significant, however, that in the literature, it has been quite normal for philosophers to appeal to apparently abnormal possible worlds to illustrate points or to bring up counterexamples. Consider, for example, Judith Jarvis Thomson's people seeds and violinist; Robert Nozick's utility monster; J. M. E. McTaggart's longeval oysters; John Rawls's veil of ignorance; Ninian Smart's benevolent world exploder; and Michael Tooley's chemical that would endow a kitten with human-like psychology—not to mention Chateaubriand's Mandarin paradox or Plato's Ring of Gyges.[33]

Nevertheless, I am not committed to the position that all possible worlds must be relevant (although I suspect that they are). It seems implausible to me that the commonsense theorist can escape charges of extreme demandingness by shifting the focus only to close possible worlds. To illustrate my point, let us take just one example of a requirement of commonsense morality—the requirement to care for one's children. For most inhabitants of affluent Western nations today, fulfilling this requirement, while certainly having its costs, will not pose

33 Thomson, "A Defense of Abortion," 47–66; Nozick, *Anarchy, State, and Utopia*, 140; McTaggart, *The Nature of Existence*, 453; Rawls, *A Theory of Justice*, 118–92; Smart, "Negative Utilitarianism," 542–43; Tooley, "Abortion and Infanticide," 37–65; Chateaubriand, *The Genius of Christianity, or Beauties of the Christian Religion*, 188; Plato, *The Republic*, 127–29.

extreme costs. But this is largely because we happen to have more than enough food, warm places to sleep, modern medicine, and strong militaries to provide protection against violence from other nations. Many people throughout history have not been so lucky. Imagine trying to fulfill this requirement during the Taiping Rebellion as a colonized subject of the British empire, as a coastal villager in the Viking Age, or during the Great Famine of 1315–17 in northern Europe. Likewise, many people today are not so lucky. Consider the extreme costs that may be involved in ensuring the safety of one's children in modern Iraq, Ukraine, or Syria, as someone who survives on under one dollar a day, or as a refugee. As such, it does not seem difficult to consider how in many close possible worlds, the requirements of commonsense morality may be extremely costly to fulfill. Those of us who feel the requirements of commonsense morality as being rather gentle may just be those lucky enough to live in very rare circumstances of affluence and safety. Perhaps the lucky ones among us have gotten so used to their circumstances that they sometimes imagine that commonsense morality is undemanding. Nonetheless, how demanding a theory is will not just depend on what that theory is like; it will also depend on what the world is like. With this in mind, we have little reason to assume that commonsense morality will be generally undemanding even in close possible worlds.

Objection 3: Commonsense Moral Obligations Have Release Conditions

Imagine that I promise to meet you for coffee but break my leg on the way to the cafe. Maybe I owe you an apology, but I am certainly released from my obligation to meet you. So, the commonsense theorist might object, it would be ridiculous to claim that my promissory obligation to meet you for coffee is too demanding because it would be too costly for me to meet you for lunch (instead of rushing to the hospital) if I broke my leg. This objection, however, relies on the assumption that all commonsense obligations have such release conditions. This is not the case. There are many commonsense obligations that remain in place even in the face of extreme costs. For example, imagine that I am a fighter in the French Resistance during World War II. I have an important mission to deliver secret information that will save many people's lives. However, I was captured by Nazis. The Nazis offer to release me if I agree to deliver false information instead—information that will kill many people. Otherwise, they will torture and kill me. Seemingly, I am obliged either to refuse this offer outright or to accept it but then betray the Nazis and deliver the real information instead. This obligation holds even if I know that I will be tortured and killed afterwards.[34]

34 Compare Stroud, "They Can't Take That Away from Me," 213–15. That my judgment here is commonsensical is perhaps supported by the fact that duress defenses are not accepted in many legal systems in cases of homicide. See, for example, "Defenses Based on Choice."

My case of the village guards serves as another example of a requirement that remains commonsensical regardless of the costs that an agent faces. After all, the guard agreed to take on the risk, usually benefits from the agreement, and would be letting many die if she did not blow the horn. If you think that it would be okay for the guard to just run and hide when attackers are coming, consider another example. Perhaps you are at the airport, and terrorists attack. You are between your helpless child and the door. You can see a terrorist start to approach. Presumably, you are obliged to collect your child before running away, even if doing so puts your life at a greater risk. Likewise, if your mother is hit by a car and paralyzed but still capable of living a good life, then you will likely be obliged to help take care of her (at least if no one else can), even if doing so is very costly to you. Finally, consider a case loosely based on the Chernobyl divers Alexei Ananenko, Valeri Bespalov, and Boris Baranov. Imagine that in order to prevent a disaster that could kill millions, a nuclear power plant worker needs to enter the plant during a meltdown to flip a certain switch. Only this worker can find the switch, which is why the responsibility falls on him. He could, if he wanted, instead choose to escape to safety before the plant explodes. It seems plausible that the worker really is required to take on this task even if it means a high chance of death. It might be argued that the very fact that we would consider this worker heroic for flipping the switch shows his action to be supererogatory rather than required. However, it is more plausible that this positive evaluation of the worker rests on a recognition of his virtues rather than on the supererogatory nature of his actions.[35] Although heroic actions are often supererogatory, these different evaluative categories are not likely to map perfectly onto each other. Thus, commonsensical obligations do not necessarily have release conditions that kick in just when costs become extreme.[36] Therefore, the "release conditions" objection also seems to fail.

In defending the cost-based demandingness objection, however, one reviewer has offered a final example that we should consider here. We might take this to be an argument about commonsensical release conditions as well. Imagine that the villagers have an agreement, upon which their lives depend, to not reveal the village's location to the enemies if they are captured on the road.

35 Compare Carbonell, "Differential Demands," 427–38.

36 One condition that might suffice for release where extreme costs would be incurred by fulfilling a comparatively trivial obligation seemingly comes from the reasons-based rather than the cost-based demandingness objection. This condition also seems to cover certain cases of over-subscription in which, in trying to discharge my duty to ϕ, I do so in such a way that makes me fail to ϕ, at least in terms of the salient description of my actions. However, there are many cases that do not fulfill such conditions. See, for example, Feinberg, "Supererogation and Rules," 276–88.

Despite the extreme stakes, a villager cannot be expected to watch his family get tortured and killed in front of him. And, the reviewer suggests, this seems true no matter what the stakes are; one cannot be expected to make this sacrifice even when protecting nuclear launch codes that would destroy civilization. Our reactions to such a case, however, may not come from intuitions about cost but instead about the limits of what we may permissibly do (or allow to be done) to others, especially close people to whom we have special relationships, in order to promote the greater good. Insofar as we would refrain from blaming the villager or soldier protecting nuclear launch codes for giving up information to save his family (assuming that doing otherwise was even psychologically possible), this plausibly is due more to our normative attitudes regarding loving relationships than to our attitudes about costs. It would after all seem very odd for the soldier or villager, if he did decide to watch his family get tortured, to bemoan the harm this has done to him rather than the harm it has done to his family. In any case, it is much easier to imagine considering someone a moral failure for giving up the nuclear launch codes and dooming civilization while only himself getting tortured. This is not, however, to imply that such a failure must warrant blame. A failure to do one's duties in such a case might be excused without having to be justified.[37] This example should remind us that even when there seem to be genuine release conditions from an obligation, it remains important to distinguish cases in which a failure to do one's duty is merely excused from cases in which there is no duty and thus no failure at all.

5. CONCLUSION

The cost-based demandingness objection needs to be aimed at the general applications of a moral theory in order to avoid allowing someone to raise the objection just when she happens to face heavy costs. This implies that cost-based demandingness objections are objections about the possible application of moral theories. But it seems possible for commonsense morality to likewise become extremely demanding. The commonsense moral theorist might reply that a moral theory needs to be judged not in terms of how it applies to particular worlds but rather across all possible worlds. This seems true, but since it does not seem possible to judge the demandingness of commonsense morality

37 One reviewer has pointed out that a theory might be less demanding insofar as one is less blameworthy for failure (presumably by the theory's own lights) to meet its requirements. This is an interesting suggestion, but it would risk undermining many of the demandingness objections that have been made against utilitarianism. After all, the prototypically utilitarian view on blame is a merely instrumental one. See, for example, Sidgwick, *The Methods of Ethics*.

across all possible worlds, we are left worrying that there is no reason to think that the cost-based demandingness objection could not also successfully apply to commonsense morality. Fulfilling the requirements of commonsense morality may also be extremely demanding in many possible worlds (as could Kantianism, virtue ethics, or satisficing consequentialism). These considerations give us strong reasons to worry about the viability of the cost-based demandingness objection. Moreover, they give the consequentialist a way to reply to the commonsense theorist that does not rely on the analogousness of moral requirements, restrictions, and permissions. Finally, these considerations suggest that it is worth thinking about what other features of moral theories might, despite first appearances, require a modal analysis.

It could turn out that consequentialism generates costly requirements that agents do not have sufficient reason to follow. If a moral principle or theory did have this result in any particular instance, this would be a problem for that theory, calling for revision or further qualification. It will not suffice, however, simply to say that consequentialism crosses a threshold of acceptable costliness. For readers who are willing to forgo cost-based demandingness objections, this result can hopefully lead to a fruitful refocus towards reasons-based demandingness objections. This means shifting our concern away from costs to agents per se and towards whether the reasons that the agents have for incurring those costs are sufficient.[38]

University of Colorado at Boulder
kyle.york@colorado.edu

REFERENCES

Ackeren, Marcel van. "How Morality Becomes Demanding: Cost vs. Difficulty and Restriction." *International Journal of Philosophical Studies* 26, no. 3 (2018): 315–34.

Ashford, Elizabeth. "The Demandingness of Scanlon's Contractualism." *Ethics*

38 I am very grateful to following people for their helpful feedback, comments, and discussions: Spencer Case, Shamik Chakravarty, Iskra Fileva, Chris Heathwood, Michael Huemer, Alastair Norcross, Merily Salura, Brian Talbot, the participants of the sixteenth Estonian Annual Philosophy Conference, and the *Journal of Ethics and Social Philosophy*'s anonymous reviewers. Thanks also to Lingnan University and the Hong Kong Research Grants Council as well as to the College of Arts and Sciences at the University of Colorado at Boulder and the American Philosophical Association's Central Division. A special thanks to Derek Baker for all of his guidance.

113, no. 2 (January 2003): 273–302.
Braddock, Matthew. "Defusing the Demandingness Objection: Unreliable Intuitions." *Journal of Social Philosophy* 44, no. 2 (Summer 2013): 169–91.
Carbonell, Vanessa. "Differential Demands." In *The Limits of Moral Obligation: Moral Demandingness and Ought Implies Can*, edited by Marcel van Ackeren and Kuhler Michael, 427–38. New York and London: Routledge, 2018.
Chappell, Richard Yetter. "Willpower Satisficing." *Noûs* 53, no. 2 (June 2019): 251–65.
Chateaubriand, François-René de. *The Genius of Christianity, or Beauties of the Christian Religion*. Baltimore: J. Murphy and Co., 1884.
"Defenses Based on Choice." In *Criminal Law*. University of Minnesota Libraries, 2015. https://open.lib.umn.edu/criminallaw/chapter/5-4-defenses-based-on-choice/.
De Lazari-Radek, Katarzyna, and Peter Singer. "How Much More Demanding Is Utilitarianism Than Common Sense Morality?" *Revue Internationale De Philosophie* 4, no. 299 (2013–14): 427–38.
Dorsey, Dale. "The Supererogatory, and How to Accommodate It." *Utilitas* 25, no. 3 (September 2013): 355–82.
Feinberg, Joel. "Supererogation and Rules." *Ethics* 71, no. 4 (July 1961): 276–88.
Fragnière, Augustin. "Climate Change and Individual Duties." *WIREs Climate Change* 7, no. 6 (November–December 2016): 798–814.
Hooker, Brad. "The Demandingness Objection." In *The Problem of Moral Demandingness: New Philosophical Essays*, edited by Timothy Chappell, 148–62. Basingstoke, UK: Palgrave Macmillan, 2009.
Kagan, Shelly. *The Limits of Morality*. Oxford: Clarendon Press, 1991.
McElwee, Brian. "Demandingness Objections in Ethics." *Philosophical Quarterly* 67, no. 266 (January 2017): 84–105.
McTaggart, John M. E. *The Nature of Existence*. Vol. 2. Cambridge: Cambridge University Press, 1927.
Murphy, Liam B. *Moral Demands in Nonideal Theory*. Oxford: Oxford University Press, 2003.
Nozick, Robert. *Anarchy, State, and Utopia.* New York: Basic Books, 1974.
Plato. *The Republic*. Edited and translated by Christopher Emlyn-Jones and William Preddy. Loeb Classical Library. Cambridge, MA: Harvard University Press, 2013.
Rawls, John. *A Theory of Justice*. Cambridge, MA: Belknap Press, 1971.
Scheffler, Samuel. *Human Morality*. Oxford: Oxford University Press, 1994.
Sidgwick, Henry. *The Methods of Ethics*. London: Macmillan, 1907.
Sin, William. "The Demandingness of Confucianism in the Case of Long-Term Caregiving." *Asian Philosophy* 23, no. 2 (2013): 166–79.

Smart, R. N. "Negative Utilitarianism." *Mind* 67, no. 268 (October 1958): 542–43.

Sobel, David. "The Impotence of the Demandingness Objection." *Philosophers' Imprint* 7 (September 2007): 1–17.

Stroud, Sarah. "They Can't Take That Away from Me: Restricting the Reach of Morality's Demands." In *Oxford Studies in Normative Ethics*, vol. 3, edited by Mark Timmons, 203–34. Oxford: Oxford University Press, 2013.

Temkin, Larry. "A 'New' Principle of Aggregation." *Philosophical Issues* 15, no. 1 (2005): 218–34.

Thomson, Judith Jarvis. "A Defense of Abortion." *Philosophy and Public Affairs* 1, no. 1 (Autumn 1971): 47–66.

Tooley, Michael. "Abortion and Infanticide." *Philosophy and Public Affairs* 2, no. 1 (Autumn 1972): 37–65.

Uzunova, Marina, and Benjamin Ferguson, "A Dilemma for Permissibility-Based Solutions to the Paradox of Supererogation." *Analysis* 80, no. 4 (October 2020): 723–31.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.2400

ON GIVING YOURSELF A SIGN

Justin Dealy

Sometimes we want a sign. We gaze heavenward hoping to spot some sign from a higher power. We scrutinize text messages for signs of affection. We peek into hot ovens hoping to see signs that our attempt at baking is going well. And so on. Consider something weird: suppose you want a sign of something you know you cannot control, and you believe you can *manufacture* that sign—specifically, you believe you can do something to make that sign happen, and it will *retain its significance* vis-à-vis the thing you cannot control. This is not normally how it works. We seldom try to manufacture signs of things we hope for but have no control over, since normally we know in advance that our best effort will only produce a dud lacking the desired significance—doctoring your email inbox will not give you a sign that your recent job application has made it to the next round of consideration, alas. But in the weird case where you can create such a sign and still have it mean what you want it to mean, do you have a practical reason to do so? I will argue on the basis of a straightforward means-end principle that the answer is yes, provided that we understand "reason" subjectively (section 2). This is intriguing in itself, but what makes it particularly noteworthy is that such reasons are grounded in a species of extrinsic desire that can exist even when you do not believe you have the means to satisfy any of your intrinsic desires—not even the intrinsic desire(s) from which the extrinsic desire derives. In other words, desires for mere signs can give rise to practical reasons that do not bottom out in moral duty or in the aim to satisfy intrinsic desire. After responding to objections aimed at undercutting my arguments for the existence of such reasons (section 3), I will argue that the reasons, though they exist, are of an inferior kind and can be trumped by reasons grounded in intrinsic desires. I finish by sketching a two-level means-end account of desire-based practical reasons (section 4).

1. SETUP AND BACKGROUND

1.1. Objective and Subjective Practical Reasons

Assertions to the effect that so-and-so has a reason to do such-and-such can be heard "objectively" or "subjectively." Bernie thinks his glass contains gin

and tonic when in fact it is filled with gasoline.[1] Bernie wants to drink gin and tonic, not gasoline—does he have any reason to take a sip? On one way of hearing the question (the "objective" way), the answer is no, he does not have a reason to take a sip. On another way of hearing the question (the "subjective" way), the answer is yes. A natural account of this is that our ordinary talk about having reasons to do things—practical reason talk—is ambiguous. There are two different sorts of practical reason our talk can be about, depending on how it is disambiguated: objective practical reasons or subjective practical reasons. Bernie has no objective practical reason to take a sip, but he does have a subjective practical reason to take a sip.

Some philosophers deny both the ambiguity and the existence of subjective practical reasons.[2] Others say that while Bernie has an "apparent" practical reason to take a sip, he in fact has no practical reason to take a sip.[3] Still others, myself included, grant the ambiguity and the existence of subjective practical reasons.[4] This paper is primarily addressed to those in the second and third camps, i.e., those who in some sense countenance subjective practical reasons.[5]

1.1.1. The Means-End Principle

I will be assuming the following:

> *Means-End Principle* (MEP): An agent *S* has a subjective practical reason, grounded in a belief *B* and desire *D*, to do option *O* iff *D* is a desire of *S*'s with content *p* and *B* is a belief of *S*'s that doing *O* is (or might be) a means to the truth of *p*.[6]

1 Williams, "Internal and External Reasons."

2 Dancy, "Response to Mark Schroeder's *Slaves of the Passions*." See also Thomson, "Imposing Risks."

3 Parfit, "Rationality and Reasons"; Parfit, *On What Matters*, vol. 1; Sylvan, "What Apparent Reasons Appear to Be" and "Respect and the Reality of Apparent Reasons."

4 Mackie, *Ethics*, 77; Joyce, *The Myth of Morality*; Schroeder, *Slaves of the Passions*, "Having Reasons," and "Getting Perspective on Objective Reasons."

5 I will often leave "practical" in "subjective practical reason" tacit.

6 I intend "grounded in" as a placeholder for the distinctive relation between subjective reasons and belief-desire complexes. This relation could be like the "in virtue of" relation familiar from the literature on metaphysical grounding, or it could be like constitution or identity.

I intend "is a means" in a broadly causal sense. In particular, the fact that *p* is or will be true if *O* is done is not sufficient for doing *O* to be a means to the truth of *p*. It is sufficient (but not necessary) for doing *O* to be a means to the truth of *p* in the intended sense that doing *O* would cause *p* to be true. It is natural, for instance, to say that my drinking coffee is a means to staying alert. This ordinary, broadly causal sense of "is a means" anchors my usage. I assume causation is transitive and irreflexive. Two notable ways for *S* doing *O* to

MEP captures a means-end view of subjective reasons that is relatively modest inasmuch as it is compatible with such reasons having sources other than desire. What it denies is the possibility of a subjective reason grounded in a belief-desire complex where the belief is not means-end.[7] On this view, if an agent wants something and believes one of their options is a means to that thing, then those facts give rise to a subjective reason for them to do that option, and that is the only kind of belief that can work together with a desire to ground a subjective reason.

MEP offers a natural explanation of Bernie's having a subjective reason to take a sip. Bernie wants to drink gin and tonic and believes taking a sip is a

---

be a means to the truth of *p* in the intended sense are (1) *S* doing *O* would increase the objective chance of *p*, and (2) *p* is the proposition that *S* does *O*.

Readers may find MEP more plausible if, on the right-hand side of the biconditional, instead of "belief," we write "justified belief" (Joyce, *The Myth of Morality*; and Gerken, "Warrant and Action"), "justified true belief" (Littlejohn, "Must We Act Only on What We Know?"), or even "piece of knowledge" (Hawthorne and Stanley, "Knowledge and Action"). If so, they should read as if this substitution has been made throughout and as if the same modification has been made in the right-hand side of the Rational Movement Condition below. Notably, readers who opt for either of the latter two substitutions have to say Bernie does not have a reason to take a sip grounded in his means-end belief and his desire for gin and tonic; they require a different sort of case to illustrate the subjective/objective reasons distinction.

7 MEP is what we might call an internalist-adjacent principle (cf. Williams, "Internal and External Reasons"; and Smith, *The Moral Problem*). But care should be taken in drawing this connection. Internalism about practical reason is sometimes understood as the view, roughly, that an agent *S* has an objective practical reason to do ϕ iff *S* would be motivated to do ϕ if *S* were fully informed and deliberating perfectly. Since MEP concerns subjective rather than objective practical reasons, the falsity of internalism as just stated does not imply (or at least, does not trivially imply) the falsity of MEP.

MEP can be fairly considered a Humean-adjacent principle as well (cf. Joyce, *The Myth of Morality*, 52–53). But similar care should be taken here. Humeanism about practical reason is sometimes understood as the view that, roughly, an agent's objective practical reasons are as a rule grounded in their desires (Schroeder, *Slaves of the Passions*). The falsity of Humeanism in this sense does not imply (at least, not trivially) the falsity of MEP.

Here is the reason for the "at least not trivially" parentheticals. Some theorists characterize subjective reasons in terms of objective reasons. For instance, Lord defends the Factoring Account ("Having Reasons and the Factoring Account"; cf. Schroeder, "Having Reasons"), on which subjective reasons are just objective reasons that one in some sense "has." A less reductive view, favored by Schroeder (*Slaves of the Passions*) and Parfit (*On What Matters*, vol. 1), says roughly that one has a subjective reason to do ϕ just in case one has a belief that if true would give one an objective reason to do ϕ. Other such "objectivist" theories have been proposed (Vogelstein, "Subjective Reasons"; Whiting, "Keep Things in Perspective"; Sylvan, "What Apparent Reasons Appear to Be"; Wodak, "Can Objectivists Account for Subjective Reasons?" and "An Objectivist's Guide to Subjective Reasons"). If such a theory is true—a question on which I take no stand—then internalist and Humean theories of objective reasons may turn out to entail MEP.

means to that. Thus, he has a subjective reason to take a sip grounded in his desire and his means-end belief.

1.1.2. The Constitutive Roles of Subjective Practical Reasons

Subjective reasons play constitutive roles vis-à-vis rationality.[8] Three such roles will be relevant. The core role has to do with rational action. It can be stated simply with the following principle:

> *Act Rationality*: An action is rational iff (and because) the subjective practical reasons to do it are not outweighed by the subjective practical reasons to do otherwise.

A second role played by subjective reasons has to do with rational criticism.[9] Roughly put, our practices of rationally criticizing agents for their actions or deliberations are constitutively bound up with perceived failures to suitably respond to subjective reasons. The following two principles capture this:

> *Act Criticizability*: An agent is rationally criticizable for doing an act *A* iff (and because) whatever subjective practical reasons they have to do *A* are outweighed by the subjective practical reasons they have to do one of their alternative options.
>
> *Deliberation Criticizability*: An agent's practical deliberation is rationally criticizable iff (and because) they failed to suitably appreciate (individually or collectively) some of the subjective practical reasons they have for or against one of their options.[10]

A third role played by subjective reasons has to do with rational motivation.[11] I will rely on the following plausible necessary condition:

> *Rational Movement Condition*: An agent is in a position to be rationally moved by a belief *B* and a desire *D* to do one of their options *O* only if they have a subjective practical reason *R*, grounded in *B* and *D*, to do *O*.[12]

8 Schroeder, *Slaves of the Passions* and "Having Reasons"; and Wodak, "Can Objectivists Account for Subjective Reasons?"

9 Schroeder, *Slaves of the Passions*; and Wodak, "Can Objectivists Account for Subjective Reasons?"

10 There might be some subjective reasons that it is suitable to ignore when deliberating. Cf. Wedgwood, "Gandalf's Solution to the Newcomb Problem."

11 Schroeder, *Slaves of the Passions* and "Having Reasons."

12 The Rational Movement Condition coheres with a plausible account of motivating reasons provided by Mark Schroeder, which states that for *R* to be an agent's motivating reason

Being rationally moved by a belief *B* and a desire *D* to do *O* is at least materially equivalent to doing *O* and having a *motivating reason* grounded in *B* and *D*. Motivating reasons are reasons that are acted on—reasons for which an agent does an action. They are not mere causes of actions. They *rationalize* actions.[13] Subjective reasons are what rationalize actions; to act and have a motivating reason is to act on a subjective reason that one has.

1.2. Desire

Subjective reasons are one main theme of this paper; the other is desire. I will assume that desires are propositional attitudes in the sense that a token desire with content *p* is essentially such that it is *satisfied* iff *p* is true. So for example, where we might ordinarily speak of Scooby wanting a club sandwich, I will understand Scooby as desiring that Scooby has a club sandwich. Having made this clear, I will sometimes state things in a way superficially inconsistent with it when doing so is convenient for presentation.

1.2.1. Intrinsic and Extrinsic Desires

To intrinsically desire something is to desire it in and of itself, for its own sake. To extrinsically desire something is to desire it for the sake of some further, logically distinct thing. Intrinsic desire is logically independent of extrinsic desire; you can have or not have an intrinsic desire for something that you extrinsically desire.

1.2.2. Extrinsic Desires Are Derivative

Extrinsic desires are derivative. They typically derive from other desires in virtue of "connecting beliefs," e.g., a belief that if an extrinsic desire is satisfied, some further desire will be satisfied.[14] The claim that extrinsic desires are derivative can be unpacked in two ways. First, extrinsic desires tend to vanish immediately if and when the beliefs in virtue of which they derive are lost—e.g., if Joe the health nut stops believing that exercise causes good health, his desire to exercise immediately vanishes. Second, extrinsic desires are—at least stereotypically—rationally explainable by appeal to further desires (i.e., those from which they derive) together with beliefs (i.e., those in virtue of which they derive). Rational explanation in this context is not merely causal. We ask questions like "Why do you want to go to med school?" and expect answers

for doing some act *A* is just for *R* to be both a subjective reason for her to do *A* and an explanatory reason for why she did *A* (*Slaves of the Passions*).

13 Davidson, "Actions, Reasons, and Causes."

14 Smith, *The Moral Problem*, 157, and "Instrumental Desires and Instrumental Rationality"; Arpaly and Schroeder, *In Praise of Desire*.

like "I want to become a doctor, and (I believe) going to med school is a means to that," which intuitively differ substantially from purely causal answers like "I was brainwashed."[15]

1.2.3. Signatory Desire

The key type of desire in this paper is a species of noninstrumental extrinsic desire that Kris McDaniel and Ben Bradley call *signatory* desire.[16] Consider Tom, who just texted his crush Judy and asked if she is into him. Tom very much wants to receive a text that reads "Yes." Tom's desire for a "yes" is not intrinsic, but neither does he believe a "yes" would be a means to anything else he wants.[17] He only wants it because it would be a sign that Judy reciprocates his feelings. His desire for the text is signatory.

The following (somewhat stipulative) definition will suffice for our purposes:

> *Signatory Desire*: A desire *D* for *p* is signatory iff there is some *q* such that *D* derives from a desire for *q* in virtue of a belief that *p*'s truth is a sign (but not a cause) of *q*'s truth.[18]

Being a sign but not a cause is a familiar notion. The failure of the bathtub to drain is a sign but not a cause of a clog in the pipe. A persistent cough is a sign but not a cause of an infection. I take being a sign but not a cause of *p* to be roughly equivalent to being a piece of evidence for but not a cause of *p*. Roughly speaking, when some *p* is a sign but not a cause of *q*, the truth of *p* implies that it is *ceteris paribus* more likely that *q*; but "more likely" in this instance does not correspond to objective physical chance in anything like the deep quantum mechanical sense (e.g., a slow drain does not imply a chance of a clog in that sense, as there being a nontrivial chance of a clog in that sense implies that it is in some deep way indeterminate or inscrutable whether there is a clog, but a slow drain never has any such implication). Rather, the relevant notion of "likely" is the familiar one used, e.g., in medicine and the special sciences,

15 Exceptions to the letter (but not the spirit) of this characterization of extrinsic desires come from cases of desires for knowledge. I will discuss these cases in section 1.2.6.

16 McDaniel and Bradley, "Desires." Intrinsic desire is often contrasted with instrumental desire, i.e., desire for something as a means to some end. But as the category of signatory desire illustrates, not all extrinsic desires are instrumental. In addition to McDaniel and Bradley's paper, see also Harman, *Explaining Value*, 128–29; and Arpaly and Schroeder, *In Praise of Desire*.

17 With the possible exception of something like knowledge or justified belief. See section 1.2.6 below.

18 McDaniel and Bradley leave out the parenthetical bit. In view of its inclusion in my definition, one might wish to think of the concept I am using as *purely signatory* desire.

when, e.g., a doctor says that cancer is more likely if certain test results turn out positive. This notion is clear enough for my purposes.

1.2.4. Desire's Causal Roles

Desire's roles in folk psychology include but are not limited to motivating and rationalizing action in combination with belief.[19] Five additional causal roles will be relevant. First, the believed satisfaction or frustration of desire is causally tied to pleasure, displeasure, and related feelings and attitudes: coming to believe a desire is or will be satisfied causes pleasure (or excitement, relief, etc.), whereas coming to believe a desire is or will be frustrated causes displeasure (or disappointment, sadness, etc.). Second, desires cause one to fantasize about their objects and dwell on what it is about their objects that is desired.[20] Third, desires direct our attention to things in our environment associated with their objects. Fourth, desires can be intensified by dwelling on vivid representations of their objects. Fifth, desires cause a distinctive phenomenological feel (a "pull") vis-à-vis their objects.

1.2.5. Idle Desire and Working Desire

Both of the following sentences are true:

1. I want a new car.
2. I want it to be the case that both (i) I have a new car and (ii) there are an even number of hairs on my head.

But only the content attributed by the first of these sentences is such that a mental state with that content plays the causal roles of desire in my psychology. (In other words, I will not be the least bit let down if I come to believe I am getting a new car but have an odd number of head hairs; I fantasize occasionally about having a new car but never about the conjunction of that and my having an even number of hairs; etc.) So we should not assume a one-to-one correspondence between an agent's desires and true attributions of desire to that agent.

Why is 2 true? Is it because I have a desire the content of which is entailed by the conjunctive content that 2 attributes? No. Say that Ralph is an American seven-year-old. It is natural to think the first but not the second of the following is true:

19 Cf. Sinhababu, "The Humean Theory of Motivation Reformulated and Defended." For methodological background, see Braddon-Mitchell and Nola, "Introducing the Canberra Plan."

20 Cf. Scanlon, *What We Owe to Each Other*, 38; Marks, "The Difference between Motivation and Desire." Note that it makes sense to reflect on "what it is about" the object of a desire that is desired only if the desire is extrinsic. An intrinsic desire's object is desired in itself!

3. Ralph wants a new toy.
4. Ralph wants it to be the case that both (i) he has a new toy and (ii) vaquitas love knafeh.[21]

It is natural to think that 4 can be false despite 3 being true because Ralph might have never heard of vaquitas or knafeh—he might not possess the required concepts.

Why then is 2 true? Here is why: the subject of 2—i.e., me—happens to believe the truth of the content attributed in 2 is a means to or a sign of the truth of the content attributed in 1, which in turn is the content of one of my desires.

What is it that prevents us from saying that I have a desire with the satisfaction condition attributed by 2? Just some causal constraint on positing psychological states. If we do not mind violating such a constraint, we can stipulate a one-to-one correspondence between true desire attributions and desires. As it turns out, doing this will be useful for us. We will simply keep in mind the distinction between the stipulated states and the states that actually play the causal roles of desire.[22] Call the stipulated states *idle desires* and the states that play the causal roles of desire *working desires*.[23] Only working desires can be intrinsic; idle desires are by nature extrinsic.

Idle desires cannot ground subjective practical reasons. MEP and other principles quantifying over desires are meant to quantify only over working desires. Henceforth, all my reference to desire will be confined to working desire, unless stated otherwise.

1.2.6. Are Signatory Desires Just Desires for Knowledge?

Perhaps signatory desires are nothing more than desires for knowledge or desires for the means to knowledge.[24] That is to say:

> *Signatory Desires Are Desires for Knowledge* (SDK): To have a signatory desire for the truth of *p* is either to desire to know *p* or to desire *p* because one believes *p* is a means to obtaining further desired knowledge.

21 Vaquitas are a rare species of porpoise. Knafeh is a type of cheesy Arabian cake.

22 Or rather, the states that play "enough" of the causal roles. Cf. Braddon-Mitchell and Nola, "Introducing the Canberra Plan," sec. 1.

23 Cf. Marks, "The Difference between Motivation and Desire" on a separate but similar distinction between "formal" and "genuine" desire.

24 Or desires for justified belief or for the possession of evidence, etc. I focus on knowledge, but the points generalize as far as I can tell. Thank you to an anonymous referee for raising this issue.

For instance, perhaps strictly speaking Tom does not want to receive a "yes" text, or perhaps the reason Tom wants a "yes" text to appear on his phone is not because it would be a sign. Rather, perhaps what he really wants is to *know* whether (or that) a "yes" has appeared. Or perhaps he wants a "yes" because he believes that would be a means to his knowing whether (or that) Judy likes him. SDK says desires like these are all that is really going on in cases of signatory desire. I think SDK is false for two reasons.

The first reason is just that we can imagine cases where someone wants a sign but clearly does not want knowledge of the sign. Examples that come to mind are weird but intelligible. Patrick the alligator lover believes an Amazing Predictor spared an alligator's life iff she predicted rain in Bangkok tomorrow and that Patrick will never know about it.[25] Patrick hopes for rain in Bangkok tomorrow, but he does not want to know about it.

A second reason to think SDK is false is that it is clearly possible to desire a sign despite not believing it to be a means to one's obtaining any desired knowledge. Tom learns that his incoming text messages have been encrypted by hackers. There is no hope of ever breaking the encryption; the messages still exist but are unreadable. So Tom is sure that if he has received a "yes" text from Judy, he cannot know of it (unless and until he sees Judy again and musters the courage to ask her). Hence, Tom does not believe that a "yes" text, if one exists in his now hopelessly encrypted account, is a means to his knowing about Judy's feelings. Still, he hopes that it is there, since if it is, that means it is likely that Judy is into him.

A third reason to think SDK is false is that agents can give up hope of knowing whether a sign or what it signifies obtains without giving up hope that the sign obtains. Damon is a dying man who wants it to be the case that Atlantis existed. He therefore wants it to be the case that Atlantean artifacts exist somewhere at the bottom of the Atlantic. Damon has given up hope of ever knowing whether there are Atlantean artifacts or indeed of ever knowing whether Atlantis existed, but he has not given up hope that there are such artifacts.[26] If one can give up hope of Desire *A* being satisfied while not giving up hope of Desire *B* being satisfied, then it seems Desire *A* and Desire *B* must be distinct. Further, if one can cease believing a sign is a means to knowing anything one

25 All the "Amazing Predictors" in this paper are, unless stated otherwise, known by the agents in the cases to be almost but not quite infallible predictors of future events.

26 Interlocutor: Why would Damon still want there to be artifacts, having given up hope of ever knowing of their existence? Me: Because he thinks that if there are artifacts, then it is likely *ceteris paribus* that Atlantis existed. Damon wants Atlantis to have existed, and so naturally he wants the former existence of Atlantis to be likely given the evidence at the bottom of the ocean. This is entirely compatible with his believing he does not and will not ever know what evidence is down there.

wants to know and nevertheless still want the sign, then desire for a sign must not necessarily be instrumentally aimed at obtaining knowledge.

I am persuaded by cases like Patrick's and Tom's and Damon's, but as far as the arguments I am going to give go, what is essential is not that SDK be false but rather that in having a signatory desire for something, one need be neither intrinsically desiring that thing nor desiring it as means to some further intrinsic desire. So if SDK turned out to be true, would that throw a wrench into my arguments? There are two ways it could. First, it could be that any time one believes an option is a means to desired knowledge, one also believes it is a means to some intrinsic desire. Second, it could be that signs one desires are desired as means to the satisfaction of intrinsic desire any time they are believed not to be means to the satisfaction of any desires aside from desires for knowledge. I will consider these possibilities in order.

Suppose Terry's only intrinsic desire is to love and to be (and to have been) loved. Terry desires to know whether his deceased relative really loved him. He believes he can gain this knowledge if he reads a special letter they wrote to him. It makes sense to suppose that Terry does not believe reading the letter to be a means to loving or to being loved—maybe Terry believes himself to not be the sort of person to behave any differently given such knowledge, or perhaps Terry is an aging recluse who foresees no future loving relationships, regardless. Cases like this show that believing that *x* is a means to desired knowledge does not imply believing that *x* is a means to intrinsic desire.

It is obvious that desire for knowledge is not always intrinsic; knowledge is frequently desired purely as a means to further desired things. But what about when the knowledge one desires is not believed to be a means to any other desire's satisfaction aside from desires for further knowledge—must any of *those* desires be intrinsic? Consider a concrete example. Suppose we ask Yancy why he wants to know whether canaries have been dying in the coal mine. He answers, "Because that would be a sign of dangerous carbon monoxide levels, and I want to know whether the miners are getting poisoned." Note that Yancy can desire the knowledge of canary deaths even if he knows he can do nothing to prevent miners getting carbon monoxide poisoning. Now, assuming Yancy has no other reasons for caring about canaries, the intensity of his desire to know whether canaries are dying will vary directly with (a) his concern for carbon monoxide levels in the mine and (b) his confidence that canary deaths are a sign of such levels. If he stops believing canary deaths are a sign of dangerous conditions for the miners, he will immediately stop caring to know about canary deaths. Further, Yancy's desire for the knowledge of canary deaths will be rationally explainable via his concern over carbon monoxide levels and his belief that canary deaths are a sign that such levels are high. Hence, Yancy's desire for the canary knowledge

is extrinsic. Analogous things can be said for Yancy's desire to know whether the miners are getting poisoned. That desire for knowledge will vary directly with his concern for the miners' lives and well-being, will immediately vanish if he loses that concern, and will be rationally explainable via that concern. Hence, Yancy's desire to know whether the miners are safe is extrinsic as well. So desires to know whether signs obtain can be extrinsic, even when such knowledge is believed not to be a means to satisfying any desires aside from desires for knowledge; and similarly, all the desires for knowledge that may happen to ground such extrinsic desires for knowledge of signs can be extrinsic as well.

Two general observations are in order here. First, it is notable that the way desires to know whether signs obtain derive seems structurally different from the way desires for means derive. In the latter case, one's belief is that the thing desired stands in a means-end relation to some further thing one desires; but in the former case, one's belief is not that the knowledge one desires stands in the is-a-sign-of relation to some further thing one desires but rather that the content of the knowledge desired stands in that relation. Second, it seems desire for knowledge can be extrinsic without deriving in virtue of any belief. If one wants a state of affairs to obtain, one can, in virtue of that alone, desire to know whether (or that) it obtains. Desire for such knowledge counts as extrinsic insofar as it waxes and wanes with and is entirely rationally explained by one's concern for the underlying state of affairs.

2. Signatory Desires Ground Subjective Practical Reasons

The last section familiarized us with signatory desires, subjective practical reasons, and the means-end principle. In this section I connect the dots. Signatory desire together with means-end belief can ground a subjective reason to do something, and this can be so despite one's having no subjective reason grounded in intrinsic desire or moral duty to do the thing—or so I will argue.

2.1. The Case of Donny

Donny wants to bowl a strike. Naturally, he thinks bowling is a means to that end. But Donny's case is unusual. His reason for wanting to bowl a strike is not that he thinks it would be pleasant. In fact, he fully expects bowling a strike would overwhelm him with restless doubts. He expects he would doubt whether the strike really happened or was some cruel hoax; he is sure his resulting anxiety would keep him awake all night. Donny is a nervous, fitful man; he trembles at great turns of fate—especially when they involve money. What is Donny's deal? It is this: Donny's reason for wanting to bowl a strike is that he believes an Amazing Predictor mailed him one million dollars iff she predicted he would bowl a

strike tonight. The money, he thinks, will arrive tomorrow, if ever. So on this night, Donny is cold to the usual appeal of bowling and does not think it will be a means to anything he desires besides the sign.[27] Donny's desire to bowl a strike is signatory. He believes bowling is a means to a strike, and his desire for a strike derives from his desire for wealth in virtue of his belief that a strike would be a sign (not a cause) that he will soon be a millionaire. In particular, Donny believes bowling is a means to a sign—one that will retain its significance as a sign—of the million dollars. In other words, Donny believes that even if he causally intervenes in the production of the sign, it will still be a sign of the million dollars.

2.2. Two Main Arguments

In this section I give two arguments for the claim that Donny has a subjective reason to bowl grounded in his means-end belief and his signatory desire to bowl a strike. Some premises have been given specific titles in parentheses.

Argument 1

P1. MEP is true.
P2. Donny has a belief B^* that bowling is a means to bowling a strike.
P3. Donny has a signatory desire D^* to bowl a strike. (*Desire to Bowl*)
P4. MEP applies to the pair consisting of B^* and D^*. (*Applies to D^**)
C1. Donny has a subjective reason to bowl grounded in his means-end belief and his signatory desire to bowl a strike. [P1, P2, and P3]

How can this argument be resisted? An ad hoc rejection of MEP would be implausible. A more serious objection would target Applies to D^* by seeking to limit application of MEP to non-signatory desires. We will consider such objections in section 3.

Argument 2

P5. Prior to making his decision, Donny is in a position to be rationally moved to bowl by his means-end belief together with his desire to bowl a strike. (*In Position to Be Rationally Moved*)

27 Interlocutor: Does Donny want the strike as a means to know that the million dollars has been sent? Me: No, he would rather it just show up in his bank account; he is sure such knowledge will only elevate his anticipation and worry over its safe arrival. Interlocutor: Is this a key assumption? Me: Not really. If Donny did want knowledge of the million dollars, his desire for that would be extrinsic as well (cf. section 1.2.6). Interlocutor: If he is not after knowledge, why does he want the sign? Me: He wants it to be likely that the million dollars is on the way (cf. note 26). Interlocutor: If he thinks he can "make it more likely," does he not think he can causally influence the million dollars? Me: He believes the likelihood would be purely evidential (cf. section 1.2.3). He thinks a strike would be a symptom of a common cause or else a mere statistical correlate of the million dollars being sent.

P6. The Rational Movement Condition is true.
C1. Donny has a subjective reason to bowl grounded in his means-end belief and his desire to bowl a strike. [P5 and P6]

In support of In Position to Be Rationally Moved, I would say that Donny is intuitively in a position to bowl and rationalize his choice by saying, "I want to bowl a strike tonight, and bowling is a way to make that happen!" Denying the Rational Movement Condition outright just to avoid the conclusion would, like a blanket rejection of MEP, seem implausible and ad hoc. It seems therefore that the best way to resist this argument is to give some reason for denying In Position to Be Rationally Moved. In section 3 we will look at attempts to motivate such a move.

Stepping back, what exactly does C1 amount to? Well, if C1 holds, Donny's signatory desire D^* to bowl a strike participates in two distinct dependence relations. One is the relation between Donny's subjective practical reason to bowl and the pair consisting of D^* and his means-end belief. The other is the relation between D^* and Donny's desire for future wealth, a relation that holds in virtue of his belief that bowling a strike tonight is a sign—but not a cause—of an impending million dollars. C1 is about the first of these two relations. The second relation is one that holds between desires in virtue of connecting beliefs. It does not take a subjective practical reason as a relatum.

2.3. Supporting Arguments: MEP Applies to Signatory Desires

In this section I defend Argument 1's fourth premise (i.e., Applies to D^*) with a set of three arguments. Once again, some premises have been given specific titles in parentheses.

Argument 3
P7. Unless Donny has some subjective practical reason to stay home and not bowl, Donny is rationally criticizable if he chooses not to bowl. (*Criticizable Choice*)
P8. Act Criticizability is true.
C2. Donny has a subjective reason to bowl. [P7 and P8]

Argument 4
P9. Donny is rationally criticizable if when deliberating over whether to bowl, he disregards his belief that bowling is a means to bowling a strike. (*Criticizable Deliberation*)
P10. Deliberation Criticizability is true.
C2. Donny has a subjective reason to bowl. [P9 and P10]

Argument 5

P11. The best explanation for Donny's having a subjective reason to bowl is that Applies to D^* is true. (*Best Explanation for a Reason to Bowl*)

C3. Applies to D^* is true. [P11 and C2]

How can these arguments be resisted? There are two main ways. The first way to respond is to motivate a denial of Criticizable Choice and Criticizable Deliberation, and then reject Best Explanation for a Reason to Bowl as falsely presupposing the existence of a subjective practical reason to bowl. The second way to respond is to accept Criticizable Choice and Criticizable Deliberation and motivate a denial of Best Explanation for a Reason to Bowl by supplying a competing explanation for Donny's having a subjective reason to bowl. We will look at both of these routes in section 3.

2.4. Summing Up

Those are my arguments about Donny. I take them to generalize. That is, I assume that Donny's case points the way to all sorts of other possible cases with the same basic structure. Hence, I take these arguments to show that signatory desires, together with associated means-end beliefs, can ground subjective practical reasons.

3. OBJECTIONS AND REPLIES

3.1. Objection 1: Refusing to Satisfy Signatory Desire Is Uncriticizable

This objection asserts that refusing to try to satisfy signatory desires is rationally uncriticizable. It directly targets Criticizable Choice and Criticizable Deliberation in Arguments 3 and 4. Best Explanation for a Reason to Bowl in Argument 5 is rejected on the grounds that it falsely presupposes that Donny has a reason to bowl. A proponent of this objection would find MEP implausible unless restricted to non-signatory desires and thus would reject Applies to D^* in Argument 1. This objection does not challenge Argument 2; so even if it is successful, its proponents have more work to do to rebut C1.

3.1.1. Reply

If you want p to be true and believe that the only way you can cause p to be true is by doing O, and you have no subjective practical reason to do anything else, then if you choose not to do O, it is just obvious that you can be rationally criticized. "Don't you want this? Don't you believe that the only way you can make it happen is to do O? Instead, you've done something you have no reason

to do! What gives?"[28] It is inadequate to reply: "I do want *p* to be true, but the only reason I want it to be true is that I want *q* to be true and I believe that the truth of *p* is a sign—but not a cause—of the truth of *q*." This is inadequate because the reason for wanting *p* to be true is irrelevant. We can reply, "Yes, that is a perfectly reasonable basis for wanting *p* to be true. What is your point? Why are not you doing the thing you believe can cause *p* to be true, given that you have no reason to do anything else?" Perhaps they might protest, "It is the *type* of basis the desire has. It does not derive from another desire in virtue of a means-end belief but derives in virtue of a belief about being a sign. *That* is the excuse." I reject this invidious distinction. What is disqualifying about the distinctive basis of a signatory desire for *p*? Does the agent have a reason for desiring *p*? Yes. Is the reason a perfectly acceptable one? Yes. Is it granted, therefore that the desire is not an irrational one? Yes.[29] I fail to see any excuse for not doing the thing they think will cause *p*'s truth, given that they have no reason to do anything else.

3.2. Objection 2: Signatory Desires Are Inherently Irrational

According to this objection, signatory desires are inherently irrational *qua* desires. Hence, it is irrational to act with the aim of satisfying them. In other words, according to this objection, mere signs are always as a rule irrational *to want*, and therefore it is always as a rule irrational to act with the aim of satisfying desires for mere signs. Like the previous objection, this objection challenges Applies to D^*; a proponent of it would likely want to restrict MEP to nonsignatory desires. But unlike the last objection, this objection also challenges Argument 2 by targeting In Position to Be Rationally Moved: it is always irrational, according to this objection, to be moved by a signatory desire (together with a means-end belief) since signatory desires are themselves always irrational. Hence, *contra* In Position to Be Rationally Moved, Donny was never in a position to be rationally moved by his signatory desire.[30]

3.2.1. Reply

Recall Tom from section 1.2.3. Is there really something inherently irrational about him wanting to receive a "yes" text? He seems in a position to give a quite ordinary and reasonable account of his desire. Receiving a "yes" text would be a

28 Is the sense of irrationality to be explained by the fact that the person has done something they have no reason to do? No, since this is not inherently irrational; it is rationally permissible to do what one has no reason to do if all of one's alternatives are such that one has no reason to do them either.

29 Well, maybe not so fast. See Objection 2.

30 That is, he may have been in a position to be moved but not *rationally* moved.

sign that something he hopes to be the case is the case: Judy is attracted to him. What is irrational about this? Perhaps the idea is that insofar as Tom really just wants knowledge, his desire is rational, and the charge of irrationality comes in only insofar as his desire for the "yes" text is a desire for neither knowledge nor a means to knowledge but purely a desire for a sign. This possibility was discussed in section 1.2.6. There I gave a version of the case where Tom's text messages have been hopelessly encrypted by a hacker, and yet he still wants there to be a "yes" text among them, despite his inability to know if it is there and its uselessness to him vis-à-vis knowing whether Judy likes him. Is it really irrational to want this inaccessible evidence to exist? I do not have space to delve too deeply into the topic of the rationality of extrinsic desire, but such delving seems unnecessary here. Desires for mere signs are at least not obviously rationally problematic—they clearly do not essentially involve desiring impossible states of affairs or desiring states of affairs that are incompatible with what one desires intrinsically. Moreover, such desires, while perhaps obscure, are not a purely hypothetical curiosity. Consider wanting certain medical tests to be negative, as that would show your newborn is healthy. To be sure, this desire would normally be associated with desires for knowledge, but even if one for some reason had little or no hope of obtaining knowledge about or via the tests, one's desire for the tests to be negative would naturally persist. Consider the following. Take any state of affairs *A* you currently desire and imagine that the likely cause of *A* would also cause some separate state of affairs *B*, where *B* would not itself be a cause of *A*. Alternatively, imagine that if *A* were to obtain, it would likely cause *B*, where *B* would be otherwise unlikely to obtain. Ask yourself, would you not naturally hope *B* obtains in either of these sorts of case, even if you knew you had no chance of knowing whether it does? Considerations such as these seem to show that at the very least, it is not the defender of the rationality of signatory desires who bears the burden of proof. Let us consider, then, how my opponent might defend their doctrine.

Prevailing accounts of rational desire come in four flavors. On content-based accounts, a desire is rational iff it has the right sort of content—e.g., it is a desire for the good, for something believed to be good, or for something there is objective reason to want.[31] On deliberative accounts, desires are rational iff they are produced, controlled, or sanctioned by an actual or idealized process of rational deliberation.[32] On information-based accounts, a desire is rational iff it is not based on false beliefs, and/or it would be maintained even

31 Anscombe, *Intention*; Audi, *The Architecture of Reason*; and Parfit, *On What Matters*, vol. 1.

32 Rawls, *A Theory of Justice*, 407–24; Brandt, *A Theory of the Good and the Right*; and Smith, *The Moral Problem*, 157–61.

if the desirer were suitably well-informed.[33] On coherence accounts, a desire is rational iff it suitably coheres with the desirer's beliefs, desires, and/or self-conception.[34] Let us look at each of these in turn.

The sort of account that seems most apt to deliver the result that signatory desires are all as a rule irrational is a content-based account that classifies a desire as irrational if its satisfaction is not as a matter of fact objectively good for the agent. One might assert:

> *The Means to My Intrinsic Desire Is My Good* (MIDIG): The objective good for an agent consists in (α) the satisfaction of their intrinsic desires and (β) any proposition whose truth would cause the satisfaction of (one or more of) their intrinsic desires.

Next one might claim that since satisfaction of signatory desires belongs to neither α nor β, it is never objectively good for an agent. Hence, on this account of rational desire, such desires are as a rule irrational. There are two problems with this.

The first problem is that the sort of content-based account just sketched seems promising as an account of when there is objective reason for an agent to desire something but does not seem promising as an account of when there is subjective reason for an agent to desire something. Recall Bernie, who thinks his glass is full of gin and tonic when in fact it is full of gasoline. Is Bernie's desire to take a sip from his glass rational? Here we encounter the familiar ambiguity. Bernie's desire seems unsupported by objective reason but nevertheless subjectively rational. Since it is surely desires' subjective rationality that would be relevant to whether they can ground subjective reasons to act so as to satisfy them, a content-based account seems like the wrong sort of account to oppose the rationality of signatory desires. One could respond by saying my opponent should instead adopt a content-based account that says a desire is irrational iff its content is not believed to be objectively good for the agent. This, together with MIDIG, handles Bernie's case but leads us to our second problem.

The second problem is that MIDIG begs the question. Why should we think it is not objectively good for you to get a sign you want? Dialectically speaking, it is hard to see why mere causes of intrinsic desire satisfaction can be part of an agent's good, but mere signs of intrinsic desire satisfaction cannot be. We could toss out condition β, but then by parallel reasoning, we wind up saying instrumental desire is inherently irrational.

33 Hume, *A Treatise of Human Nature*; Brandt, *A Theory of the Good and the Right*; and Savulescu, "Rational Desires and the Limitation of Life-Sustaining Treatment."

34 Velleman, *Practical Reflection*; Smith, "In Defense of *The Moral Problem*" and "Instrumental Desires and Instrumental Rationality"; and Verdejo, "Norms for Pure Desire."

If content-based accounts do not support the claim that signatory desires are as a rule irrational, can some combination of the other three types of account do so? On the contrary, these accounts seem quite friendly to the rationality of such desires. It would be odd to deny that an informed and ideally reflective agent can desire a sign of something else they want and that that desire might cohere with the agent's beliefs and other desires.[35]

Finally, it bears mentioning that even if signatory desires are irrational, that alone does not trivially imply that seeking to bring about their satisfaction would be irrational. That inference is mediated by the significant assumption that it is as a rule *practically* irrational to seek to satisfy irrational *desires*. But it is one thing for a desire to be irrational; it is another for behavior aimed at satisfying that desire to be irrational. I am in favor of sharply separating these two domains of rationality. An agent whose actions are sanctioned by the beliefs they have about how to best effect their desires' satisfaction seems to me one whose behavior makes rational sense, even if the desires they aim to satisfy are irrational.[36] Such a position seems in keeping with a broadly Humean perspective on rationality.

3.3. *Objection 3: Do You Really Want What You Extrinsically Want?*

This objection claims that when pressed, we admit that we do not *really* want the things that we only extrinsically want. Therefore, extrinsic desires do not exist.[37] For instance, if we press Donny by asking, "Is bowling a strike *really* what you want?" it would be natural for him to respond, "Well, no. All I *really* want is wealth." So perhaps Donny's desire for a sign is not a real desire. If that is right, I was confused when I stipulated Desire to Bowl (i.e., P3), and Argument 1 fails for the simple reason that Donny does not want to bowl a strike—indeed he could not desire to bowl a strike unless he intrinsically desired to bowl a

35 Interlocutor: Wouldn't being *sufficiently* informed mean not needing or wanting the sign? Me: If we understand "sufficient information" in that strong a way, it will tend to make instrumental desires irrational. Suppose I think Pete's attendance will cause the party to be fun, and hence I want Pete to attend. Would my desire for Pete's attendance survive my being sufficiently informed if that meant knowing whether Pete will attend? Interlocutor: Yes, it would. If you learn he will not be there, you will wish he were, and if you learn he will be there, you will be glad. Me: The same can be said about signs. Suppose I want my X-ray to be clear because I believe that would be a sign I am cancer free. If I learn that it is clear, I will be glad, and if I learn it is not, I will wish it were.

36 One might protest that for a desire to be irrational just is for it to be inherently irrational to seek to satisfy. This reduces Objection 2 to the nakedly question-begging claim that signatory desires are irrational to seek to satisfy.

37 Finlay, "Responding to Normativity." See also Marks, "The Difference between Motivation and Desire."

strike. Argument 2 fails because In Position to Be Rationally Moved is false—Donny cannot be in a position to be moved by a desire that is not real.

3.3.1. Reply

It is undeniable that sometimes people say they want things only to admit under pressure that they do not *really* want them. But the fact that someone admits (or is disposed to admit) under pressure that they do not *really* want something does not immediately imply that the desire does not exist, for as I will now explain, the meaning of questions like "Is that *really* something you want?" and "What do you *really* want?" is context sensitive.

Asking what someone *really* wants or whether something they profess to want is something they *really* want can be interpreted in at least three ways depending on conversational context.[38] First, it can be interpreted as asking whether a salient professed desire is a working desire. Second, it can be interpreted as asking what desire the salient professed desire proximally derives from. Third, it can be interpreted as asking whether the salient professed desire is intrinsic and/or what intrinsic desire the salient professed desire distally derives from. Thus, in some contexts you *can* truthfully say you *really* want something that you merely extrinsically desire, and in contexts where you cannot, the reason is that the context is such that "what you *really* want" just denotes what you intrinsically desire. Allow me to explain more fully.

Sometimes in the face of pressure we steadfastly affirm that we *really* want things that we only extrinsically want. If we ask Donny whether he *really* wants to bowl a strike tonight, he might say, "Absolutely!" We might continue, "But isn't there some further thing you want that explains *why* you want to bowl a strike tonight?" He might reply, "Of course, but that doesn't mean I don't really want a strike—I *really do*!" There is nothing unnatural about this. A competent speaker would not judge his assertions to be baffling or incoherent. It would seem that Donny is taking "*really* want" to denote working desire as opposed to idle desire (section 1.2.5), which is a perfectly ordinary and acceptable interpretation.[39] A real-world example might be helpful. Suppose it is the day after the 2020 US presidential election. Votes are still being counted in swing states like Michigan. I want Joe Biden to win the presidency. If he is not going to win the presidency, I do not care if he wins Michigan. His winning Michigan is desirable to me solely as a potential means to his winning the election—it is an extrinsic desire. I am being sincere when I say I really, genuinely want

38 My discussion will be nontechnical, but it fits well with the contextualist Karttunen-style semantics for questions (see Karttunen, "Syntax and Semantics of Questions") and embedded wh-complements drawn on by Stanley and Williamson, "Knowing How."

39 Cf. Marks, "The Difference between Motivation and Desire."

Biden to win Michigan. What do I mean? I take myself to mean that my desire is a working desire. I fantasize about Biden winning Michigan. My attention is drawn to news alerts about Biden and Michigan. If I learn that Biden wins Michigan, I will be excited and pleased; if I learn that he lost Michigan, I will be displeased and troubled. And so on.

We have just seen that there is one natural interpretation of "*really* want" on which one can truthfully assert that one *really* wants something that one only extrinsically wants and that Donny can, in that sense, affirm that he *really* wants to bowl a strike. For purposes of blocking the objection, this is all that is needed. But for the sake of completeness, let us consider contexts where a speaker accedes to not *really* wanting something they formerly professed to want. Are all such contexts ones in which you cannot truthfully assert that you *really* want what you only extrinsically want? I will argue that the answer is no. Such contexts come in three varieties. Let us look at them one by one.

The first kind of context is one we have already seen: it is one where the point of asking whether someone *really* wants something is to find out whether the desire is a working desire. Suppose that it is the week after the 2020 election, and someone asks me, "Hey Justin, do you want to hear Wolf Blitzer announce that Biden won Michigan?" and I answer, "Yes, please!" Then they ask, "But is hearing Wolf make the call *really* what you want?" and I reply, "Well, no, I guess not. What I *really* want is for Biden to win Michigan. I don't care who announces it." In this example, my professed desire to hear Wolf Blitzer announce a Biden win in Michigan is idle. I believe that its satisfaction would be a sign that Biden won in Michigan, but I have no working desire to hear Wolf Blitzer make the announcement—I do not fantasize about Wolf in particular making the announcement, etc. Note that in this case, I truthfully say what I *really* want is for Biden to win Michigan—but this of course is an extrinsic desire.

In the second kind of context, asking whether someone *really* wants what they profess to want is meant to discover what desire their professed desire proximally derives from. Suppose someone asks, "But is Biden winning Michigan what you *really* want? What are you *ultimately* hoping for?" In this updated context, I cannot truthfully reply by saying that what I *really* want is for Biden to win Michigan. However, I can say, "Well no, what I *really* want is a Biden presidency." If this satisfies my interlocutor, then it seems what they were after is what desire my desire for a Biden win in Michigan proximally derives from. Note that I truthfully said that what I *really* want is a Biden presidency, which is still an extrinsic desire.

In the third kind of context, the point of asking whether someone *really* desires something is to find out whether their desire is intrinsic or to find out what intrinsic desire stands at the end of the chain of desires from which the

professed desire ultimately derives. Suppose my interlocutor keeps up their pressure. Soon enough I catch on and say, "Well, I suppose what I *really* want is for there to be happiness, health, peace, and justice in human society." In the context created by the repeat questioning, it seems I cannot truthfully say I *really* want what I only extrinsically want. But note that this is not the most natural interpretation of "*really* want"; it can take a bit of nudging to convey. Perhaps this is because it can be hard to pin down one's intrinsic desires. Pressing someone to do so is liable to seem weirdly demanding. Further, note that in this sort of context, "what *S really* wants" does not denote the set of all *S*'s desires or even the set of all *S*'s working desires—it just denotes a set of *S*'s intrinsic desires.[40] Saying you do not *really* want something in such a context thus does not entail lacking a working desire for it.

3.4. Objection 4: Only Intrinsic Desires Are Rationally Relevant

The rough idea of this objection is that intrinsic desires can do all the work we need from desires in a theory of practical reason, and therefore, extrinsic desires, including signatory desires, do not ground subjective reasons. This idea can take four forms, each targeting a different subset of my premises. What all four forms have in common is a denial of Applies to *D**, of In a Position to Be Rationally Moved, and of Best Explanation for a Reason to Bowl. There are two main variants, each with two subvariants. One main variant is neutral on whether extrinsic desires are real psychological states; its distinctive assertion is that when it comes to practical reason and rational motivation, such desires are explanatory third wheels.[41] This variant's first subvariant attacks Applies to *D** by claiming that MEP needs to be restricted to intrinsic desires; it then denies Criticizable Choice and Criticizable Deliberation and rejects Best Explanation for a Reason to Bowl as having a false presupposition. Its second subvariant agrees that MEP should be restricted to intrinsic desires but adds that even when this restriction is made, MEP is still true only as a sufficient—not a necessary—condition; this is so because a subjective practical reason, according to this line of thinking, can be grounded in a pair consisting in an intrinsic desire and a belief that an action would itself be a sign that that intrinsic desire is satisfied. One would go this route if one accepts Criticizable Choice and Criticizable Deliberation and wishes to account for them without appealing to signatory desire. Those are the two subvariants of this objection's first main variant. The other main variant's distinctive feature is that it denies Desire to Bowl and

40 Maybe it denotes the set of all *S*'s intrinsic desires, or maybe it denotes the set containing the intrinsic desire(s) from which the desire *S* originally professed ultimately derives.

41 Cf. Marks, "The Difference between Motivation and Desire"; and Smith, "Instrumental Desires and Instrumental Rationality."

asserts that extrinsic desires simply do not exist; that is to say, this second main variant "Quines" extrinsic desires.[42] As with the previous variant, this variant faces a choice with respect to whether Donny is susceptible to rational criticism. On the one hand, the Quiner of extrinsic desires might reject Donny's susceptibility to rational criticism. If they go this route, there is no need for them to restrict MEP to intrinsic desires—that sort of move would be called for only if there might be an extrinsic desire that Donny takes bowling to be a means to, and the Quiner says such desires do not exist. On the other hand, if the Quiner agrees Donny is susceptible to criticism, then by Act Criticizability, there must be a subjective reason to bowl, and MEP must be denied as a necessary condition; it is again natural to deny Best Explanation for a Reason to Bowl and offer in its place the sort of non-means-end-based explanation mentioned above. Regardless of which of the four forms just summarized this objection takes, it will deny In Position to Be Rationally Moved on the grounds that only intrinsic desires can rationally move people.

Table 1. Summary of the Four Variants of Objection 4

	Neutral on extrinsic Desires	Extrinsic desires do not exist
Donny *is* susceptible to rational criticism	Denies P1, P4, P5, and P11 (Variant 1A)	Denies P1, P3, P4, P5, and P11 (Variant 2A)
Donny is *not* susceptible to rational criticism	Denies P4, P5, P7, P9, and P11 (Variant 1B)	Denies P3, P4, P5, P7, P9, and P11 (Variant 2B)

3.4.1. Reply

Let us start with those who accept my claims about Donny's susceptibility to rational criticism but reject MEP as a necessary condition and deny Best Explanation for a Reason to Bowl.[43] Someone going this route must explain Donny's susceptibility to rational criticism, and the most natural way to do so is to appeal to Donny's intrinsic desire for wealth together with the fact that he believes the act of bowling would itself be a sign of impending wealth. Thus, it seems someone going this route will accept:

> *Evidential Sufficient Condition* (ESC): If an agent *S* has a desire *D* for *p* and a belief *B* that one of their options *O* is a sign (but not a cause) of the truth of *p*, then the agent has a subjective practical reason, grounded in *B* and *D*, to do *O*.

42 Chan, "Are There Extrinsic Desires?"; and Finlay, "Responding to Normativity." Here I am using "Quine" as a verb in the way popularized by Dennett, "Quining Qualia."

43 This covers Variants 1A and 2A in table 1.

Here I think we are encountering an intuitive fault line. All I can say is that to my mind, it is primitive that insofar as rational action relates to desire, its aim is to do the most the agent thinks they can to *make* the world *into* what they want it to be. If you believe an action would be a mere sign that some desire of yours is satisfied but would do nothing to affect whether that desire is satisfied, then that desire does not ground any subjective reason to do that action. James Joyce puts the idea nicely when he writes:

> Rational decision makers should choose actions on the basis of their *efficacy in bringing about desirable results* rather than their auspiciousness as harbingers of these results. Efficacy and auspiciousness often go together, of course, since most actions get to be good or bad news only by causally promoting good or bad things. In cases where causing and indicating come apart, however, . . . it is the causal properties of the act, rather than its purely evidential features, that should serve as the guide to rational conduct.[44]

Is this a distinction without a difference? Is denying ESC while maintaining that MEP applies to signatory desires in some sense not meaningfully different from denying MEP as a necessary condition, accepting MEP as a sufficient condition, and maintaining ESC?

The claim that my rejection of ESC amounts to a superficial distinction has force only if the following principle is a necessary truth:

> *Signatory Belief to Desire* (SBD): If an agent with option *O* has beliefs that entail that doing *O* would be a sign (but not a cause) of the satisfaction of one of their desires, then that agent desires to do *O*.

If SBD is a necessary truth, then necessarily, if an agent *S* is minimally rational, ESC implies *S* has a subjective reason to do ϕ only if MEP also implies *S* has a subjective reason to do ϕ.[45] But SBD is not a necessary truth. There could be a hard-nosed sort of person, a "stoic causalist," who (on at least one occasion) lacks any desire to do a thing they are certain is not a means to anything they want, despite believing that to do it would be a sign of something they want.

44 Joyce, *The Foundations of Causal Decision Theory*, 150.

45 If ESC implies that an agent *S* has a subjective practical reason to do an option *O*, then *S* believes doing *O* would be a sign (but not a cause) of something they desire. Thus, given that SBD is necessary, if ESC implies that *S* has a subjective practical reason to do *O*, then *S* desires to do *O*. But if *S* desires to do *O*, then provided that *S* is minimally rational, *S* believes that doing *O* is a means to satisfy one of their desires, since it is analytic that doing *O* is a means to its being that case that one does *O*.

Should we take SBD to be a norm and hence conclude that the only time ESC and MEP clash is in cases where the agent is "already" irrational? I think not. For one thing, SBD supports the connection between ESC and MEP only if it is restricted to working desire. But given that restriction, SBD is a questionable norm. (Would the stoic causalist be criticizable for, say, feeling no "pull" to doing the thing that would have no causal impact on what they care about?) For another, even if SBD were a norm, violating it would make one guilty of conative, not practical, irrationality. That is, the violation would consist in lacking a rationally required desire.[46] It seems not implausible that an agent could be conatively irrational while still being practically rational, and vice versa. If these are indeed separate normative domains (cf. section 3.2), then the disagreement between ESC and MEP over what subjective practical reasons are had by the stoic causalist is substantive regardless of whether SBD is a norm.

So much for the rejection of Best Explanation for a Reason to Bowl by way of ESC and SBD. If that route is a dead end to anyone who accepts MEP as a necessary condition, where does that leave this objection? It seems the only available alternative is to deny Criticizable Choice and Criticizable Deliberation and to claim Donny is not susceptible to rational criticism in the ways I suggest in section 2.3. There are two possible routes. One route is to remain neutral on the existence of extrinsic desires and claim that it is rationally uncriticizable to not do what you think would satisfy such desires, provided you do not think the rejected action might also lead to the satisfaction of any intrinsic desire.[47] I have already presented my case for the untenability of this route in sections 3.1 and 3.2. Setting that route aside leaves the final route that this objection might take. One might claim that extrinsic desires simply do not exist and that no one can be rationally criticized for failing to aim to satisfy desires that are not there.[48] This route denies Desire to Bowl, In Position to Be Rationally Moved, Criticizable Choice, and Criticizable Deliberation; it dismisses Applies to D^* and Best Explanation for a Reason to Bowl as each having (separate) false presuppositions. I assume it accepts MEP and does not restrict it (since it denies any need for a restriction—extrinsic desires do not exist). I have already considered and rejected an argument that could be used to support this position in section 3.3. Let us consider a different sort of argument.

The sort of argument we turn to now uses inference to the best explanation (IBE). Examples in print focus on motivation. They claim that since rational motivation can be accounted for by intention and intrinsic desire or by belief

46 Cf. Audi, *The Architecture of Reason*, 69.

47 This is Variant 1B in table 1.

48 This is Variant 2B in table 1.

and intrinsic desire, there is no reason to posit extrinsic desires.[49] Using IBE to deny the existence of extrinsic desires puts my opponent in a somewhat awkward position, as there appear to be in Donny's case two things that intrinsic desire is ill-suited to explain: (1) Donny's meriting blame if he does not bowl (and praise if he bowls); and (2) Donny's being in a position to be motivated to bowl. Additionally, setting aside cases like Donny's, there is good reason for thinking extrinsic desires play the causal roles that are characteristic of desire (section 1.2.4) in many cases better than intrinsic desires—this too presents a serious difficulty for an IBE-based rejection of extrinsic desires.

Let us start with Donny's susceptibility to criticism. The opponent we are now considering thinks there is no such susceptibility. But why not? Well, one kind of inference to the best explanation takes a liberal stance with regard to *explananda*. It is happy to deny certain would-be *explananda* if doing so results in a pattern of data that admits of a more parsimonious and unifying explanation. So the Quiner of extrinsic desire might argue that we should just reject Criticizable Choice and Criticizable Deliberation since in so doing we get rid of a few stray intuitive *explananda* that are not readily captured via intrinsic desire. The result of rejecting the intuition that Donny merits criticism if he refuses to bowl, goes the thought, is a simpler, more elegant, less gerrymandered combination of *explananda* and *explanans*.

The first thing I will say in response to this is that it seems born out of too severe a lust for parsimony. To be sure, a theory that posits only intrinsic desires is just plain simpler, but brute simplicity is a slim basis for rejecting the intuitions regarding Donny's susceptibility to criticism. The tradeoff appears less like good abduction and more like fetishism. But aside from this, there is another, perhaps firmer reason to resist the Quiner: there are additional *explananda*, apart from Donny's susceptibility to criticism, that intrinsic desires alone are ill suited to handle.

Consider motivation. Even if, to avoid begging the question against this objection, we do not assume Donny is in a position to be moved to go bowling by a signatory desire to bowl a strike, surely it must be granted that he is *somehow* in a position to be moved to go bowling. The battle should be over what the best explanation for this potential motivation is. My opponent will claim the best explanation involves Donny's intrinsic desire for wealth rather than any signatory desire to bowl a strike. Given that Donny does not believe bowling is a means to satisfy his intrinsic desire for wealth or indeed any intrinsic desire of his, how can such a desire get involved in his motivation to bowl? The reply

49 Chan, "Are There Extrinsic Desires?"; Finlay, "Responding to Normativity" and "Motivation to the Means."

is that Donny can be motivated by the complex of his intrinsic desire for wealth and his belief that bowling, specifically bowling a strike, would be a sign of the satisfaction of that desire.[50] On this picture, Donny can be moved to bowl despite not thereby being moved to acquire, make manifest, realize, or, in short, to cause something he wants to take place.[51] This does not square with the sort of explanation we would most naturally expect Donny to give for why he chose to bowl, should he so choose, or why he was tempted to bowl, should he not. Our natural expectation would be that he would explain his motivation in terms of a desire to bowl a strike: he is thinking about that strike and what it would mean, and he wants it to happen. It is less natural to suggest that on the contrary, he does not want the strike to happen—indeed he does not think bowling would yield anything he wants—and instead he believes bowling would be a sign of something he wants and is moved by *that alone*.[52] In truth, such motivation, wherein there is no prospect of getting anything one wants—and further, no belief that what one is doing is morally required—seems to me incoherent. Stepping back, though, it is unnecessary to press this incoherence claim; to rebut the IBE-based argument, it suffices to point out that the explanatory picture the Quiner offers is less natural than that Donny's motivation would come from a common means-end belief and his desire to bowl a strike.

There is one final set of data that intrinsic desires are ill suited to help explain and for which it seems extrinsic desires ought to be posited. Extrinsic desires, including signatory desires like Donny's, are in some cases the most natural states to play the causal roles of desire (cf. section 1.2.4). That is, extrinsic desires can intuitively be working desires (cf. sections 1.2.5 and 3.3).

Not everyone agrees. Notably, David Chan has presented several cases meant to show that only intrinsic desires can play the causal roles stereotypical of desire.[53] Here is one of his cases:

> John may be afraid of dogs and dislike the sight of dogs. But John may be in love with a woman whom he regularly sees walking her dog around the block. John may therefore welcome, and even look forward to, seeing her dog coming around the corner on its leash, as he knows that

50 A dialectically equivalent alternative says his motivation would come via an intention suitably formed on the basis of the belief(s) just mentioned.

51 Notably, the opponent we are now imagining would hold such motivation to be irrational, since deeming it rational requires ESC, which we are assuming has been rejected.

52 Interlocutor: The claim is not that he can be moved by belief alone but that he can be moved by belief together with his intrinsic desire. Me: How can an intrinsic desire in any way move him to do something he is sure will not lead to its satisfaction?

53 Chan, "Are There Extrinsic Desires?"

> it will be followed by his beloved coming into view. Since John does not have an intrinsic desire to see the dog, it is suggested that his positive feelings in favor of seeing the dog can be attributed to an extrinsic desire.

Speaking of this case, Chan has said the following:

> It is not the sight of the dog but the thought of seeing his beloved that gives John pleasure, and he would be not in the least disappointed if his beloved appeared around the corner without her dog. The daydreaming that is motivated by John's desires will be about meeting his beloved without the dog appearing, even if this cannot happen in the real world.

Chan's claims strike me as unappealing. I see no reason why imagining or experiencing the sight of the dog rounding the corner cannot be a source of pleasure for John. ("Oh boy, there is her dog! In just a moment, she will be rounding the corner as well!") To be sure, John's belief that the dog is a sign of the woman's impending presence is (*ceteris paribus*) a necessary background condition for dog sighting (imagined or real) to cause pleasure. But nevertheless, the sight itself still intuitively can cause pleasure. By the same token, it is implausible to insist that John cannot daydream about the dog's appearing (only to be followed closely by his crush) absent an intrinsic desire to see the dog. Here my opponent might argue that all that is needed to explain John's pleasure at the dog sighting (imagined or real) is an intrinsic desire to see his crush together with a belief that the dog is a sign that he will soon see his crush. They might say the same thing about the other causal effects of the supposed extrinsic desire, e.g., John's attention being directed toward the sound of a dog approaching and the "pull" he feels toward the thought of seeing the dog. This seems to me best taken as an argument for the view, defended by Michael Smith, that extrinsic desires can be reduced to "suitably related" complexes of intrinsic desire and belief.[54] It seems

54 Smith, "Instrumental Desires and Instrumental Rationality." Notably, Smith's "suitably related" qualification, which looks like a problem for a pure reduction, is indispensable. This is because it seems possible (cf. the "stoic causalist" described earlier) to have an intrinsic desire, believe that the truth of some proposition would be a sign or means to the satisfaction of that desire, and yet just not want the truth of the proposition (cf. Smith, 97–98). But further, leaving out "suitably related" seems to commit one to something stronger than (the already implausible) SBD. Whereas SBD is limited to desires to do things, the view in question minus the "suitably related" qualification would entail that in general, whenever one has an intrinsic desire and a belief that the truth of a proposition would be a sign or cause of that desire's satisfaction, one thereby has an extrinsic desire for the truth of the proposition. The resulting proliferation of (mostly gerrymandered) extrinsic desires is unattractive, as they would all, it seems, have to count as working rather than idle desires (cf. section 1.2.5). One avoids this by saying that only the "suitably related" belief-plus-intrinsic-desire complexes count as working extrinsic desires.

to give us no reason to eliminate John's extrinsic desire altogether. Whether Smith's reduction succeeds or not, extrinsic desire is in no danger of elimination.

The reader might be unsure of my take on Chan's case. The sort of intuitions I am aiming to prompt are liable to be muddied due to the "closeness," causally and spatiotemporally, of the objects of John's extrinsic and intrinsic desire (i.e., the dog and John's crush). I suggest we consider a case that lacks this closeness.

Recall my Biden example from section 3.3. Consider me in the days after the 2020 election but before Biden's victory was announced. At that time, finding out that Biden had won Michigan would have been (and indeed later was) a source of great pleasure; I fantasized about a Biden win in Michigan; my attention was drawn to news alerts about Michigan; etc. Some desire ought to have been playing these roles. My claim is that it was an extrinsic desire for a Biden win in Michigan. But notice: all the ready-to-hand alternative desires that might have played those roles were *also* extrinsic. What were those desires? Well, a desire for a Biden presidency, for a reversal of Trump's policies on climate change, for more humane policies on immigration, etc. Assuming those desires existed, none of them were for things I wanted *in themselves*, and in fact, I am not even sure exactly what intrinsic desires of mine would have ultimately supported them. If forced to specify the intrinsic desires, they would have been something like desires for general human happiness, equality, health, flourishing—but those too might ultimately have turned out to be extrinsic desires. My immediate answer would have been nebulous and tentative, and I would have needed to reflect on it to sort it out.[55] In this case, it is, I expect, clearer than in Chan's case both that an extrinsic desire—my desire for a Biden win in Michigan—is playing the desire roles and that there are no obvious intrinsic desires at hand to usurp those roles.

4. THE TWO-LEVEL ACCOUNT

I have argued that an extrinsic desire—specifically, a signatory desire—together with a means-end belief can ground a subjective practical reason. The cases I have used show further that such reasons can occur even if the agent is certain none of their options are means to any of their intrinsic desires. So there can be a subjective reason to do something grounded solely in extrinsic desire, unaccompanied by any reason grounded in intrinsic desire to do the thing. What is the status of such reasons? In particular, what happens when they compete

55 Facts about what people intrinsically desire seem like they will ordinarily be subject to a good deal more indeterminacy than facts about what they extrinsically desire. Cf. Rawls, *A Theory of Justice*: "Whether an aim is final or derivative is not always easy to ascertain. The distinction is made on the basis of a person's rational plan of life and the structure of this plan is not generally obvious, even to him" (494).

with reasons grounded in intrinsic desire? In this section I defend the view that subjective reasons grounded in extrinsic desires can be *trumped* by reasons grounded in intrinsic desire, and this can happen even when the desires and beliefs grounding the respective reasons are of equal respective intensities.

Suppose you have only two options, *A* and *B*. You believe *A* is a means to satisfy an extrinsic desire but not any intrinsic desire, and *B* is a means to satisfy an intrinsic desire. Suppose the desires are of equal intensity, and the beliefs carry equal confidence. My claim is that you rationally must do *B*. Why? Not because of any *quantitative* difference between your desires or reasons but rather because of the *qualitative* difference between intrinsic and extrinsic desires. If you intrinsically desire something, you want it for its own sake—it is one of your basic ends. If you only extrinsically desire something, you do not want it for its own sake—it is not one of your basic ends. It is irrational to pursue something that is not one of your basic ends at the expense of something that is—or at least this is so when the desires for the two things are of close to equal intensity, and your levels of confidence in your ability to obtain each desire are close to equal.

Let us check this with a case. Walter believes a predictor whose predictions he is sure are roughly 90 percent reliable sent him a coupon for a club sandwich, a beer, and a dessert iff she predicted he would not bowl today. He believes the coupon will arrive at his house tonight if ever. What Walter intrinsically desires are pleasurable taste sensations, which for simplicity we will represent in terms of gustatory hedons. He does not believe staying home is a means to such sensations, since the coupon is either already in the mail, or it is not coming regardless. His other notable option is to go bowling with Donny and the Dude. He is confident that if and only if he bowls with them, they will buy him a club sandwich and a beer. A club sandwich and a beer would yield nine gustatory hedons; an added dessert would make it ten. So Walter does not believe staying home will cause him to get any gustatory hedons, but it will give him a *sign* (with roughly 90 percent reliability) that he will soon be able to get ten hedons. By contrast, he is more or less certain that hanging out with his friends will procure him nine gustatory hedons. His desire for the sign of ten hedons and his desire for nine hedons are of close to equal intensity, but they are of different quality: one is extrinsic, the other intrinsic. His confidence in the claim that the predictor is 90 percent reliable is roughly equal to his confidence in his friends' willingness to buy him dinner. There are no other desires that Walter believes might be satisfied as a result of anything he can do tonight. He cannot both stay home and go bowling; he must choose one.

By MEP, Walter has a subjective practical reason to stay home that is grounded in his signatory desire together with the associated means-end belief, as well as a subjective practical reason to go bowling with his friends that is grounded in his

intrinsic desire for nine gustatory hedons together with the associated means-end belief. What I claim is that the reason grounded in the signatory desire is trumped by the reason grounded in the intrinsic desire, even if the two desires have equal intensity. In other words, Walter is not faced with a toss-up; the only rational option is to go bowling with his friends. To reiterate the core intuition, there is an inherent irrationality in choosing to make desirable but causally impotent signs of your ends happen rather than to make your ends happen—or this is so at least in a case like Walter's, where the desire for the sign and the intrinsic desire are close to equal intensity, and the respective means-end beliefs carry close to equal confidence. It bears emphasizing that trumping in cases like Walter's is not due to one reason's arising from a means-end belief and the other's arising from a "signatory belief." Both reasons arise from means-end beliefs.

Are there any cases where trumping does not occur? If the extrinsic desire is far more intense than the intrinsic desire or if the agent is far more confident in their ability to satisfy the extrinsic desire, should we say that the reason grounded in the intrinsic desire still trumps the reason grounded in the extrinsic desire? In short, are there thresholds that prevent trumping? It seems unlikely that everyone who considers this will arrive at the same intuitive answer. In the interest of setting forth the securest aspects of an account of subjective reasons grounded in signatory desire, I will take a neutral approach. I will incorporate thresholds into my account but make no assumptions as to where these thresholds should be set or even whether they should be finite rather than infinite. The outlines of a view involving trumping can be drawn so long as trumping occurs in cases like Walter's, where thresholds clearly do not come into play given (a) the near equality in the intensities of the signatory desire and the intrinsic desire and (b) the near equality in the levels of confidence in the respective means-end beliefs.

Let us say that a subjective practical reason grounded in an extrinsic desire is *subordinate*, and a subjective practical reason grounded in an intrinsic desire is *superior*. Let the *weight* of a subjective practical reason grounded in a means-end belief and a desire be an increasing function of the confidence the belief carries and the desire's intensity.[56] For any given option O, I assume there is some (possibly empty) set containing all and only superior (/subordinate) reasons that favor O.[57] I call this set the set of superior (/subordinate) reasons favoring O, and I assume the reasons in that set have a total weight that is an

56 I intend "increasing" in this sense: letting $w(c, i)$ be weight as a function of belief confidence and desire intensity, I assume that if $0 \leq n, m$ then $w(c, i) \leq w(c + n, i + m)$ and if both $0 \leq n, m$ and $0 < n + m$ then $w(c, i) < w(c + n, i + m)$.

57 For convenience, I treat a single-member set as interchangeable with its member, and I speak collectively about the members of a set by speaking of the set.

increasing function of their individual weights.[58] I say that an agent's superior reasons favor some option O iff there is no O^* such that the total weight of superior reasons favoring O^* is greater than the total weight of superior reasons favoring O. Let every positive number be a *threshold*. Say that the set of subordinate practical reasons to do an option O is *trumped* if its total weight is below the *appropriate* threshold, and the superior subjective practical reasons favor some alternative option(s) distinct from O.[59] Say the set of subordinate reasons to do an option is *active* iff it is not trumped.

With these definitions in hand, I propose the following.

> *The Two-Level Account* (TLA): If the set of subordinate subjective practical reasons $\mathbb{R}$ favoring an option O is active, its members play the distinctive roles of subjective practical reasons. If $\mathbb{R}$ is trumped,
>
> 1. The members of $\mathbb{R}$ do not contribute their weight to O; the weight, individually and collectively, of the members of $\mathbb{R}$ is irrelevant to what it is rational for the agent to do.
> 2. The agent cannot be rationally criticized for failing to act in accord with $\mathbb{R}$.
> 3. If the agent has suitably involved all their superior subjective practical reasons in their deliberation, failure to also consider, individually or collectively, the members of $\mathbb{R}$ is not rationally criticizable.[60]
> 4. The agent cannot be rationally motivated by $\mathbb{R}$.[61]

58 It is natural to think of this function as summation, but I do not assume this. I also do not assume favoring is one to one; a reason might favor two or more options equally—that is, it might favor doing any member of a set of options, or it might favor two or more options to different degrees. In either of these cases, it might be better to speak of a "total favoring amount" rather than a total weight.

59 There are two things to note here. First, this is only a sufficient condition. There may be other circumstances in which a subordinate reason is trumped; I am currently uncertain about this. Second, this account is meant to be neutral on what thresholds are appropriate. It can be squared with the view that trumping always occurs by using the extended reals and taking the appropriate threshold to always be positive infinity. If appropriate thresholds should ever be finite, I make no assumptions about what determines them (beyond what I have already said), about whether they should be the same for every agent at every time, or about whether they should be vague. What complications arise if they are treated as vague depends on what theory of vagueness is chosen. (E.g., a trivalent approach requires dealing with reasons that are neither trumped nor not trumped.) As far as I see, such complications are orthogonal to my claims.

60 Note that this is only a sufficient condition. I tend to think there are other kinds of conditions under which it is rationally uncriticizable to ignore trumped reasons.

61 This is not to say that the agent cannot be *irrationally* motivated by $\mathbb{R}$. Indeed, I would say that they can.

TLA and MEP are logically independent. One could accept MEP only as a sufficient condition while also accepting ESC (cf. section 3.4) and consistently conjoin both of those with TLA. Since I accept MEP as a necessary and sufficient condition and reject ESC, I prefer to take MEP and TLA as a package view, on which you can have a superior subjective reason to do some option *O* only if you believe that *O* is a means to the satisfaction of an intrinsic desire. I call the conjunction of TLA with MEP the *Two-Level Means-End Account*.

Given the Two-Level Means-End Account, Walter subjectively rationally ought to hang out with his friends. By MEP, Walter has a superior reason to hang out and a subordinate reason to stay home. The latter reason is trumped. (In other words, its weight is below the appropriate threshold, and Walter's superior reasons favor a different option.) So by TLA, Walter's subjective reason to stay home does not contribute its weight. Since his only other subjective reason favors hanging out with his friends, that is the rational choice.

There are interesting cases where the subordinate reasons for one option are trumped but those for another option are not. Naturally, one way this can occur is if the weight of the former but not the latter is below the appropriate threshold. It can also occur if superior reasons do not favor a unique option. Suppose there are subordinate reasons R_1 to do *A* and R_2 to do *B*, where R_1 and R_2 have weights below the appropriate threshold (so they are both "trumpable"), and superior reasons favor doing either *B* or *C*. In this case, R_1 is trumped, but R_2 is not. Let us look at this more concretely.

The Dude's only intrinsic desire is for pleasure. He is certain that a Predictor mailed him a free day-pass to a spa iff she predicted he would go bowling at Alley *A*. He is also certain she mailed him eight dollars iff she predicted he would go bowling at Alley *B*. The Dude wants to bowl at Alley *A* (/*B*), since he believes that would be a sign of an incoming spa pass (/eight bucks). The Dude thinks a spa day would be good for ten hedons; as for the eight bucks, he can use it to buy a White Russian, which would be good for one hedon. Now, unlike Alley *A*, to which he is indifferent, the Dude likes Alley *B*; he digs the ambiance. He is certain going to Alley *B* is a means to nine hedons. If he stays home, the Dude will take a long bath, which would also be nine hedons. Like Walter, the Dude takes the Predictor to be 90 percent reliable. See table 2 below for a summary of the Dude's hedons given each action-prediction combination.

Table 2. The Dude's Hedons Depending on Which Prediction Was Made and What He Does

	Alley *A* predicted	Alley *B* predicted	Stay home predicted
Bowl at Alley *A*	10	1	0
Bowl at Alley *B*	18	10	9
Stay home	18	10	9

For definiteness, let us assume that the weight of a reason grounded in a desire *D* and a belief *B* is equal to the strength of the desire times the confidence (understood as a percentage) carried by the belief. Let us also assume that the Dude's desire for *n* hedons has an intensity of *n*. The Dude has no superior reason to bowl at Alley *A*. He has a subordinate reason to bowl at Alley *A* with weight 9. He has a subordinate reason to bowl at Alley *B* with weight 0.9. He has a superior reason to bowl at Alley *B* with weight 9. Last, he has a superior reason to stay home with weight 9. See table 3 for a summary of the Dude's reasons and their weights.

Table 3. Summary of Weights of Reasons

	Bowling at Alley *A*	Bowling at Alley *B*	Staying home
Total weight of superior reasons		9	9
Total weight of subordinate reasons	9	0.9	

Should the Dude go bowling at Alley *A* and get the sign of the spa pass? Or is he permitted to either bowl at Alley *B* or stay home, since either one is a means to more intrinsically desired pleasure than bowling at Alley *A* is a means to? None of the above! Given its equality in weight to the superior reason to bowl at Alley *B*, the Dude's subordinate reason to bowl at Alley *A* is trumped. But the subordinate reason to go bowling at Alley *B* is active and breaks the tie with staying home. The Dude should go bowling at Alley *B*: it is a means to just as much intrinsic desire as soaking in the tub, but it also satisfies a signatory desire.[62]

62 If we assume agents' utility functions over possible worlds reflect intrinsic desire only (cf. Weirich, *Decision Space*, "Intrinsic Utility's Compositionality" and *Models of Decision-Making*), we get an interesting result for decision theory. Evidential Decision Theory says the Dude must go bowling at either Alley *A* or Alley *B*, since according to EDT, ExpectedUtility(Alley*A*) = $\Sigma_n n[\Pr(n \text{ hedons}|\text{Alley}A)] \approx 10$ = ExpectedUtility(Alley*B*) = $\Sigma_n n[\Pr(n \text{ hedons}|\text{Alley}B)] \approx 10 >$ ExpectedUtility(Home) = $\Sigma_n n[\Pr(n \text{ hedons}|\text{Home})] \approx 9$. By contrast, Causal Decision Theory (CDT) says he can either go bowling at Alley B or stay home, since according to CDT, given the aforementioned assumption, ExpectedUtility(Alley*A*) + 9 ≤ ExpectedUtility(Alley*B*) = ExpectedUtility(Home). Hence, by the lights of the Two-Level Means-End Account, both Evidential and Causal Decision Theories are wrong, for the Dude rationally must go bowling at Alley *B*. For background on decision theory and the debate between the Evidential and Causal theories, see Jeffrey, *The Logic of Decision*; Nozick, "Newcomb's Problem and Two Principles of Choice"; Gibbard and Harper, "Counterfactuals and Two Kinds of Expected Utility"; Sobel, "Probability, Chance, and Choice" and "Notes on Decision Theory"; Skyrms, *Causal Necessity* and *Pragmatics and Empiricism*; Lewis, "Causal Decision Theory"; and Joyce, *The Foundations of Causal Decision Theory*.

5. CONCLUSION

As a general rule, when you want a mere sign, you do not also believe you can make that sign happen without undermining its significance. Cases like Donny's, Walter's, and the Dude's show that deviations from this rule are not inconceivable. I have argued that in such deviant cases, the agent has a subjective reason to bring about the sign they desire, and this reason can be accounted for by the straightforward means-end principle. What is more, such unusual cases make it clear that it is possible to have a subjective practical reason to do something despite believing it to be neither morally required nor a means to satisfying any of one's intrinsic desires. The view that subjective reasons can be grounded in signatory desires implies that they can be grounded in extrinsic desires; I have therefore responded to philosophers who argue that extrinsic desires are not rationally relevant as well as to those who argue that they simply do not exist. Last, I have argued that subjective reasons grounded in extrinsic desires can be trumped by ones grounded in intrinsic desires and have presented a Two-Level Account to express this idea. Conjoining this account with the aforementioned means-end principle yields an outline of a theory of desire-based subjective reasons: the Two-Level Means-End Account.[63]

University of Massachusetts Amherst
dealyjustins@gmail.com

REFERENCES

Anscombe, G. E. M. *Intention*. 2nd ed. Cambridge, MA: Harvard University Press, 1963.

Arpaly, Nomy, and Timothy Schroeder. *In Praise of Desire*. Oxford: Oxford University Press, 2014.

Audi, Robert. *The Architecture of Reason: The Structure and Substance of Rationality*. Oxford: Oxford University Press, 2001.

Braddon-Mitchell, David, and Robert Nola. "Introducing the Canberra Plan." In *Conceptual Analysis and Philosophical Naturalism*, edited by David Braddon-Mitchell and Robert Nola, 1–20. Cambridge, MA: MIT Press, 2009.

Brandt, Richard B. *A Theory of the Good and the Right*. Oxford: Clarendon Press, 1979.

63 I would like to thank Phil Bricker, Louis Gularte, Chris Meacham, Andrew Morgan, and two anonymous referees for generous and insightful criticisms and suggestions.

Chan, David K. "Are There Extrinsic Desires?" *Noûs* 38, no. 2 (June 2004): 326–50.

Dancy, Jonathan. "Response to Mark Schroeder's *Slaves of the Passions*." *Philosophical Studies* 157, no. 3 (February 2012): 455–62.

Davidson, Donald. "Actions, Reasons, and Causes." *Journal of Philosophy* 60, no. 23 (1963): 685–700.

Dennett, Daniel C. "Quining Qualia." In *Consciousness in Contemporary Science*, edited by Anthony J. Marcel and Edoardo Bisiach, 42–77. New York: Oxford University Press, 1988.

Finlay, Stephen. "Motivation to the Means." In *Moral Psychology Today: Essays on Values, Rational Choice, and the Will*, edited by David K. Chan, 173–92. New York: Springer, 2008.

———. "Responding to Normativity." In *Oxford Studies in Metaethics*, vol. 2, edited by Russ Shafer-Landau, 220–39. Oxford: Oxford University Press, 2007.

Gerken, Mikkel. "Warrant and Action." *Synthese* 178, no. 3 (February 2011): 529–47.

Gibbard, Allan, and William L. Harper. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, edited by C. A. Hooker, J. J. Leach, and E. F. McClennon, 125–62. Dordrecht: Reidel, 1978.

Harman, Gilbert. *Explaining Value and Other Essays in Moral Philosophy*. Oxford: Clarendon Press, 2000.

Hawthorne, John, and Jason Stanley. "Knowledge and Action." *Journal of Philosophy* 105, no. 10 (October 2008): 571–90.

Hume, David. *A Treatise of Human Nature*. Edited by L. A. Selby-Bigge and P. H. Nidditch. Oxford: Clarendon Press, 1978.

Jeffrey, Richard C. *The Logic of Decision*. New York: McGraw-Hill, 1965.

Joyce, James M. *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press, 1999.

Joyce, Richard. *The Myth of Morality*. Cambridge: Cambridge University Press, 2001.

Karttunen, Lauri. "Syntax and Semantics of Questions." *Linguistics and Philosophy* 1, no. 1 (January 1977): 3–44.

Lewis, David. "Causal Decision Theory." *Australasian Journal of Philosophy* 59, no. 1 (1981): 5–30.

Littlejohn, Clayton. "Must We Act Only on What We Know?" *Journal of Philosophy* 106, no. 8 (August 2009): 463–73.

Lord, Errol. "Having Reasons and the Factoring Account." *Philosophical Studies* 149, no. 3 (July 2010): 283–96.

Mackie, J. L. *Ethics: Inventing Right and Wrong*. London: Penguin, 1977.
Marks, Joel. "The Difference between Motivation and Desire." In *The Ways of Desire*, edited by Joel Marks, 133–47. London: Routledge, 1986.
McDaniel, Kris, and Ben Bradley. "Desires." *Mind* 117, no. 466 (April 2008): 267–302.
Nozick, Robert. "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, 114–46. Dordrecht: Springer, 1969.
Parfit, Derek. *On What Matters*. Vol. 1. Oxford: Oxford University Press, 2011.
———. "Rationality and Reasons." In *Exploring Practical Philosophy: From Action to Values*, edited by Dan Egonsson, Jonas Josefsson, Björn Petersson, and Toni Rønnow-Rasmussen, 17–40. London: Routledge, 2001.
Rawls, John. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press, 1971.
Savulescu, Julian. "Rational Desires and the Limitation of Life-Sustaining Treatment." *Biolethics* 8, no. 3 (July 1994): 191–222.
Scanlon, T. M. *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press, 1998.
Schroeder, Mark. "Getting Perspective on Objective Reasons." *Ethics* 128, no. 2 (January 2018): 289–319.
———. "Having Reasons." *Philosophical Studies* 139, no. 1 (May 2008): 57–71.
———. *Slaves of the Passions*. Oxford: Oxford University Press, 2007.
Sinhababu, Neil. "The Humean Theory of Motivation Reformulated and Defended." *Philosophical Review* 118, no. 4 (October 2009): 465–500.
Skyrms, Brian. *Causal Necessity*. New Haven: Yale University Press, 1980.
———. *Pragmatics and Empiricism*. New Haven: Yale University Press, 1984.
Smith, Michael. "In Defense of 'the Moral Problem': A Reply to Brink, Copp, and Sayre-McCord." *Ethics* 108, no. 1 (1997): 84–119.
———. "Instrumental Desires and Instrumental Rationality." *Aristotelian Society Supplementary Volume* 78, no. 1 (July 2004): 93–109.
———. *The Moral Problem*. Malden, MA: Blackwell, 1994.
Sobel, Jordan Howard. "Notes on Decision Theory: Old Wine in New Bottles." *Australasian Journal of Philosophy* 64, no. 4 (1986): 407–37.
———. "Probability, Chance, and Choice: A Theory of Rational Agency." Unpublished manuscript, 1978.
Stanley, Jason, and Timothy Williamson. "Knowing How." *Journal of Philosophy* 98, no. 8 (August 2001): 411–44.
Sylvan, Kurt. "Respect and the Reality of Apparent Reasons." *Philosophical Studies* 178, no. 3 (October 2021): 3129–56.
———. "What Apparent Reasons Appear to Be." *Philosophical Studies* 172, no. 3

(March 2015): 587–606.
Thomson, Judith Jarvis. "Imposing Risks." In *Rights, Restitution, and Risk*, edited by William Parent, 173–91. Cambridge, MA: Harvard University Press, 1986.
Velleman, J. David. *Practical Reflection*. Princeton: Princeton University Press, 1989.
Verdejo, Víctor M. "Norms for Pure Desire." *Theoria* 35, no. 1 (2020): 95–112.
Vogelstein, Eric. "Subjective Reasons." *Ethical Theory and Moral Practice* 15, no. 2 (April 2012): 239–57.
Wedgwood, Ralph. "Gandalf's Solution to the Newcomb Problem." *Synthese* 190, no. 14 (September 2013): 2643–75.
Weirich, Paul. *Decision Space: Multidimensional Utility Analysis*. Cambridge: Cambridge University Press, 2001.
———. "Intrinsic Utility's Compositionality." *Journal of the American Philosophical Association* 1, no. 3 (Fall 2015): 545–63.
———. *Models of Decision-Making: Simplifying Choices*. Cambridge: Cambridge University Press, 2015.
Whiting, Daniel. "Keep Things in Perspective: Reasons, Rationality, and the *A Priori*." *Journal of Ethics and Social Philosophy* 8, no. 1 (March 2014): 1–22.
Williams, Bernard. "Internal and External Reasons." In *Moral Luck: Philosophical Papers, 1973–1980*, 101–13. Cambridge: Cambridge University Press, 1981.
Wodak, Daniel. "Can Objectivists Account for Subjective Reasons?" *Journal of Ethics and Social Philosophy* 12, no. 3 (December 2017): 259–79.
———. "An Objectivist's Guide to Subjective Reasons." *Res Philosophica* 96, no. 2 (April 2019): 229–44.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3359

AMBIGUOUS THREATS

"DEATH TO" STATEMENTS AND THE MODERATION OF ONLINE SPEECH ACTS

Sarah A. Fisher and Jeffrey W. Howard

In July 2022, a Facebook user in Iran posted the phrase "death to Khamenei" ("marg bar Khamenei" in the original Farsi).[1] The content was posted in a public group that described itself as supporting freedom for Iran. The post shared a cartoon (dating from a 2011 blog post) of the Iranian supreme leader, Ayatollah Khamenei, his beard forming a fist, which grasped a woman wearing a hijab, a blindfold, and a chain around her ankles. A speech bubble next to the cartoon stated that being a woman is forbidden. The user's caption, accompanying the cartoon, included the text "death to the anti-women Islamic government and its filthy leader Khamenei" alongside wider criticism of the Iranian regime for its treatment of women. The post appeared shortly before Iran's National Day of Hijab and Chastity, an occasion used by critics of the government to protest against the mandatory hijab and other illiberal policies.

Another Facebook user complained about this post to Meta's content moderation teams, which govern the speech of users on the social media platforms Facebook and Instagram. In response, the post was removed, having been deemed to violate Facebook's Violence and Incitement Community Standard.

Did Meta make the right decision by removing the post? This single question, it turns out, illuminates a litany of deeper philosophical issues concerning what sort of speech is properly targeted by platforms' content moderation systems. Upon review, both Meta and its Oversight Board concluded that the platform had erred in removing the post. But the arguments that brought them to this conclusion were fundamentally in conflict and involved crucial confusions about the criteria for removing speech. This article pinpoints an important source of instability in how speech is currently governed online—namely, a failure to distinguish the illocutionary and perlocutionary dimensions of speech acts (very roughly, their communicative force versus their downstream effects).

1 The details of this case are taken from the Meta Oversight Board's decision 2022-013-FB-UA, available at https://www.oversightboard.com/decision/FB-ZT6AJS4X/.

By offering a solution, our goal is to provide broader normative guidance for the proper design and enforcement of platform rules.

In section 1 we describe further details of the case. In section 2 we map the conclusions of Meta and the Oversight Board to a philosophical distinction between illocution and perlocution. Section 3 identifies confusions in both parties' reasoning about the case and traces these back to Meta's policy. In section 4 we show why the confusions are objectionable and set out three possible solutions, whereby content moderation targets (1) the probability of the speech having harmful (perlocutionary) effects, (2) the probability of its (illocutionary) force belonging to a prohibited category, or (3) both of the above aspects, using a systematic combinatory procedure. We conclude by defending an "illocutionary-first" version of target 3.

1. CASE BACKGROUND

1.1. At-Scale Moderation Decision

Like all administrators of social media, Meta specifies what speech is and is not allowed on its platforms via a suite of "Community Standards." The policies therein cover topics ranging from privacy, nudity, and sexual exploitation to hate speech, misinformation, and incitement, among many others. They are enforced through proactive and reactive moderation, undertaken by a combination of automated systems and human reviewers. Meta's content moderation activity (and that of large social media companies in general) often proves controversial, given its impacts on billions of active users.[2] The "death to Khamenei" post involves one such controversy.

Although the complainant reported the post as hate speech, Meta's at-scale moderation team did not judge it to violate Facebook's Hate Speech Community Standard; it was removed for violating the platform's Violence and Incitement Community Standard. That policy stated:

> We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence. . . . We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety. In determining whether a threat is

2 As of December 31, 2022, Facebook had 2.96 billion monthly active users, according to Meta's Q4 2022 Earnings Report, available at https://investor.fb.com/home/default.aspx.

> credible, we may also consider additional information such as a person's public visibility and the risks to their physical safety.[3]

Subsequent investigation by Meta's Oversight Board shed light on the moderators' decision-making. On one hand, "death to *X*" statements are usually allowed on the platform, in that they typically express a mere wish or hope that *X* dies rather than a credible threat or call for lethal violence. On the other hand, Meta removes "death to *X*" statements wherever *X* is a member of a high-risk category. The reason why "death to Khamenei" triggered enforcement action was the fact that the target was a head of state whose safety was therefore deemed to be at elevated risk.[4] Accordingly, the post was classified as a violation of the Violence and Incitement Community Standard, resulting in enforcement action.[5]

1.2. Appeal and Review

The moderation decision was immediately appealed by the user. However, that appeal was not prioritized by Meta's automated systems (which take account of signals concerning the type, virality, severity, and recency of the content), and it was subsequently closed without further review. The user then appealed to the Oversight Board, an independent body established by Meta in 2020 to review and adjudicate content rules and decisions on Facebook and Instagram.

While the Oversight Board was deliberating about the case, Meta conducted its own internal review of the decision. The company continued to maintain that the user had indeed violated the Violence and Incitement Community Standard. However, it nevertheless concluded that the post should have been allowed to stand on the grounds of its "newsworthiness." Meta controversially grants "newsworthiness allowances" to disallowed speech if it is in the public interest for people to see it. To decide whether violating speech merits a

3 The Community Standard has been available in Farsi since February 2022. Note that the Violence and Incitement Community Standard encompasses both *threats* of violence and *incitements* to violence, even though these are distinct categories causing different harms and subject to differing normative analyses in free-speech literature. For the full statement of the current policy and the relevant change log, see https://transparency.fb.com/en-gb/policies/community-standards/violence-incitement/.

4 According to the Oversight Board's report, other categories of high-risk targets include former heads of state, candidates and former candidates for head of state, candidates in national and supranational elections for up to thirty days after election if not elected, people with a history of assassination attempts, activists, and journalists.

5 In addition to the post being removed, the user who posted it received one "strike" against their account (where accumulating more strikes leads to greater restrictions on one's ability to create content on the platform) and two "feature limits," meaning that they could not post or comment in groups for the next thirty days and could not post or comment on other Facebook surfaces (except Messenger) for seven days.

newsworthiness allowance, the company putatively weighs the public interest in seeing the speech against the risk of harm caused by the speech. In this case, the company deemed the risk of offline harm to the Iranian supreme leader to be outweighed by the significant public interest in hearing the antiregime perspective.[6] That perspective was considered especially newsworthy given the timing of the post (when protests were being organized in the run-up to the National Day of Hijab and Chastity) and its contribution to political and religious discourse (in criticizing Iran's mandatory hijab laws and the government's overall stance towards women). The argument for allowing the speech was further compounded by the Iranian government's history of suppressing free expression, including online (with Facebook having been banned in Iran since 2009 and only accessible via technical workarounds). As a result, the post was reinstated on the platform in August 2022.[7]

1.3. Oversight Board Ruling

In January 2023, the Oversight Board published the results of its own investigation into the case. Although the board similarly determined that the post should be allowed to stand, it marshaled quite different reasons and argumentation in support of that conclusion. Specifically, the Oversight Board ruled that "death to Khamenei" here was better interpreted as "down with Khamenei," a "clearly rhetorical" expression of criticism, disdain, or disgust for Khamenei, rather than a genuine threat or call for his assassination. In other words, it ruled that the speech was not a violation of the Violence and Incitement Community Standard at all.

6 At the time of this article's publication, Meta's description of the newsworthiness allowance (available at https://transparency.fb.com/en-gb/features/approach-to-newsworthy-content/) states:

> When making a newsworthy determination, we assess whether that content surfaces an imminent threat to public health or safety, or gives voice to perspectives currently being debated as part of a political process. We also consider other factors, such as:
>
> - Country-specific circumstances (for example, whether there is an election underway, or the country is at war)
> - The nature of the speech, including whether it relates to governance or politics
> - The political structure of the country, including whether it has a free press
>
> We remove content, even if it has some degree of newsworthiness, when leaving it up presents a risk of harm, such as physical, emotional and financial harm, or a direct threat to public safety.

On this occasion, a *narrow* newsworthiness allowance was granted, not a *scaled* one, meaning that other posts of "death to Khamenei" were still subject to removal.

7 The strike against the user's account was also removed. However, it was impossible to reverse the temporary feature restrictions, which had already run their course.

The Oversight Board's interpretation of the user's speech was based on testimony from Farsi speakers familiar with Iranian current affairs (including language experts and members of the public who submitted comments on the case), who attested to a common rhetorical use of "death to Khamenei" as a political slogan in contemporary public discourse. Deployed in this way, the board ruled, the phrase does not threaten or call for violence against Khamenei but instead falls within the Community Standard's exemption for speech that expresses disdain or disagreement by threatening or calling for violence in *non-serious* ways.[8] The board judged that the author of the post was using "death to Khamenei" in its rhetorical sense, merely criticizing rather than threatening or calling for violence. As such, the post did not violate the Community Standard, nor did it require any special allowance to appear on the platform.[9]

In light of its findings, the Oversight Board issued a series of recommendations to Meta, including to revise the public-facing Violence and Incitement Community Standard and the internal guidance for moderators, so as to implement a more contextually nuanced treatment of seemingly threatening or inciting speech.

Meta has accepted the Oversight Board's guidance that "death to Khamenei" is a political slogan that is integral to the ongoing protests in Iran. The phrase is now allowed on the platform in that context (and previous enforcement actions are being reversed). At the time of writing, Meta is still considering several of the other recommendations made by the Oversight Board, including those to revise the Violence and Incitement Community Standard and internal guidance for moderators.[10]

How exactly should Meta proceed? Drawing on longstanding distinctions in the philosophy of language, we argue that the answer depends on whether platform rules should primarily target an utterance's illocutionary force or instead its perlocutionary effects. The analysis we provide in what follows is

8 In fact, the Community Standard states only that Meta "understands that" people commonly express disdain or disagreement by threatening or calling for violence in nonserious ways, not that such speech is permitted. This would seem to leave room for a stricter interpretation than the Oversight Board's presumption of an exemption, although we will not pursue the point here. We assume the policy contains a deliberate carveout for "nonserious" speech.

9 This view is reminiscent of the position adopted by the US Supreme Court, which has held that many statements that may *appear* to threaten are in fact mere hyperbole and do not constitute what it calls "true threats." Such cases of hyperbole are ubiquitous in political discourse, which is often "vituperative, abusive, and inexact." See *Watts v. United States* 394 U.S. 705 (1969).

10 Details of Meta's response are taken from https://transparency.fb.com/en-gb/oversight/oversight-board-cases/caricature-of-ayatollah-ali-khamenei (accessed July 12, 2023).

intended to support the development of a principled, fair, and appropriate approach in content moderation policy and enforcement.

2. THE ILLOCUTION/PERLOCUTION DISTINCTION

It is striking how the different reasoning applied by Meta and its Oversight Board tracks a longstanding distinction in speech act theory. Yet this distinction is currently obscured in online speech governance, leading to inconsistent—and, we will argue, unfair—results for users. By demonstrating the normative significance of the distinction, we aim to offer guidance on how content rules should be interpreted and enforced by platforms. Given that large online platforms are now arguably the most important fora for the exercise of our communicative liberties, it is vital that they be governed by clear and defensible principles.

What exactly is the distinction we have in mind? Austin famously described several different kinds of acts we perform when speaking—or things we *do* with words.[11] Most relevant to the current discussion are his categories of "illocution" and "perlocution." We will present each in turn and explain how they pertain to the "death to Khamenei" case, before proceeding to offer our central normative argument.[12]

2.1. Illocution

We perform illocutionary acts in using words with a particular *force*, such as telling, instructing, warning, promising, or thanking. Consider the sentence "The window is open." Even after accounting for the meanings of its constituent words and the way they are combined, there is still scope for this sentence to be uttered with differing force. In one scenario, a speaker could simply be informing her audience that the window is open. On a different occasion, though, uttering "The window is open" could be a request or a command to close it (say, where the speaker manages a building in which it is forbidden to open the windows, and she has just walked in on a tenant who is breaking the rules). In yet another kind of context, "The window is open" could be a warning (say,

11 Austin, *How to Do Things with Words*.

12 The claim that distinctions within Austinian speech act theory can shed light on pressing normative issues is by now familiar in social philosophy. See, for example, Langton, "Speech Acts and Unspeakable Acts" and "Blocking as Counter-speech"; Langton and Hornsby, "Free Speech and Illocution"; Hornsby, "Disempowered Speech"; McGowan, "Conversational Exercitives" and "Oppressive Speech"; Maitra, "Silencing Speech"; Dotson, "Tracking Epistemic Violence, Tracking Practices of Silencing"; Kukla, "Performative Force, Convention, and Discursive Injustice"; Hesni, "Illocutionary Frustration"; and Schiller, "Illocutionary Harm."

to the parent of a small child who is entering a fourth-floor apartment with floor-to-ceiling windows).

Returning to our case, the phrase "death to Khamenei" similarly underdetermines the illocutionary act being performed. What act, exactly, did the user in question perform in posting it? Were they threatening the life of the Iranian leader or calling for others to assassinate him?[13] Were they criticizing the leader and his regime? Plainly, this question is at the heart of the disagreement between Meta and the Oversight Board. Whereas Meta took itself to be dealing with a threat or call for violence against Khamenei (later deciding that the speech should be allowed anyway), the Oversight Board interpreted the utterance as an act of mere criticism.[14]

How should the dispute be adjudicated? This depends in part on a contested question in speech act theory concerning which criteria determine illocutionary force. Broadly speaking, theorists have appealed to three conditions for performing illocutionary acts: (i) the speaker's *intending* to perform that act (in this case, intending to threaten or intending to criticize); (ii) the audience's "uptake" of the act (whether they *interpreted* the speaker to be either threatening or criticizing); and/or (iii) the prevailing social conventions for performing the act being met (including the speaker's having used a particular kind of linguistic formulation, in a particular kind of context, with the requisite degree of authority). As is to be expected, philosophers of language disagree as to which of conditions i–iii are necessary or sufficient for the performance of any given illocutionary act.[15]

While we cannot resolve that debate here, we will need to make *some* minimal assumptions in order to apply the notion of illocution in our context. We assume that what is relevant for online speech governance is *how the illocutionary force of a post would reasonably be interpreted by its audience*. In standard cases, what audiences would reasonably interpret the speech to be doing will be sensitive to shared evidence (available to audience members, speakers, and platform adjudicators) about the relevant language, speaker, and wider context.

13 Following Searle's classification in "A Taxonomy of Illocutionary Acts," threats are standardly treated as illocutionary acts.

14 Those familiar with Austin's work might worry that rhetorical uses of language simply fall outside the scope of the theory, being an instance of nonliteral or nonserious speech (akin to acting in a play, making a joke, or writing a poem, which Austin explicitly excluded from consideration). However, even if one did not believe that any felicitous illocutionary act was performed with "death to Khamenei," that would still be sufficient to undercut Meta's claim that the user was threatening or calling for violence, and that is the important point for our purposes here.

15 For relevant discussion, see Strawson, "Intention and Convention in Speech Acts"; Searle, *Speech Acts*; and Sbisà, "How to Read Austin." For an overview of the contemporary literature on speech act theory, see Fogal, Harris, and Moss, *New Work on Speech Acts*.

The proposed approach is broadly friendly to those who emphasize the importance of the speaker's intention, insofar as reasonable interpretations of speech are those that attempt to recover what the speaker was trying to do. The approach also recognizes a role for uptake, insofar as it is the audience's interpretation that is being estimated. Finally, our approach respects the importance of social conventions, insofar as reasonable interpretations rely on what a speaker of a particular kind would normally be doing with those words in that context.[16] What we end up with then is an illocutionary category tailored to the specific case of speech governance that tracks the speaker's (likely) intentions, the (reasonable) audience's uptake, and the surrounding conventions (other things being equal) without giving priority to any one criterion in its pure form.[17]

2.2. Perlocution

Perlocutionary acts are things we do *by means of* our utterances—the effects our speech achieves in the world. For example, by saying "The window is open" I might cause you to believe that fact; alternatively, I might cause you to close the window; or I might cause you to take additional care to avoid accidents. Across these cases, the focus is on the *consequences* of speech. Likewise, in the "death to Khamenei" case, the perlocutionary question concerns what effects the speaker was (likely to be) producing by means of their speech. Were they causing Khamenei to experience intimidation, or inducing audience members to attempt an assassination? Or alternatively, would the words simply have the effect of raising awareness, causing audiences to reflect on the oppressive government and perhaps take up protest against it?

16 There are complications across all three dimensions. First, a clumsy or misguided speaker might intend to do one thing but reasonably be interpreted as doing another. In some cases, this will be due to negligence or recklessness on the speaker's part, as will be discussed in section 4. Further, we note the possibility that an unreasonable audience might interpret the speaker as doing one thing while a reasonable audience interprets them as doing another. In section 4 we will discuss why we think it is appropriate for content moderation to track the latter. Finally, we note the possibility that available evidence about a particular user or context might affect what interpretation is most reasonable, all things considered, even if that diverges from convention.

17 Even so, the argument below would apply *mutatis mutandis* if we were to adopt (implausibly, we suspect) a single necessary-and-sufficient criterion of illocutionary force. For example, a hardcore intentionalist version would require content moderators to focus on only what the speaker was trying to say, whereas an uptake-centric version would require them to focus on how audiences actually interpret the speech.

There is an obvious and close connection between the illocutionary and perlocutionary aspects of an utterance.[18] For instance, telling *S* that *p* tends to result in *S* coming to believe that *p*; requesting *S* to ϕ tends to result in *S* ϕ-ing; and so on. More precisely, illocutionary acts have the *normal function* of producing the corresponding perlocutionary effects. Yet they will not produce those effects on every occasion, meaning that illocutionary force and perlocutionary effects can diverge in practice. For example, I might tell you that the window is open, but you, seeing it clearly closed, may refuse to believe me. By the same token, a particular threat or call for violence might fail to result in harm.[19] Conceptually, then, the illocutionary force of an utterance can be held apart from its perlocutionary effects, even if the two normally go hand in hand. We will see later why this is so important for thinking about the governance of speech.

With the illocution/perlocution distinction in place, we can rationally reconstruct the positions of Meta and the Oversight Board. For Meta, the author of the "death to Khamenei" post performed a particular illocutionary act—the act of threatening or calling for the killing of Khamenei. The perlocutionary effects of this act included some risk of actual harm, but they also included beneficial awareness raising. The newsworthiness allowance was applied following a cost-benefit analysis of these perlocutionary effects, which determined that the benefits of the speech outweighed the costs. Yet that determination did not alter the illocutionary status of the speech as a threat or a call for murder; the speech constituted a violation, even though it was exempted due to its instrumental benefits.

In contrast, the Oversight Board interpreted the same speech as an illocutionary act of criticism, which was legitimate and protected by Facebook policy, independent of whatever downstream perlocutionary effects (good or bad) it might have had in the situation at hand.

All of that said, it turns out that neither party managed to hold apart illocutionary and perlocutionary issues in a fully consistent way. By analyzing those

18 There can also be disagreements over whether a particular verb is illocutionary or perlocutionary. Indeed, "inciting" seems to be one of these; for relevant discussion, see Kurzon, "The Speech Act Status of Incitement." Because of that, we prefer to contrast threatening or calling for violence (illocutionary) with instigating violence (perlocutionary).

19 It remains a contested issue in speech act theory whether perlocutionary effects should include only those that the speaker intended or also any unintended effects of their speech. For the purposes of applying the theory to content moderation, we wish to include all downstream effects, whether intended or unintended. Those who object to this use of "perlocution" are free to substitute a different label.

inconsistencies in the next section, we will then be in position to recommend our preferred systematic approach.

3. Conflations, Complications, and Confusions

Platforms must decide whether a given utterance violates or does not violate their rules. When deciding this, they might focus on the illocutionary force of the utterance—i.e., whether it constitutes a speech act belonging to a prohibited category, such as an act of threatening or an act of calling for murder. Alternatively, they might focus on the perlocutionary force of the utterance—i.e., its (likely) causal effects in the world. Or they might opt for some principled combination of the two. Soon we shall explain what we think the right position on that is. Before doing so, we will linger on our core case slightly longer to show why platforms' current thinking on this issue is unhelpfully muddled.

We start with the Oversight Board's conflation of illocutionary and perlocutionary issues. The board notes, quite reasonably, the importance of context in determining the force of "death to *X*" utterances. However, it then considers a hypothetical case in which the target is Salman Rushdie rather than the Ayatollah Khamenei. The board argues that this case would pose a much more significant risk of harm (due to the fatwa against Rushdie, the recent attempt on his life, and ongoing concerns for his safety) and would therefore need to be taken more seriously. While it is no doubt true that a threat against Rushdie would be more credible than one against Khamenei, this is a fact about the downstream dangerousness of the threat *once issued*. It does not tell us whether such a speech act (the act of threatening) was performed in the first place. The risks facing Rushdie do not themselves make "death to Rushdie" more likely to be an illocutionary act of threatening or calling for violence than "death to Khamenei." On the contrary, the Oversight Board's decision on the "death to Khamenei" case rests on the claim that "death to" here has a rhetorical use equivalent to "down with." The fact that the "death to Khamenei" post was deemed to instantiate this rhetorical use was what exempted the post from enforcement action. Presumably, then, moderators must rule out the possibility of any other "death to" statement being merely rhetorical before taking enforcement action under the Violence and Incitement Community Standard. This applies equally to a post of "death to Rushdie" as to a post of "death to Khamenei." In sum, changing the identity of the target cannot settle what the (violating or nonviolating) illocutionary force of a "death to" statement is.[20]

20 Perhaps the board would ultimately want to say that the rhetorical use of "death to" is possible only for particular targets, say those who hold political office or those who feature

It seems then that the Oversight Board *does* want moderators to consider perlocutionary effects, despite its claim that acts of criticism should be excluded from consideration. In its discussion of Rushdie, the board implies that even rhetorical uses of "death to *X*" might sometimes be appropriate targets of enforcement, so long as the risks to *X*'s safety are sufficiently high. If this is the right interpretation, the board's reasoning is unclear. Should Meta take a view on the illocutionary force of a "death to" speech act before deciding whether to take enforcement action? Or should it first assess the likely risk of harm to the target? Or should it do both concurrently? The answer is not clear.

That confusion, visible at the level of enforcement and review, is, we think, a direct result of the conflation of illocutionary force and perlocutionary effects in Meta's underlying policy. On one hand, the overarching rationale for moderating content (under the Violence and Incitement Community Standard and other policy provisions) is to prevent harm, i.e., damaging perlocutionary effects of speech. In line with this, Meta states that it will take enforcement action where speech "facilitates serious violence," where there is believed to be a "genuine risk of physical harm or direct threats to public safety," and where these threats are considered "credible." On the other hand, there are carve-outs for users to "express disdain or disagreement by threatening or calling for violence in non-serious ways"—speech that does not in fact threaten or call for violence but has some other illocutionary force.

One might think that nonserious threats are permitted precisely because they tend not to cause harm. As we saw in section 2.2, though, the connection between illocutionary and perlocutionary acts is not one that is guaranteed to hold on every occasion, being contingent on a range of ad hoc contextual factors. There could, for example, be particular illocutionary acts of criticism that are in fact likely to lead to violence; on the flipside, there could be particular illocutionary acts of threatening that are unlikely to do so.

Imagine, first, a celebrity who has a coterie of zealous fans. The celebrity posts something that is intended as a mere criticism (say, "down with Khamenei," in the context of a peaceful political protest movement) and is universally interpreted that way. Yet the zeal of her hardcore fans is such that they will take it upon themselves to attempt serious harm on anyone their idol disagrees with. We can even imagine that this fact is known to the platform's moderation team, making the offline harm entirely foreseeable to them. In this case, the post is likely to instigate violence without having threatened or called for it. Should such speech acts be protected, regardless of their potentially harmful effects?

---

in particular protest slogans. However, even if such an argument could be sustained, it is not provided in the case ruling.

On the flipside, a post that is intended—and widely understood to be—a call for violence (say "One of you should kill Khamenei when he leaves the Beit Rahbari compound at 10 AM tomorrow") may nevertheless carry a very low probability of causing harm to the target. We can imagine that the user has very few followers, none of whom is in a position to attempt an attack, and the compound in question is extremely well protected. In this case, illocutionary and perlocutionary aspects come apart in the opposite direction: the speech has the illocutionary force of a threat or call for violence but no prospect of generating harmful perlocutionary effects. Should such a post be removed despite its extremely low risk?

These two limit cases illustrate the need to clarify whether illocutionary or perlocutionary factors are driving content moderation. Even though illocutionary force and perlocutionary effects are closely connected, we have seen that they can and do diverge. Thus, the question of whether a post constitutes a call for violence is simply not the same as the question of whether it is likely to cause offline harm. By conflating both aspects, policies like Facebook's Violence and Incitement Community Standard fail to give us clear principles for handling the limit cases sketched above and a vast spectrum of intermediate cases that arise at enormous scale and frequency on social media platforms. As we have seen, this confusion at the policy level destabilizes the reasoning of Meta and the Oversight Board during enforcement and review.

4. THREE MODELS OF CONTENT MODERATION

This brings us to the foundational question raised through our case study: Should content moderation on social media platforms target the illocutionary or perlocutionary aspects of speech? For example, when specifying what counts as a forbidden "threat," should platforms emphasize the illocutionary aspect, the perlocutionary aspect, or some combination thereof?[21] At a minimum, we submit that platforms should take a consistent and transparent line in their policy and enforcement. This is both for the sake of moderators, who need a procedure for resolving cases where illocutionary force and perlocutionary effects diverge, and for the sake of users, who have an interest in knowing how their speech will be handled. Indeed, given the important role played by platforms like Facebook in facilitating the speech of billions of people, the

21 Specifying what should count as a forbidden threat for the purposes of online content moderation (or, relatedly but distinctly, what should count as a forbidden threat for the purposes of criminal law) is different from the question of what counts as a threat for the purpose of conceptual or linguistic analysis. Among other reasons, concerns of free speech are relevant for specifying the former but not the latter.

current lack of a clear moderation principle arguably impairs their enjoyment of an intrinsic good—i.e., their ability to express themselves. Meanwhile, at the social level, we are denied a proper balance between the good of free expression in increasingly important online environments and the bad of potentially harmful effects. Accordingly, it is a moral imperative for platforms to get clear on whether their content moderation targets illocutionary force or perlocutionary effects.[22]

In what follows, we contrast three possible approaches to content moderation that focus respectively on perlocutionary effects, illocutionary force, and (as we prefer) a systematic combination of the two.

4.1. Moderating Perlocution

Given that the end goal of content moderation is to minimize the harmful effects of speech, a natural suggestion would be to adopt a purely perlocutionary strategy whereby each utterance is moderated based on its likely resulting harm. On such an approach, enforcement action would be taken against posts that exceed some threshold of risk, as a function of the probability of the harmful effect occurring and its degree of harmfulness. Meanwhile, posts deemed unlikely to cause harm would be left to stand and spread.

Such a strategy immediately runs into two central difficulties. First, it fails to offer clear normative guidance about what exactly platforms ought to do. That is because our ability to predict the effects of particular utterances is highly limited. But second, a purely perlocutionary approach would lead to objectionable overmoderation, licensing the removal of potentially large swaths of legitimate speech. If all that matters is whether an utterance has a contingent effect of causing some harmful result downstream, plenty of legitimate speech that merits protection will be vulnerable to censorship.

22 It might be objected at this point that because platforms are private companies, they do not wrong their users when they remove legitimate speech. Even if a state would wrong them by removing such speech, a *company* does not wrong them by removing it. In reply, though, we note that platforms have themselves committed to respect the value of users' expressive and communicative interests. Meta has expressly committed itself to respecting "voice" as a paramount value. In this way, platforms have voluntarily taken on such a moral obligation. While this modest point is sufficient to establish that they can wrong their users by removing legitimate speech, we are further tempted by a stronger thesis: that platforms exercise a kind of governance power over the public discourse and, in virtue of this power, are bound by similar principles to respect free speech as states are. That stronger thesis underpins the Oversight Board's provocative contention that principles of international human rights law should govern content moderation decisions by the large platforms. For discussion, see Howard, "The Ethics of Social Media."

To see the problem, consider again the limit case raised above involving a celebrity with overly zealous fans. We saw there how an entirely innocent illocutionary act, intended and interpreted as a mere criticism, could nevertheless have foreseeably harmful effects due to the unreasonableness of a small number of audience members. Giving this unreasonable minority the normative power to silence the speaker, rendering her speech eligible for removal, would entirely misallocate the burdens: the unreasonable fans, not the celebrity, are chiefly responsible for any violence that ensues.[23] Thus, a purely consequentialist focus on harm leaves platform users' speech rights hostage to factors that are intuitively irrelevant, such as how many unreasonable people follow them or how much security the subject of their critical speech happens to have.

By the same token, ignoring illocutionary force also disregards the expressive interests of speakers. Suppose that it is true that by posting "death to Khamenei," the user was engaging in speech that was both intended and (more importantly) likely to be understood as vociferous protest against an authoritarian regime. Such expression is at the heart of what the right to free speech protects. It is not plausible to suggest that Iranian users on Facebook have a moral duty to refrain from such expression.[24]

On the flipside, a purely perlocutionary approach would also lead to objectionable undermoderation of illegitimate speech. For instance, if a user issues a threat or call to violence, it is much less plausible to think that they have a moral right to engage in such speech. There is a moral duty to refrain from issuing death threats and encouragements to murder; such activity familiarly falls outside the protective scope of free speech. Therefore it is plausibly an abuse of a platform to engage in such speech, which, while having the normal function of instigating harmful behavior, does not depend for its wrongness and unprotected status on causing harm in every instance.[25]

23 This is compatible with the claim that speakers have moral duties to look out for ways in which unreasonable listeners might misinterpret their speech and to rearticulate or clarify their messages in ways that reduce such risks. Just because a speaker has a right to express a certain message does not mean that there are no moral considerations that bear on *how* she ought to express that message. But we are skeptical that such highly nuanced issues arising from occasionally foreseeing the unreasonable responses of deranged audience members should limit what people are allowed to say. Rather, they bear on how speakers might use discretion when exercising their speech rights.

24 The intuitive force of this claim is doubtlessly bolstered by the fact that the Iranian regime is seriously unjust. But even if the regime were in fact just, free speech theories familiarly protect citizens' general right to protest regardless of the moral standing of their state.

25 See Howard, "Dangerous Speech." Some might think assassination of tyrants is legitimate, such that the illocutionary act of threatening or calling for such assassination is also legitimate. It is possible that people's intuitions on this issue may influence their reactions to

These considerations lead us to reject a purely consequentialist strategy of moderating the perlocutionary effects of speech (whether under Facebook's Violence and Incitement Community Standard or any other platform policy): such an approach unsurprisingly fails to attend to the nonconsequentialist significance of users' rights and duties, which militate in favor of centering illocutionary force.

4.2. Moderating Illocution

At the other extreme, we might be tempted to focus entirely on the illocutionary act performed by a piece of online speech. Thus, a post would be removed if deemed to be a violating illocutionary act, such as threatening or calling for violence (even if it is in fact unlikely to lead to harm in a given case). Meanwhile, legitimate illocutionary acts would be left alone (regardless of any harmful effects they might have).

On our view, this approach is nearly correct. It respects users' freedom of speech by protecting utterances that have valuable or innocuous illocutionary force; speakers have no moral duty to refrain from such speech and indeed have a right against its restriction. In contrast, speakers do not enjoy rights to achieve perlocutionary effects—there is no plausible right, for example, to convince, persuade, or instigate a particular chain of events, whereas there is a right to assert, argue, criticize, and so on. The right to free expression sits at the illocutionary level; and speech enjoys a blanket protection wherever it is reasonably interpreted as having innocent illocutionary force (even if it could end up being harmful for reasons beyond the speaker's control).

At the same time, the illocutionary approach enables platforms to target categories of speech acts—like wrongful threats and calls for violence—that normally function to inflict serious wrongful harms on others. Partly in virtue of the harms that such utterances are calculated (and often tend) to produce, they lack value *qua* free expression since (on the standard view) such unprotected speech is utterly disconnected from the values (such as autonomous expression, democratic citizenship, and the search for truth) that justify free speech in the first place.[26] This is especially clear in cases where the user *intends*

---

the case under discussion. We set aside this complication here, as it plays no explicit role in the official reasoning about the case. Meta disallows calls for assassination of political leaders, even seriously autocratic ones.

26 We assume that freedom of speech is a moral principle that makes it very difficult to justify restrictions on communication by public (and some private) institutions, especially restrictions that silence particular viewpoints. We further assume that this principle is justified by a plurality of interests (of both speakers and audiences), including autonomous self-expression and self-development, education, and democratic self-government. Such

to engage in the relevant speech act; she is not contributing to public discourse by sharing her opinion on matters of public concern but rather endeavoring to cause harm. Such speech is indeed a violation of duties she owes to others. But even in cases where there is no such intention, if the reasonable construal of the utterance is that it constitutes a speech act of threatening or calling for violence, the user may be violating her duty through recklessness or negligence. Such a user cannot stand on her free-speech rights to immunize her speech from interference; she too violates duties she owes to others not to perpetrate wrongful utterances.[27]

Notwithstanding its virtues, a purely illocutionary strategy depends upon a simplifying assumption: that any given utterance admits of only one reasonable interpretation. Yet this is plainly false. In reality, a single utterance can often have multiple competing illocutionary interpretations, each of which may be deemed similarly plausible given the available evidence about the language, the speaker, and the wider context. In fact, the "death to Khamenei" post seems to exhibit just this kind of ambiguity, despite the Oversight Board's confident assertion that it was mere criticism.

The complexities of establishing an utterance's illocutionary force are evident from ongoing debates in speech act theory and should in no way be underestimated. Across a wide range of cases, there are likely to be genuine difficulties in determining which illocutionary act a speaker has performed. After all, Austin's core insight (discussed in section 2.1) was that a speaker's form of words commonly underdetermines what is done in uttering them. This immediately creates room for uncertainty and disagreement about illocutionary force.

Moreover, the context for a social media post is often so sparse that homing in on a unique illocutionary act is simply impossible. Accordingly, we will often need to hold open different possibilities.[28] This points to the need to nuance

---

a pluralist view of the grounds of free speech is widely held in the scholarly literature and avoids the challenge of having to decide what the *real* justification for free speech is. For one pluralist view, see Cohen, "Freedom of Expression."

27 See Howard, "Dangerous Speech," for an argument along these lines in the context of speech calling for wrongful violence.

28 For some speech act theorists, the question of which illocutionary act is performed may even be more deeply and metaphysically indeterminate. Some argue, for example, that a single utterance can have multiple illocutionary forces—and not just because of a divergence between the primary and secondary speech act (as when a speaker asks a question in order to make a request) but also because of different interpretations at one of those levels of analysis. For relevant discussion, see Sbisà, "Some Remarks about Speech Act Pluralism"; Johnson, "Investigating Illocutionary Monism," "Mansplaining and Illocutionary Force," and "Illocutionary Relativism"; and Lewiński, "Illocutionary Pluralism." Even

the illocutionary approach in its application to online content moderation: a pure version of the approach can tell us only what to do with speech deemed to perform violating or nonviolating illocutionary acts; it has nothing to say about the (potentially many) cases where there is genuine uncertainty about which type of act was performed.

4.3. A Hybrid Approach

We suggest that a hybrid approach to content moderation will be most defensible. Such an approach, we argue, should synthesize illocutionary and perlocutionary considerations in a systematic and predictable way rather than haphazardly conflating the two, as in current platform policy. There are potentially several different ways to pursue a hybrid approach. For example, one might treat violating illocutionary force and harmful perlocutionary effects as individually necessary and jointly sufficient conditions for enforcement action to be taken (such that a post must be reasonably interpreted as a threat or call to violence *and* likely to lead to violence, in order to be a legitimate target for moderation). Or one might treat each condition as individually sufficient (such that a post must be reasonably interpreted as a threat or call to violence *or* likely to lead to violence, in order to be a legitimate target for moderation).

The arguments in the previous two subsections, however, point to an important asymmetry between illocutionary and perlocutionary aspects of speech: illocutionary force is far more closely linked to speakers' rights and duties and thus to predictable, justifiable restrictions of their speech, whereas perlocutionary effects are more arbitrary. Accordingly, we propose an "illocutionary-first" moderation strategy. The strategy is comprised of three rules:

1. If an utterance performs a legitimate illocutionary act (such as a criticism), it should not be moderated.
2. If an utterance performs an illegitimate illocutionary act (such as a threat or a call for violence), it should be moderated.
3. If an utterance's illocutionary status is ambiguous between legitimate and illegitimate acts, it should be moderated only if it has net harmful perlocutionary effects.

In effect, we propose to supplement the illocutionary approach described in section 4.2 with a procedure for resolving cases of uncertainty—namely, to

---

if the pluralist view is correct, we believe our proposed approach remains the right one: where multiple interpretations are available (regardless of whether they are understood in terms of illocutionary ambiguity or illocutionary pluralism), content moderators will need to decide what to do about the potentially violating speech; and this is when they should look to perlocutionary effects.

conduct a cost-benefit analysis, weighing the probability of harm against the probability of benefit (roughly along the lines of Meta's "newsworthiness" test).

In principle, other resolution procedures are available at step iii. Most obviously, one could simply apply a blanket approach in cases of uncertainty, either subjecting them all to moderation (on the basis that the speakers should have been clearer about their illocutionary intents) or exempting them all (and giving speakers the benefit of the doubt). However, we believe these blanket approaches fail to track speakers' rights and responsibilities sufficiently closely.

First, speakers have substantial interests in expressing themselves, even if they are not always perfectly clear about what they are saying or doing with their words. Given the scope for performing different illocutionary acts with the same words, even in similar contexts, and given humans' bounded capacities to optimize language choice (considering, for example, our imperfect knowledge, finite vocabularies, performance frailties, and so on), it would be unreasonable to expect all utterances to have clear illocutionary force. A blanket policy of shutting down all *potentially* violating speech would therefore involve unacceptable costs for speakers' rights.

On the other side, a blanket exemption for such speech would be too risky. In certain cases, ambiguous speech is dangerous. Consider President Donald Trump's exhortation to his followers to "fight like hell."[29] Was this an illocutionary act of calling for his supporters to attack the Capitol? Or was it merely advocating vigorous but peaceful protest? It was ambiguous. Trump, we suggest, had a moral duty to be clearer about what precisely he meant in that case; it is this duty that explains why Trump would not have been wronged had the ambiguous speech been taken down in such a case. Here the crucial distinction is between *innocuous ambiguity*—where the illocutionary force is ambiguous, but it is no big deal because perlocutionary risks are low—and *dangerous ambiguity*—where the illocutionary force is ambiguous, and there is indeed a serious risk of harm. In the latter cases (but not the former), speakers have a duty to clarify what they mean. If they do not, they are liable to have their speech moderated. (One could imagine a mechanism whereby ambiguous posts trigger an auto-prompt, encouraging speakers to clarify the meaning of their post.) This is the rationale for appealing to perlocutionary effects as the arbitrating factor: *when the stakes are high, users must take pains to clarify the legitimate illocutionary function of their speech.*

What does this mean for the "death to Khamenei" case? If the Oversight Board was right that this speech, in context, was clearly a mere criticism, the first step of our illocutionary-first approach requires that the post be left

29 See Naylor, "Read Trump's Jan. 6 Speech."

unmoderated. However, if, as we suspect, it was genuinely unclear whether the post was threatening or calling for violence, the third step in our procedure turns to the likely perlocutionary effects. At this point, we agree with Meta that the risk of harm to Khamenei was negligible and outweighed by the opportunity to raise awareness and foster political resistance.

Our illocutionary-first strategy nicely reflects the underlying purpose of content moderation. Ultimately, platforms seek to restrict speech because of its connection to wrongful harm. The reason why illocutionary acts are appropriate targets for moderation is because of the perlocutionary effects they normally function to produce without being subject to all sorts of contingent factors affecting whether or not those effects are produced on any given occasion. In other words, the focus on illocution never takes harmfulness out of the picture but seeks to counter it in a way that best respects a speaker's freedom of expression. This reveals the illocutionary-first strategy to be thoroughly in the business of finding the appropriate balance between free speech and the avoidance of harm—and makes the appeal to perlocutionary effects in cases of uncertainty a very natural one.

Further, our approach avoids counterintuitive implications of a purely perlocutionary approach. Wherever ambiguous speech is removed under an illocutionary-first approach, the user can typically rearticulate their post so as to produce an illocutionary act that is more clearly permissible (and thus exempt from moderation). In contrast, under a purely perlocutionary strategy, this option would not be available in the same way, since no illocutionary acts would be automatically protected (instead requiring case-by-case assessment of perlocutionary effects).

In closing, our proposal is decidedly *not* that platforms should start adding more philosophical jargon to their rules. Rarefied Austinian terminology is, we suspect, best left to philosophy journals. The point instead is that platforms should recognize that when enforcing their rule against, for example, threats, they should focus first on whether the speech *constitutes* a threat—something determined by what reasonable audiences would likely infer. If it does not, it should be allowed. If it definitely constitutes a threat, it should be taken down. And if it is unclear, a cost-benefit analysis of likely effects is necessary. Platform employees should be able to understand that order of operations without any fancy nomenclature.

5. CONCLUSION

We have argued for an illocutionary-first approach to online content moderation that primarily enforces against violating illocutionary acts while protecting

those that are nonviolating. The proposed approach has direct implications for moderation practices: human reviewers should first seek to establish the illocutionary force of a piece of online speech (say, threatening, calling for violence, or merely criticizing) and assess its perlocutionary effects (say, instigating violence or peaceful protest) only in cases of genuine illocutionary uncertainty. By the same token, automated moderation systems should not integrate signals relating to perlocutionary effects (such as virality) with signals relating to illocutionary force (such as common rhetorical use in political protest).

While we have arrived at this position by closely examining the specific case of a Facebook user posting "death to Khamenei," it clearly applies well beyond the individual post, the Violence and Incitement Community Standard, and the Facebook platform to policies adopted by social media platforms in general.[30] In this way, we take ourselves to be putting forward a foundational principle that will help ensure that ever larger swaths of speech in the increasingly online world can be justly and robustly supported.[31]

University College London
sarah.a.fisher@ucl.ac.uk
jeffrey.howard@ucl.ac.uk

REFERENCES

Austin, J. L. *How to Do Things with Words*. Oxford: Clarendon Press, 1962.

Cohen, Joshua. "Freedom of Expression." *Philosophy and Public Affairs* 22, no. 3 (Summer 1993): 207–63.

Dotson, Kristie. "Tracking Epistemic Violence, Tracking Practices of Silencing." *Hypatia* 26, no. 2 (Spring 2011): 236–57.

Fogal, Daniel, Daniel W. Harris, and Matt Moss, eds. *New Work on Speech Acts.* Oxford: Oxford University Press, 2018.

Hesni, Samia. "Illocutionary Frustration." *Mind* 127, no. 508 (October 2018): 947–76.

Hornsby, Jennifer. "Disempowered Speech." *Philosophical Topics* 23, no. 2 (Fall

30 As this paper was headed for publication, the Meta Oversight Board published a new decision exhibiting many of the same issues that we identify here. See decision 2023-032-IG-UA, available at https://www.oversightboard.com/decision/IG-6BZ783WQ.

31 We are grateful for feedback from audiences at the 2023 conference of the Society for Applied Philosophy and the 2023 Joint Session of the Aristotelian Society and the Mind Association, as well as from two anonymous reviewers for this journal. The work was funded by UK Research and Innovation (grant reference MR/V025600/1).

1995): 127–48.
Howard, Jeffrey W. "Dangerous Speech." *Philosophy and Public Affairs* 47, no. 2 (2019): 208–54.
———. "The Ethics of Social Media: Why Content Moderation Is a Moral Duty." *Journal of Practical Ethics* (forthcoming).
Johnson, Casey Rebecca. "Illocutionary Relativism." *Synthese* 202, no. 3 (2023): 1–18.
———. "Investigating Illocutionary Monism." *Synthese* 196, no. 3 (March 2019): 1151–65.
———. "Mansplaining and Illocutionary Force." *Feminist Philosophy Quarterly* 6, no. 4 (2020).
Kukla, Rebecca. "Performative Force, Convention, and Discursive Injustice." *Hypatia* 29, no. 2 (Spring 2014): 440–57.
Kurzon, Dennis. "The Speech Act Status of Incitement: Perlocutionary Acts Revisited." *Journal of Pragmatics* 29, no. 5 (May 1998): 571–96.
Langton, Rae. "Blocking as Counter-speech." In Fogal, Harris, and Moss, *New Work on Speech Acts,* 144–64.
———. "Speech Acts and Unspeakable Acts." *Philosophy and Public Affairs* 22, no. 4 (Autumn 1993): 293–330.
Langton, Rae, and Jennifer Hornsby. "Free Speech and Illocution." *Legal Theory* 4, no. 1 (1998): 21–37.
Lewiński, Marcin. "Illocutionary Pluralism." *Synthese* 199, nos. 3–4 (2021): 6687–714.
Maitra, Ishani. "Silencing Speech." *Canadian Journal of Philosophy* 39, no. 2 (June 2009): 309–38.
McGowan, Mary Kate. "Conversational Exercitives: Something Else We Do with Our Words." *Linguistics and Philosophy* 27, no. 1 (February 2004): 93–111.
———. "Oppressive Speech." *Australasian Journal of Philosophy* 87, no. 3 (2009): 389–407.
Naylor, Brian. "Read Trump's Jan 6. Speech, A Key Part of Impeachment Trial." NPR, February 10, 2021. https://www.npr.org/2021/02/10/966396848/read-trumps-jan-6-speech-a-key-part-of-impeachment-trial.
Sbisà, Marina. "How to Read Austin." *Pragmatics* 17, no. 3 (January 2007): 461–73.
———. "Some Remarks about Speech Act Pluralism." In *Perspectives on Pragmatics and Philosophy*, edited by Alessandro Capone, Franco Lo Piparo, and Marco Carapezza, 227–44. New York: Springer, 2013.
Schiller, Henry Ian. "Illocutionary Harm." *Philosophical Studies* 178, no. 5 (May 2021): 1631–46.

Searle, John R. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press, 1969.

———. "A Taxonomy of Illocutionary Acts." In *Language, Mind and Knowledge*, edited by Keith Gunderson, 344–69. Minneapolis: University of Minnesota Press, 1975.

Strawson, P. F. "Intention and Convention in Speech Acts." *Philosophical Review* 73, no. 4 (October 1964): 439–60.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.2562

THE NEED FOR MERELY POSSIBLE PEOPLE

Johan E. Gustafsson

In his brief study of population ethics, W. V. Quine declared that the only interests that matter are those of actual people. While the interests of future people matter, the interests of possible yet nonactual future people do not. That is, merely possible people do not matter.[1] This *Actual-Population Restriction* is needed in order to avoid recognizing any present yet unactual possibilities, something that Quine—for independent reasons—was eager to resist.[2]

It is well known that the Actual-Population Restriction, which depends on what is actual, can lead to *normative variance*—that is, that what ought to be done in a situation can depend on what would be done in that situation.[3] This, however, is a shared problem for actualist forms of consequentialism. In this paper, I will present a new problem for the Actual-Population Restriction.

1 Quine writes:

> A formulation is ready to hand which sustains the moral values that favour limiting the population while still safeguarding the environment. Namely, it is a matter of respecting the future interests of people now unborn, but only of future actual people. We recognize no present unactualized possibilities. ("On the Nature of Moral Values," 45)

Much the same restriction is defended by Warren, "Do Potential People Have Moral Rights?" 285; Bigelow and Pargetter, "Morality, Potential Persons and Abortion," 173–75; Harman, "Creation Ethics," 311; Parsons, "Axiological Actualism," 142; and Cohen, "An Actualist Explanation of the Procreation Asymmetry," 72–73.

2 See Quine, "On What There Is," 23–4.

3 Arrhenius, "Future Generations," 140–41; and Hare, "Voices from Another World," 503. For some objections to normative variance, see Prichard, *Duty and Ignorance of Fact*, 26; and Carlson, *Consequentialism Reconsidered*, 100–2. Hare also objects that there will be moral dilemmas if we accept the Actual-Population Restriction—for instance, a choice between (i) creating Alice at a negative level of well-being and (ii) creating Bob at a negative level of well-being (503–8). But Hare's examples are neither *obligation dilemmas*—that is, situations where more than one option is obligatory—nor *prohibition dilemmas*—that is, situations where each option is wrong. (See Vallentyne, "Two Types of Moral Dilemmas," 302.) Rather, Hare's examples are *variance dilemmas*—that is, situations where (i) there is at least one option that is not wrong, (ii) there are not two or more options each of which is obligatory, and (iii) each option would be wrong if it were chosen. See Gustafsson, "Is Objective Act Consequentialism Satisfiable?" 194n3.

As for Quine's overall view of ethics, he favored consequentialism, and he suggested (but did not endorse) that utilitarianism may be a systematization of our values.[4] While we will show that the problem for the Actual-Population Restriction also applies if the restriction is combined with nonutilitarian views, we start by combining the restriction with utilitarianism.

Consider the following case:

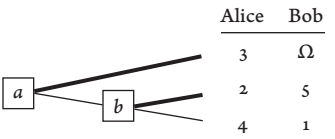

FIGURE 1 Case One

Here, the boxes represent choice nodes. If you were to go up at node *a*, only Alice would exist and she would have a well-being of 3. If you were to go down at node *a*, there would be a second choice at node *b*. At node *a*, you do not (at the time you face node *a*) have any voluntary control of what you would choose later at node *b*. At node *b*, you have a choice between going up, which gives Alice a well-being of 2 and Bob a well-being of 5, and going down, which gives Alice a well-being of 4 and Bob a well-being of 1. The thicker lines represent the choices you would make at each choice node if you were to reach that node. Hence, in this case, it is stipulated that you would go up (even though you *could* go down) at each choice node. Moreover, you know this in advance—that is, you know at node *a* that you would go up at node *b*.

The sequential form of Case One is crucial.[5] The choice at node *a* determines whether Bob will exist. At node *b*, Bob already exists (or at least his

4 Quine writes:

> There is a legitimate mixture of ethics with science that somewhat mitigates the methodological predicament of ethics. Anyone who is involved in moral issues relies on causal connections. Ethical axioms can be minimized by reducing some values causally to others; that is, by showing that some of the valued acts would already count as valuable anyway as means to ulterior ends. Utilitarianism is a notable example of such systematization. ("On the Nature of Moral Values," 43–44)

In Bergström and Føllesdal, "Interview with Willard Van Orman Quine in November 1993," however, Quine withholds truth from ethical statements altogether (202–4).

5 If this were a synchronic choice between the three potential outcomes, we would merely have a case of normative variance—that is, we would find that, if the top outcome is chosen, only the bottom outcome would be permitted but, if either of the two lower outcomes were chosen, only the middle outcome would be permitted. This may be weird, but it is not a violation of Weak Sequential (or Weak Anonymous) Status-Confined Pareto.

existence is guaranteed). Accordingly, Bob can only be created in one way (that is, by going down at node *a*).

If you were to reach node *b*, both Alice and Bob must then be actual since they would exist in all of the then still possible outcomes.[6] Hence the interests of both Alice and Bob would matter at node *b*. Accordingly, Actual-Population Utilitarianism would prescribe going up at node *b* (since going up has a greater sum total of well-being than going down for the people whose interests would matter).

Note that you *will* go up at node *a* (as stipulated in the description of the case). So the only person who will actually exist is Alice. Hence she is the only person whose interests matter at node *a*. Since you know at node *a* what you would do if you were to reach node *b*, you can use *backward induction*, which is to predict what you would choose at later choice nodes and to take those predictions into account when you choose at earlier nodes.[7] To use backward

6 We are assuming here that, if you were to reach node *b*, the people who would then be actual (and thus whose interest would matter) would be the people who would exist if you were to do what you would do at that node. This is consistent with Hare's Strong Actualism (roughly, the view that you should assess all options at a choice node, taking into account only the people who would exist if you did what you would do at that node), since this is a separate application of the theory at a new choice node ("Voices from Another World," 503). Understood in this way, the actualism we consider in this paper is Strong Actualism. Hare's Weak Actualism (roughly, the view that you should assess each option at a choice node, taking into account only the people who would exist if you chose that option) leads to violations of some compelling principles of deontic logic ("Voices from Another World," 502, 504–6). An alternative way of applying the restriction—call it *Super-Strong Actualism*—is to apply it rigidly to the actual world even at choice nodes that will not be reached—for instance, to apply it relative to the world where you go up at node *a* even at node *b*. But how could this be how you would apply the theory at node *b*? How would you know that you are not in the actual world? What is relevant for reasoning by backward induction is how the theory would be applied at each node. Furthermore, applying the restrictions this way may lead to Pareto violations at the choice nodes that would not be reached. To see this, suppose that Alice would, instead of 4, get a well-being of 2 if you went down at node *b*. Then going down at node *b* would not affect those whose interests matter (that is, Alice), but Bob will be worse off than if you went up at that node.

7 For backward induction, see von Neumann and Morgenstern, *Theory of Games and Economic Behavior*, 116–17; Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games"; and Rosenthal, "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," 94–95. The first two cases in this paper are *BI-terminating* decision problems—that is, the choices that are prescribed by backward induction in these cases are final in the sense that they are not followed by any further choices. (See Rabinowicz, "Grappling with the Centipede," 101.) Crucially, in BI-terminating decision problems, the choices that are prescribed by backward induction can be defended with very minimal assumptions. Notably, we do not need the controversial assumption that agents would choose rationally at nodes that can only be reached through

induction to determine what you ought to do is a form of *actualism*, since it makes what you ought to do now depend on what (you predict) you would do in the future.[8] Actualism is a rival to *possibilism*—the view that what you ought to do now depends on what you could (rather than just what you would) do in the future.[9] (To avoid terminological confusion, note that actualism, in the sense of taking into account what you would do in the future, is distinct from the Actual-Population Restriction—that only the interests of actual people matter.) Actualism fits better with the motivation for the Actual-Population Restriction than possibilism, since actualism restricts the morally relevant acts to those that are actual (or would be actual). So we assume actualism for now, but later on we will explore the Actual-Population Restriction given possibilism.

Using backward induction, you take into account that you would follow Actual-Population Utilitarianism's prescription to go up at node *b*. So you find that going up at node *a* is better than going down for everyone whose interests matter, because Alice (the only person whose interests matter at node *a*) would get a well-being of 3 if you were to go up and a well-being of just 2 if you were to go down (since you would go up at node *b*). Accordingly, Actual-Population Utilitarianism entails that you *ought to* go up at node *a*.

The trouble is that going up is worse for everyone whose interests matter (that is, Alice) than the alternative sequence of choices consisting in going down at both choice nodes. If you were to go up at node *a*, Alice would get a well-being of 3, but, if you were to go down at both choice nodes, Alice would get a well-being of 4. Thus the choices that Actual-Population Utilitarianism prescribes in this case (going up at each choice node) are worse than the opposite choices (going down at each choice node) for everyone whose interests matter (everyone who actually exists). We have a violation of the following principle:

> *Weak Sequential Status-Confined Pareto*: If (i) outcome *X* is better than outcome *Y* for everyone whose interests matter and (ii) *X* is the outcome of an available sequence of choices, then it is not the case that each choice in a sequence of choices with outcome *Y* ought to be made.

Violations of this principle are worrying, since they entail that, for the only people whose interests matter, the prescriptions of the violating theory would make things worse. The choices that Actual-Population Utilitarianism

---

irrational choices. See Broome and Rabinowicz, "Backwards Induction in the Centipede Game," 240–41.

8 Jackson and Pargetter, "Oughts, Options, and Actualism," 233.

9 Jackson and Pargetter, "Oughts, Options, and Actualism," 233.

prescribes in Case One are worse than the opposite choices for everyone whose interests matter according to the Actual-Population Restriction. This cannot be right.[10]

So far, we have assumed that the Actual-Population Restriction would be combined with utilitarianism. But the objection to the Actual-Population Restriction needs only fairly minimal ethical assumptions.

To reach the conclusion that the outcome of going up is better than the outcome of going down at node *b*, we need only the following principle:[11]

> *Weak Anonymous Status-Confined Pareto*: If (i) outcome *X* is better than outcome *Y* for everyone whose interests matter and (ii) *Y* is just like outcome *Z* except that the identities of some people whose interests matter have been permuted, then *X* is better than *Z*.

Let $\langle u, v \rangle$ denote an outcome where Alice gets a well-being of *u* and Bob gets a well-being of *v*. Given that the interests of both Alice and Bob matter (since they would both be actual at node *b*), we find that $\langle 2, 5 \rangle$ is like $\langle 5, 2 \rangle$ except that the identities of some people whose interests matter have been permuted. Since $\langle 5, 2 \rangle$ is better than $\langle 4, 1 \rangle$ for everyone whose interests would matter at node *b*, Weak Anonymous Status-Confined Pareto entails that $\langle 2, 5 \rangle$ is better than $\langle 4, 1 \rangle$. Accordingly, the outcome of going up at node *b*, $\langle 2, 5 \rangle$, is better than the outcome of going down at that node, $\langle 4, 1 \rangle$.

Similarly, to reach the conclusion that the outcome of going up at node *a* is better than the outcome of going down at that node, we need only backward induction and Weak Anonymous Status-Confined Pareto. Since you would go up at node *b*, we find by backward induction that Alice would get a well-being of 2 if you were to go down at node *a*, which is lower than her well-being would be if you were to go up at node *a*. Since Alice is the only person whose interests matter (she is the only one who actually exists), we find by Weak Anonymous Status-Confined Pareto that the outcome of going up at node *a* is better than the outcome of going down at that node.

Hence the above objection works against the Actual-Population Restriction in combination with backward induction and any consequentialist theory that satisfies Weak Anonymous Status-Confined Pareto. (As we shall see later on, the objection also works against possibilist consequentialist theories.)

10 Do not be distracted by the observation that, if you were to go down at each node, then both Alice and Bob would actually exist and the interests of both of them would matter. The crucial thing for Actual-Population Utilitarianism is that you actually will not go down at node *a*, and therefore Bob does not actually exist—and his interests do not matter.

11 This principle is a variation of a principle in Sen, *Collective Choice and Social Welfare*, 153.

So far, we have not relied on any specific version of utilitarianism. But, if we do, we can strengthen the objection. Given either a total or an average version of Actual-Population Utilitarianism, it violates the following principle:

> *Weak Sequential Fixed-Population Pareto*: If (i) outcome *X* is better than outcome *Y* for everyone who exists in these outcomes, (ii) the same people exist in *X* and *Y*, and (iii) *Y* is the outcome of an available sequence of choices, then it is not the case that each choice in a sequence of choices whose outcome is *X* ought to be made.

Violations of this principle should be even more worrying than the violation of Weak Sequential Status-Confined Pareto in Case One. If you go up at node *a* of Case One and thereby violate Weak Sequential Status-Confined Pareto, the only person whose interests matter (Alice) is worse off than if you had gone down at all choice nodes. But, if you had gone down at all choice nodes and realized the dominating outcome, there would have been an additional person (Bob) whose interests would have mattered. On the other hand, if Weak Sequential Fixed-Population Pareto is violated, the dominating outcome has the same population as the dominated outcome.

To see how we get violations of Weak Sequential Fixed-Population Pareto, we start with the total version of Actual-Population Utilitarianism:

> *Total Actual-Population Utilitarianism*: An outcome *X* is at least as good as an outcome *Y* if and only if the sum total of well-being in *X* for people who actually exist and who also exist in *X* is at least as great as the sum total of well-being in *Y* for people who actually exist and who also exist in *Y*.

In other words, this view is the same as standard total utilitarianism except that the well-being of people who do not belong to the actual population is ignored.

Consider the following case:

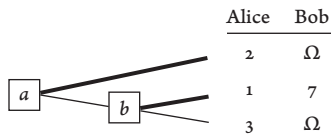

FIGURE 2 Case Two

At node *a*, you do not (at the time you face node *a*) have any voluntary control of what you would choose later at node *b*. At node *b*, it is stipulated that you would go up if you were to reach that node. And you know this in advance—that is, you know at node *a* that you would go up at node *b*. So, if you were

to reach node *b*, the actual population would include both Alice and Bob. Accordingly, Total Actual-Population Utilitarianism would prescribe going up at node *b*, since the total well-being for Alice and Bob is 8 if you go up but only 3 if you go down.

At node *a*, it is likewise stipulated that you will go up at that node. So the actual population includes only Alice. Using backward induction, you take into account that, if you were to reach node *b*, you would (following Total Actual-Population Utilitarianism) go up at that node. Alice gets a well-being of 2 if you go up at node *a*, and she would get a well-being of 1 if you were to go down at node *a* (since you would go up at node *b*). Accordingly, Total Actual-Population Utilitarianism prescribes going up at node *a*, since Alice is the only person in the actual population. But then we have a violation of Weak Sequential Fixed-Population Pareto, since we end up with an outcome where only Alice exists and her well-being is 2 but, if you had gone down at all choice nodes, only Alice would exist and her well-being would have been 3.

Next, we turn to the average version of Actual-Population Utilitarianism.

> *Average Actual-Population Utilitarianism*: An outcome *X* is at least as good as an outcome *Y* if and only if the average of well-being in *X* for people who actually exist and who also exist in *X* is at least as great as the average of well-being in *Y* for people who actually exist and who also exist in *Y*.

In other words, this view is the same as standard average utilitarianism except that people who do not belong to the actual population do not count toward the average of well-being.

Consider once more Case Two. At node *b*, since you would go up at that node, the actual population would include both Alice and Bob. So the average well-being in the outcome of going up for the actual population is 4. And the average well-being in the outcome of going down for those in the actual population who also exist in that outcome (namely, just Alice) is 3. (Bob would be actual if you were to reach node *b*, but he does not exist in the outcome of going down.) Accordingly, Average Actual-Population Utilitarianism prescribes going up at node *b*.

At node *a*, since you will go up at that node, the actual population includes only Alice. Using backward induction, you take into account that, if you were to reach node *b*, you would (following Average Actual-Population Utilitarianism) go up at that node. Hence the average well-being in the outcome of going up for the actual population is 2, and the average well-being in the outcome of going down for those in the actual population who also exist in that outcome (namely, Alice) is 1. (Bob would exist if you were to go down at node *a*, but he is

not actual.) Accordingly, Average Actual-Population Utilitarianism prescribes going up at node *a*.

We find that Average Actual-Population Utilitarianism prescribes the same options in Case Two as Total Actual-Population Utilitarianism. So, like Total Actual-Population Utilitarianism, Average Actual-Population Utilitarianism violates Weak Sequential Fixed-Population Pareto.

So far, we have relied on backward induction and actualism. It may be objected that the Actual-Population Restriction and Actual-Population Utilitarianism avoid trouble in Cases One and Two if they are instead coupled with possibilism. Given possibilism, you take into account all the things you could do and choose according to one of the optimal plans you could possibly follow. That is, (i) you consider the outcomes of all available plans and assess which of these outcomes is optimal in a choice between all of them, and (ii) you ought to choose in accordance with a plan whose outcome is optimal—without taking into consideration whether you would later depart from that plan.[12]

Given this form of possibilism, it is no longer the case that you ought to go up at node *a* in Case One. Does this bar the earlier objection to the Actual-Population Restriction? To see that it does not, consider once more Case One—but now we mark what ought to be done given possibilism with dashed lines (the thick lines still denote what you would do at each choice node):

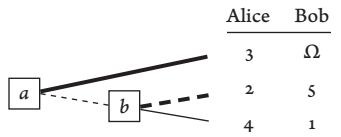

FIGURE 3 Case One (with Possibilist Prescriptions)

Since you will actually go up at node *a*, only Alice is actual. So only Alice's interests matter at node *a*. The best you can do for Alice is to go down at each choice node. At node *a*, Weak Anonymous Status-Confined Pareto entails that the outcome of going down at each choice node is better than the outcome of any other available sequence of choices. So, given possibilism, you ought to go down at node *a*. And, if you were to reach node *b*, you would go up at that node. So both Alice and Bob would be actual, and their interests would matter if you were to reach node *b*. So, at node *b*, it follows (in the same way as before) by Weak Anonymous Status-Confined Pareto that the outcome of going up is better than the outcome of going down. So, given possibilism, you ought to go

12 In decision theory, this approach is known as *naive choice*. See Pollak, "Consistent Planning," 202–3; and Hammond, "Changing Tastes and Coherent Dynamic Choice," 162.

up at node *b*. Hence each choice in the sequence of choices consisting in going down at node *a* and going up at node *b* ought to be made, given possibilism. But the outcome of going up at node *a* is better for everyone whose interests matter (namely, Alice) than the outcome of the sequence of going down at node *a* and up at node *b*. Hence we still have a violation of Weak Sequential Status-Confined Pareto in Case One.

But how about the objection that Total Actual-Population Utilitarianism and Average Actual-Population Utilitarianism both violate Weak Sequential Fixed-Population Pareto in Case Two? In fact, given possibilism, that objection no longer works in Case Two. To see this, consider that case once more but with what ought to be done given possibilism marked with dashed lines:

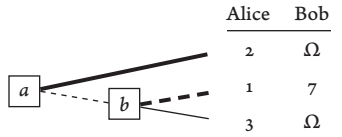

FIGURE 4 Case Two (with Possibilist Prescriptions)

At node *a*, you will go up. So, at node *a*, only Alice is actual, and so only her interests matter. Then, at node *a*, Total Actual-Population Utilitarianism and Average Actual-Population Utilitarianism each entails that the outcome of going down at each choice node is better than the outcome of any other available sequence of choices. So, given possibilism, you ought to go down at node *a*. But then we have no violation of Weak Sequential Fixed-Population Pareto.

Nonetheless, Total Actual-Population Utilitarianism and Average Actual-Population Utilitarianism still violate Weak Sequential Fixed-Population Pareto in another case. Consider the following:

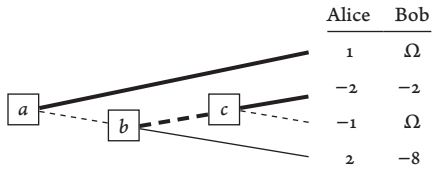

FIGURE 5 Case Three

At each choice node, you do not (at the time you face that node) have any voluntary control of what you would choose later at future nodes. It is stipulated that, at each choice node, you would go up (even though you could go down). So, at node *a*, only Alice is actual, and so only her interests matter. Therefore, according to both Total Actual-Population Utilitarianism and Average

Actual-Population Utilitarianism, the best outcome of any available sequence is the outcome of going down at both node *a* and node *b*. Hence, given possibilism, you ought to go down at node *a*.

If you were to reach node *b* or node *c*, both Alice and Bob would be actual (since you would go up at those nodes). So, at nodes *b* and *c*, the best outcome of any available sequence of choices is the outcome of going up at node *b* and then down at node *c*. Therefore, according to both Total Actual-Population Utilitarianism and Average Actual-Population Utilitarianism, you ought to go up at node *b* and then down at node *c*. Hence, given possibilism, each choice in the sequence of choices consisting in going down at node *a*, going up at node *b*, and going down at node *c* ought to be made. But the outcome of going up at node *a* is better for everyone whose interests matter (namely, Alice) than the outcome of the sequence of going down at node *a*, up at node *b*, and down at node *c*. Hence, even given possibilism, we have a case where Total Actual-Population Utilitarianism and Average Actual-Population Utilitarianism violate Weak Sequential Fixed-Population Pareto.[13]

In summation, Quine's Actual-Population Restriction leaves us with an implausible population ethics. The silencing of the interests of merely possible people comes with a cost to actual people.[14]

University of Texas at Austin
University of York
Institute for Futures Studies
johan.eric.gustafsson@gmail.com

13 It may be objected that you actually violate the prescriptions of the possibilist versions of Total and Average Actual-Population Utilitarianism in Case Three. But note that all available sequences of choices in this case would violate those prescriptions. Going up at node *a* violates the prescription at that node. Going down at node *a*, up at node *b*, and up at node *c* would violate the prescription at node *c*. Going down at node *a*, up at node *b*, and down at node *c* would violate the prescription at node *b*. Going down at node *a* and down at node *b* would violate the prescription at node *b*. Hence these possibilist theories are sequentially unsatisfiable: there is no available sequence of choices in Case Three such that, if you were to make that sequence of choices, you would not violate the theory (even though these theories are satisfiable at each choice node).

14 I wish to thank John Broome, Krister Bykvist, Hilary Greaves, Caspar Hare, Wlodek Rabinowicz, Melinda Roberts, and an anonymous referee for valuable comments. Financial support from the Musk Foundation and the Swedish Foundation for Humanities and Social Sciences is gratefully acknowledged.

REFERENCES

Arrhenius, Gustaf. "Future Generations: A Challenge for Moral Theory." PhD diss., Uppsala University, 2000.

Bergström, Lars, and Dagfinn Føllesdal. "Interview with Willard Van Orman Quine in November 1993." *Theoria* 60, no. 3 (December 1994): 193–206.

Bigelow, John, and Robert Pargetter. "Morality, Potential Persons and Abortion." *American Philosophical Quarterly* 25, no. 2 (April 1988): 173–81.

Broome, John, and Wlodek Rabinowicz. "Backwards Induction in the Centipede Game." *Analysis* 59, no. 4 (October 1999): 237–42.

Carlson, Erik. *Consequentialism Reconsidered*. Dordrecht: Kluwer, 1995.

Cohen, Daniel. "An Actualist Explanation of the Procreation Asymmetry." *Utilitas* 32, no. 1 (March 2020): 70–89.

Gustafsson, Johan E. "Is Objective Act Consequentialism Satisfiable?" *Analysis* 79, no. 2 (April 2019): 193–202.

Hammond, Peter J. "Changing Tastes and Coherent Dynamic Choice." *Review of Economic Studies* 43, no. 1 (February 1976): 159–73.

Hare, Caspar. "Voices from Another World: Must We Respect the Interests of People Who Do Not, and Will Never, Exist?" *Ethics* 117, no. 3 (April 2007): 498–523.

Harman, Elizabeth. "Creation Ethics: The Moral Status of Early Fetuses and the Ethics of Abortion." *Philosophy and Public Affairs* 28, no. 4 (October 1999): 310–24.

Jackson, Frank, and Robert Pargetter. "Oughts, Options, and Actualism." *Philosophical Review* 95, no. 2 (April 1986): 233–55.

Parsons, Josh. "Axiological Actualism." *Australasian Journal of Philosophy* 80, no. 2 (June 2002): 137–47.

Pollak, R. A. "Consistent Planning." *Review of Economic Studies* 35, no. 2 (April 1968): 201–8.

Prichard, H. A. *Duty and Ignorance of Fact*. London: Humphrey Milford, 1932.

Quine, W. V. "On the Nature of Moral Values." In *Values and Morals: Essays in Honor of William Frankena, Charles Stevenson, and Richard Brandt*, edited by Alvin I. Goldman and Jaegwon Kim, 37–45. Dordrecht: Reidel, 1978.

———. "On What There Is." *Review of Metaphysics* 2, no. 5 (September 1948): 21–38.

Rabinowicz, Wlodek. "Grappling with the Centipede: Defence of Backward Induction for BI-Terminating Games." *Economics and Philosophy* 14, no. 1 (April 1998): 95–126.

Rosenthal, Robert W. "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox." *Journal of Economic Theory* 25, no. 1 (August 1981):

92–100.
Selten, R. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4, no. 1 (March 1975): 25–55.
Sen, Amartya K. *Collective Choice and Social Welfare*. San Francisco: Holden-Day, 1970.
Vallentyne, Peter. "Two Types of Moral Dilemmas." *Erkenntnis* 30, no. 3 (May 1989): 301–18.
Von Neumann, John, and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1944.
Warren, Mary Anne. "Do Potential People Have Moral Rights?" *Canadian Journal of Philosophy* 7, no. 2 (June 1977): 275–89.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3131

ADDICTION, RESPONSIBILITY, AND A SORITES PROBLEM

Jeesoo Nam

Suppose we want to find out whether Deanna, a heroin-addicted person, is morally responsible for the death of her two-year-old daughter. As a result of being high on heroin, Deanna forgot about her kid she left unattended in her hot car, and the toddler tragically passed away. Should we condemn Deanna for her reckless heroin use and the resultant harm? In defense of Deanna, she was a heroin-addicted person; her addiction acted as a compulsion to use heroin. Compulsion is one of two commonly accepted excuses to moral responsibility.[1] Thus, Deanna is excused for her heroin use and its resultant harm to her daughter.

A critic of Deanna could respond that Deanna's responsibility can be traced back to a time before she became addicted. In the early days of Deanna's drug use, she exercised free choice to use heroin and knew full well that continued drug use would lead to addiction. Younger Deanna's use of drugs was not compelled by addiction (because she was not yet addicted) but nevertheless began the causal chain ending in the toddler's death. Since Deanna caused the harm as an exercise of free choice, she is morally responsible. In rebuttal, Deanna's defense might appeal to the notion that although younger Deanna knew her continued drug use would lead to addiction, she did not know that it would lead to her daughter's death.[2] Ignorance is the second of the two commonly accepted excuses to moral responsibility.[3]

The line of reasoning just set out is controversial, but much of it seems right to me. For instance, it seems right that one cannot be held responsible for the downstream consequences of one's actions if one did not believe that those consequences would arise. Like most others, I believe ignorance excuses. But is ignorance the only defense available to addicted people? Surely many people starting to use drugs know that becoming addicted will risk severe harm to their families. Are they morally responsible when those harms arise?

1 Fischer and Tognazzini, "The Truth about Tracing," 531.

2 See Moore, "Addiction," 26.

3 Fischer and Tognazzini, "The Truth about Tracing," 531.

This article argues in favor of the following surprising result: Deanna is likely excused from moral responsibility even if at the time she first began using drugs she knew with 100 percent certainty that continued, long-term drug use would lead to her daughter's death. Even if someone knows that long-term drug use leads to harmful consequences, almost no one chooses long-term drug use.[4] Instead, they only choose to do small amounts of drugs such as smoking tonight or opening up a bottle of wine, and those small amounts of drugs happen to end up over time constituting long-term drug use. Since these individuals have never chosen long-term drug use, they cannot be held responsible for the consequences of their long-term drug use.

Secton 1 of this article details the *prima facie* argument that addiction excuses certain wrongful behaviors because addiction is an irresistible compulsion. Section 2 details the rebuttal of tracing the addicted person's responsibility to before he became addicted. The addicted person is supposedly responsible for *starting* to use drugs with the knowledge that continued drug use leads to bad consequences. Section 3 examines how addicted people become addicted. Addicted people do not plan out long-term drug use. Instead, they decide to use drugs in small doses, and those decisions merely add up over time. Section 4 argues that addicted people are not responsible for the bad consequences of their addiction because they never chose long-term drug use. Much of section 4 responds to various objections to this core argument. Last, I generalize the moral framework so that it can be used to analyze a wide variety of similar cases.

1. ADDICTION SOMETIMES EXCUSES

Addicted people cannot be punished for their wrongful acts when their addiction served as a compulsion to do such acts. Take, for instance, the crime of illicit drug use.[5] Some addictions are so strong as to render the addicted person unable to resist using drugs. Punishing an unavoidable act would frustrate the retributivist since inability to do otherwise excuses the actor from moral responsibility.

This line of thinking can be seen in Justice White's opinion in *Powell v. Texas*: "the chronic alcoholic with an irresistible urge to consume alcohol should not be punishable for drinking or for being drunk."[6] For Justice White, the addicted person is excused not only for the unavoidable act, i.e. drinking, but thereby

4 Heyman, *Addiction*, 131.

5 E.g., Cal. Health and Safety Code § 11550 (West 2021). Also consider the criminal acts of underage drinking and "taking pain medications in excess of their prescribed dosages." Yaffe, "Compromised Addicts," 209.

6 *Powell v. Texas*, 392 U.S. 514, 548–49 (1968).

also for the consequence of that act, i.e. being drunk. Other consequences of drug use can be more sinister, such as forgetting about a child left in a hot car. Addiction might also be raised as a defense to crimes committed in pursuit of drug use, such as stealing money in order to buy drugs. It may very well be the case that an addicted person who feels the compulsion to use drugs but has no drugs to use will also feel a compulsion to steal in order to buy drugs. Thus, there are at least three distinct categories of wrongful acts for which addiction could be seen to excuse: the crime of drug use, the acts that result from drug use, and the acts that are undertaken in pursuit of drug use.[7]

There is general agreement among experts in this area that drug addiction (at least sometimes) diminishes (at least partly) moral responsibility for one's wrongful acts.[8] Compulsion is one reason for thinking addicted people are excused, but it is certainly not the only candidate theory in the literature concerning why addicted people are excused.[9] The topic of this article is a counterargument (laid out in the next section) that is supposed to apply to all of the various arguments that addiction excuses, including the compulsion view. Thus, I wish to remain theory neutral at this level. As long as the reader thinks there is at least one category of wrongful acts that addiction sometimes at least partly excuses, this article will have meaningful implications. However, for my own ease of exposition in this article, I will simply take *arguendo* that all addicted people have a compulsion to use drugs and that this compulsion should be considered as corroding the freedom ordinarily necessary for moral responsibility.

2. TRACING

Arguendo, addicted people are yielding to irresistible compulsions. Even so, there is a common rebuttal to the proposition that addiction excuses. Applying a tracing principle, the rebuttal states that most addicted people are responsible for their behavior after they become addicted because, at some earlier point in time, they chose to place themselves in that situation.[10] Even if one would ordinarily be permitted an excuse of compulsion for a wrongful act he did, no such excuse is permitted if one earlier freely put himself in that situation, knowing that he would then be committed to doing that wrongful act. Tracing back

7 Yaffe, "Compromised Addicts," 209–10.

8 For a survey, see Moore, "Addiction," 26.

9 Moore, "Addiction," 26.

10 Moore, "Addiction," 23.

moral responsibility in this way is both plausible and an important component of ethical theory in explaining our intuitions about moral responsibility.[11]

Illustrating by example, suppose that there is a pill that delivers an instant release of dopamine but also drives one to murderous psychopathy—if someone takes the pill, he develops an irresistible lust for killing. Suppose that Daniel took the pill and went into a murderous psychopathy, the consequence of which was the killing of his wife, Valerie. Suppose further that Daniel took the pill for his own pleasure but nevertheless knew that it would drive him to kill someone. Daniel is responsible for Valerie's murder regardless of the fact that Daniel's compulsion to kill Valerie at the time of her murder was so strong as to be irresistible.[12]

A necessary condition of tracing is that the actor was morally responsible for having put himself in the binding situation. If someone force-fed Daniel the pill or Daniel made a nonculpable mistake about the pill's negative effects, he could not be blamed for Valerie's murder.

In the literature, the most prominent charge against tracing back moral responsibility has been that the knowledge requirement is rarely satisfied.[13] For example, defenders of addicted people argue that someone like Deanna (who left her daughter in the car while high on heroin) did not know at the time she started doing drugs that addiction would lead to the eventual death of her child. Her lack of knowledge means that responsibility cannot be traced back.

But, of course, the fact that the knowledge requirement is rarely satisfied does not mean the knowledge requirement is never satisfied. Some people do have knowledge that their continued drug use would lead to an irresistible wrongdoing. For example, it is plausible that someone living in poverty has a justified true belief that his continued drug use will eventually commit him to stealing from others in order to finance his drug addiction. Addicted people

11 Fischer and Tognazzini, "The Truth about Tracing," 552–53. But see Agule, "Resisting Tracing's Siren Song," 1; Alexander, "Causing the Conditions of One's Defense," 623; Khoury, "Responsibility, Tracing, and Consequences," 187; and King, "Traction without Tracing," 463. On the competing account, putting oneself in the dangerous incapacitated state is the wrongful act, for which one is responsible if one does so with the mental state necessary for culpability. The competing account can be more fine grained, since tracing's knowledge requirement can be replaced with the more flexible *mens rea* standard applicable to the wrong for which we are trying to hold the actor responsible. Thus, for example, if theft requires specific intent for moral responsibility, the earlier act of putting oneself into the incapacitated state must also have been done with the specific intent of taking another's property in order for that actor to be responsible for theft. This article's conclusions, which deal with responsibility generally, will apply not only to tracing but also to this alternative to tracing.

12 See also Moore and Hurd, "Punishing the Awkward, the Stupid, the Weak, and the Selfish," 179.

13 Moore, "Addiction," 26; and Vargas, "The Trouble with Tracing," 277.

like these cannot rely on the nonsatisfaction of the knowledge requirement to defend themselves from moral responsibility.

In the tracing literature, there is some disagreement about how narrowly/broadly the boundaries of the knowledge requirement should be drawn.[14] Thus, well-informed readers may disagree with my rough characterization of how often drug users know that continued drug use will commit them to future illicit activity. One might, for example, say that it was not necessary for Deanna to know that she would end up leaving her daughter in her car as a result of her addiction but rather that she merely needed to know that she would end up hurting the people around her. In this article, I want to be theory independent with regard to that debate. I would like to argue that moral responsibility fails to trace back for addicted people regardless of which theory of the knowledge requirement we adopt. Even if Deanna satisfied the knowledge requirement—whatever the knowledge requirement is—she is still likely excused from moral responsibility.

3. VAGUE DESIRE: WHAT'S THE HARM OF JUST ONE MORE?

To see whether moral responsibility can be traced back for addicted people, we need to look at why and how drug users become addicted. The inquiry is complicated by the fact that addiction often has such harmful consequences.[15] Typically, when thinking about why someone performed some action, we would try to figure out why that individual took himself to have an all-things-considered reason to perform that action. But the state of being addicted to the point of compulsion can be such a nightmare that it just does not seem likely that many would consider the benefits of prolonged drug use to outweigh the costs of addiction.[16] If addiction is so harmful to the person addicted, why does anyone use drugs at levels that result in addiction? Call this the *puzzle of addiction*.[17]

In trying to solve this puzzle and other similar puzzles about why people act in ways contrary to their interests, some philosophers and psychologists have

14 Vargas, "The Trouble with Tracing," 277; and Fischer and Tognazzini, "The Truth about Tracing," 537.

15 "Addiction ravages lives and communities" (Elster and Skog, "Introduction," 1). See also Heyman, *Addiction*, 133.

16 Elster and Skog, "Introduction," 1.

17 Elster and Skog, "Introduction," 1. A similar question can be asked of why addicted people continue to use drugs. See Pickard, "The Puzzle of Addiction," 9–10. The more relevant question for this article is why recreational/nonaddicted drug users use drugs at levels that result in addiction. These two questions are distinct.

proposed that such self-harm can result from a series of individually rational behaviors that, collectively, lead to an irrational result.[18] In this section, I will present this general idea and argue that it is a good explanation of how people become addicted even when they desire not to be addicted.

Before getting to addiction, let us begin with an innocent example. Gus is a thin, rational man (exhibiting no weakness of will) with a desire of first lexical priority not to be fat. No other desire could trump his desire not to be fat. So long as he is thin, however, he does not particularly care how much he weighs; whether he weighs 130 pounds or 135 pounds, for instance, makes no difference to him. He merely desires not to be fat. The postman knocks at Gus's door. The postman informs Gus that Gus just won the sweepstakes for a lifetime supply of fun-sized Snickers, an absolute delight for this sweet tooth. As the candy is carried into his living room by the box, he takes out a single piece and considers whether he should eat it. Given his desire not to be fat, should he allow this indulgence?

Gus is a thin man. A fun-sized Snickers has eighty calories, a miniscule difference maker in his weight. Surely, eighty calories do not separate a thin man from a fat man. Eating the Snickers will serve his desire for sweets, but it certainly will not make him fat. And so Gus eats the Snickers and finds it a delight. He has made a rational decision. Then he reaches for a second fun-sized Snickers. Should he eat this one? We can apply the same reasoning here again. Since the previous one did not make him fat, he is still thin. As clear as day, eighty calories cannot make a thin man fat. And so he eats the second Snickers. And the third Snickers? Same here. And so he eats the third. And the fourth. The fifth. And so on. Gus spends the next month popping one Snickers after another. At the end of the month, when he steps on the scale, he comes to find the truth of the following proposition: Gus is fat.[19]

The astute reader will see that this is a variation on the sorites problem. It is the defining feature of vague predicates such as *fat* that if there is some object *a* to which the vague predicate applies and another object *b* that is qualitatively identical to *a* but for a miniscule difference, then the vague predicate applies also to *b*. Following Crispin Wright, call this the *tolerance principle* of vague predicates.[20] There is no point, no boundary line, at which adding another eighty calories will make a thin man fat, just as there is no hair that falls off to make a man bald, no day that passes to make a man old.[21] When the content

18 E.g., Heyman, *Addiction*, 133; Andreou, "Environmental Damage and the Puzzle of the Self-Torturer," 100–1; Edgington, "Vagueness by Degrees," 296; and Schwartz, "Soritic Thinking, Vagueness, and Weakness of Will," 18.

19 Edgington, "Vagueness by Degrees," 296 (setting out the "dieter's paradox").

20 Wright, "On the Coherence of Vague Predicates," 333–34.

21 It would also be easy to reword the whole example using "not fat" in place of "thin."

of a desire employs a vague predicate, the vague desire cannot motivate even a rational person to stop him from making an incremental change.

More precisely, the issue is the result of an interaction between two phenomena: *de minimus* contribution and vagueness. *De minimus* contribution can complicate ethical inquiry in two ways. The first way is when multiple actors each contribute in a *de minimus* manner to some harmful result. For instance, ordinary individuals contribute in extremely minor ways to global warming that, across our global population of eight billion, add up to an existential threat. This first way has been the subject of much discussion in ethics. The second way in which *de minimus* contribution complicates ethical inquiry is the topic of this article: its interaction with vagueness. In this second way, there is only one actor, but he performs many *de minimus* actions. The trouble is that when some undesired outcome is vague, the actor has no reason to refrain from doing those actions that make only *de minimus* contributions towards the undesired outcome. This second problem has received comparatively little attention.

A similar story can be spun about addiction. Suppose that all addiction is is a very high level of desire for a drug.[22] Each additional intake of nicotine increases one's desire for nicotine. At stake in a decision about whether to smoke another cigarette is an extremely minor increase in such a desire. Applying the tolerance principle, smoking an additional cigarette cannot be the difference between addiction and nonaddiction. Therefore, the desire not to be addicted remains causally inert in deciding whether to smoke another cigarette. Both smoking one more cigarette and not smoking the cigarette equally satisfy the desire not to become addicted.

Call the picture described thus far the *Vague Desire* solution to the puzzle of addiction and those people who are accurately described by the picture *Vague Desirers*. How do people become addicted despite their desire not to? The vagueness of their desire not to be addicted renders such desires irrelevant to decision-making about using small dosages of drugs.

The Vague Desire picture relies on two key premises to establish that the desire not to become addicted will not stop the user from using drugs.

22 "Addictive desires are just strong desires toward pleasure" (Foddy and Savulescu, "A Liberal Account of Addiction," 16–17). My main argument will go through no matter what the right conceptual analysis of addiction may turn out to be so long as addiction is a vague predicate and so long as the property that serves as the analysans is increased only incrementally by the performance of the addictive act. So, for example, if addiction turned out to be defined by the severity of withdrawal symptoms, then withdrawal symptoms could take the place of desire for drugs in this article so long as performing the addictive act increased the severity of withdrawal symptoms only incrementally.

Premise 1: The user believes in the tolerance principle with respect to addiction.

By the tolerance principle, an incremental change at the margin does not affect the application of a vague predicate. The tolerance principle is intuitive and the dominant viewpoint of both laymen and experts.[23] Consider, for example, the Roman philosopher Galen's stance on the principle two millennia ago: "I know of nothing worse and more absurd than that the being and nonbeing of a heap is determined by a grain of corn."[24] The tolerance principle sets out in lucid terms what most of us take for granted about vague language. The tolerance principle is so closely intertwined with vagueness that many would consider it a defining feature of vague predicates.[25] Thus, I take both the truth of the tolerance principle as well as the proposition that most people believe the tolerance principle as received wisdom, though the Vague Desire picture is only contingent on the drug user's belief in the tolerance principle.

One way to explicate the tolerance principle is to contrast it with a competing theory's denial of the principle. Under epistemicism, the tolerance principle is false, and there is some incremental change that does make a difference in the application of a vague predicate; some *n*th candy bar is the difference between fat and thin, but we simply do not know which one it will be.[26] Thus, an epistemicist would believe that each candy bar he eats has some minute probability that it would cause him to be fat, whereas someone who believes the tolerance principle thinks of any one candy bar that it has zero chance of making him fat.

Not all of our language and thought is vague—take, for example, the predicate of numerical identity—but insofar as there are vague predicates, addiction is a good candidate. Vagueness is characterized by borderline cases, and there are clearly borderline cases of addiction.[27]

Premise 2: The user chooses to do drugs in small, incremental units.

The Vague Desire picture does not explain the behavior of a recreational drug user who makes the decision that he will smoke a pack a day for the next year. The Vague Desire picture relies essentially on the tolerance principle, and a pack

23 "Leibniz shared the view which has dominated twentieth-century discussions: that the ignorance associated with vagueness is not a matter merely of not knowing where the cut-off lies, but of there being nothing to know" (Sainsbury and Williamson, "Sorites," 741).

24 See Sainsbury and Williamson, "Sorites," 740.

25 See Sainsbury and Williamson, "Sorites," 745.

26 See Sainsbury and Williamson, "Sorites," 752; Sorensen, *Vagueness and Contradiction*; and Williamson, *Vagueness*.

27 "There is no sharp line determining when problem use becomes addiction" (Pickard, "The Puzzle of Addiction," 14). See also Sinnott-Armstrong and Summers, "Defining Addiction," 123.

a day for a year is too large a quantity of nicotine to plausibly apply the principle. In contrast, the Vague Desire picture does explain the recreational user who thinks "just one more" over and over again until he ends up smoking a pack a day for a year. Which kind of drug user seems more typical—the user who one day decides "I'll smoke a pack a day for the next year" or the user who says, "What's the harm of just one more"? Surely the latter. Plausibly, many drug users consider whether to do drugs in small units rather than planning out copious usage over the long term.[28] Premise 2 roughly comports with the common characterization of drug users, both by psychologists and by users themselves, as narrowly focused on the next high.[29] Contrast this to a bodybuilder who in his meal planning decides to have fifty chicken breasts in a month and binds himself to that decision like Ulysses at the mast. I would be surprised to find a drug user like the bodybuilding Ulysses at the mast. In fact, the drug user could even be contrasted against those who successfully resist the temptation for drugs. For those who stay clean, it may very well be the case that they plan to be clean for the next week, the next month, the rest of their life, etc.

One clear exception, however, is when individuals plan out their drug usage with the goal of avoiding addiction. For instance, someone might plan to smoke exactly three cigarettes a day, with the belief that this amount will avoid addiction, but nevertheless become addicted because he made a miscalculation. His addiction would not be explained by the Vague Desire picture. On the other hand, the tracing argument for moral responsibility does not apply to him either because he had the belief that his actions would not lead to addiction, so the knowledge requirement has not been fulfilled.[30]

If the above two premises are true—as seems likely the case for many—then the Vague Desire picture is a good explanation of how these people can become addicted despite their concerns about the tragic consequences of addiction. These users think that the small amounts of drugs they use in any one instance will not make a difference to whether or not they become addicted, but they make these decisions again and again until they end up addicted.

28 Heyman, *Addiction*, 133.

29 Heyman, *Addiction*, 131; and Kirst, "Social Capital and Beyond," 663.

30 The case becomes more complex if one thinks risking is sufficient for the tracing argument to go through. If so, then people who make long-term decisions to use a certain amount of drugs while risking addiction may be outside of the Vague Desire picture and within the ambit of tracing.

4. Moral and Legal Implications of Vague Desires

If the Vague Desire picture is right, it presents an interesting question about moral responsibility. The drug user made decisions about using small amounts of drugs, but using small amounts of drugs does not cause addiction; using large amounts of drugs causes addiction, but the drug user never made a decision to use large amounts of drugs.

The more general puzzle is this: suppose some wrongful harm is caused by a series of actions, but none of the individual actions causes any wrongful harm. The actor made a free and knowing decision to commit each action that composes the series but never made a decision about the series as a whole. That the actions formed the series was in a sense accidental. Can we still hold the actor to be responsible for having performed the series of actions? In this section, I will resolve the problem for the special case of addiction and tease out the principles relevant to resolving the general case.

4.1. Moral Responsibility for Becoming Addicted

It used to be a crime in California to be addicted to the use of narcotics.[31] Suppose that were still the law. Would it be morally permissible to punish a Vague Desirer for his addiction to narcotics? Retributive justice dictates that punishment is only permissible when applied to those who are culpable for some moral failure. The first hurdle that the law against addiction faces is the act requirement. Our moral rules, on various conceptions, prescribe or forbid acts. Likewise, criminal liability requires some voluntary act.[32] Addiction, insofar as it is a mere status, is not an act. Punishing addiction violates the act requirement.

The obvious revision is to legislate into the penal code an *actus reus* that prohibits the act of entering into that status. What we should punish, an alternate universe California legislature says, is any act that causes one to become addicted to the use of narcotics. Would it be just to punish Vague Desirers under this new statute?

Suppose the candidate act for criminal offense is the Vague Desirer's smoking of any one particular cigarette laced with prohibited narcotics. Such an act fails to violate the prohibition against acts that cause addiction because smoking a single cigarette does not cause addiction for anyone. Applying the tolerance principle, just as no one hair falls to make a man bald, no one cigarette makes a man addicted.

31 Robinson v. California, 370 U.S. 660, 660 (1962).

32 Model Penal Code § 2.01 (Am. L. Inst., Proposed Official Draft 1962).

The real question of concern is raised by a different candidate for criminal offense: the conjunction of every act of smoking done by the addicted person. Suppose an addicted person has smoked a thousand narcotic-laced cigarettes thus far. Whereas no individual smoke caused his addiction, the thousand surely did. Unloading a dump truck full of sand creates a heap. Thus, the Vague Desirer, through his own actions, caused his addiction.

Once it is clear that the Vague Desirer committed the crime, the second step is to figure out whether he was morally responsible for it. To establish responsibility, it must also be the case that the individual knew that his actions would lead to the prohibited state of affairs (or knew that there was a substantial likelihood that it would lead to the prohibited state of affairs). Criminal law calls this the *mens rea* requirement.[33] The Vague Desirer knew that smoking a thousand cigarettes would cause him to be addicted to narcotics so, the hypothetical California legislature would claim, he is morally responsible for his becoming addicted and therefore can be punished.

But *ex hypothesi*, the Vague Desirer made decisions only about whether or not to use small units of drugs; there was never a decision-making process regarding whether or not he would smoke a thousand cigarettes. Thus, even though the Vague Desirer knows that his smoking a thousand cigarettes will cause addiction, that knowledge is disconnected from his smoking the thousand cigarettes.[34] This disconnect, I will argue, is why the Vague Desirer does not satisfy the *mens rea* requirement despite his knowledge. Moral responsibility requires not only the knowledge of the consequences of one's actions but that this knowledge be connected to one's acts in the appropriate ways. Consider the following examples.

> *Railway*$_1$: Bam knows there is a train that runs through Princeton every day at 3:00 PM. The train is hard to see, has no breaks and no warnings, so one must be very careful to avoid the railways around 3:00 PM. Bam knows that the street he is driving on, Nassau Street, will soon intersect with the railway. It is 2:59 PM. Not wanting to be late for his hot date, Bam decides to take a chance and continues to drive down Nassau. Bam crashes into the train, killing a passenger.

> *Railway*$_2$: Dayvid believes that the railway line and Nassau street are parallel to one another. However, he also believes that the railway runs north–south and that Nassau street runs east–west. Despite the

33 In actuality, *mens rea* is a bit more complicated. See generally Model Penal Code § 2.02 (Am. L. Inst., Proposed Official Draft 1962).

34 Cf. Shabo, "More Trouble with Tracing," 998 (arguing that the knowledge requirement for tracing is not satisfied when the causal steps are gradual with no single "pivotal moment").

> metaphysical impossibility of all three propositions being true at once, Dayvid holds these three beliefs because he never thought about all three at once—his beliefs are compartmentalized.[35] Dayvid knows there is a train that runs through Princeton every day at 3:00 PM. The train is hard to see, has no breaks and no warnings, so one must be very careful to avoid the railways around 3:00 PM. Dayvid, driving down Nassau Street, sees that it is 2:59 PM. Attending only to his belief that Nassau is parallel to the railway, he concludes that Nassau is free from any train-related danger. Consequently, he continues driving down Nassau. At 3:00 PM, Dayvid collides with the train and kills a passenger.

Despite the fact that both Bam and Dayvid had the belief that Nassau and the railway would intersect, my intuition is that Bam is morally responsible for crashing into the train and Dayvid is not. The mere fact that Dayvid knew that the railway runs north–south and the street runs east–west is irrelevant if that knowledge was separated from the decision-making for his actions. The relevant belief is that Dayvid thought Nassau was parallel to the railway since that was the belief that factored into his decision-making. It is pure accident that Dayvid crashed into the train.

My intuition here comports with the uncontroversial proposition that moral responsibility requires not only the right sorts of mental states but also that the wrongful acts of an individual arise from those mental states in the right ways. Not as uncontroversial is what exactly those right ways are.

Should the right way be analyzed as counterfactual dependence between the wrongful act and the mental state? No. In Railway$_1$, Bam is clearly morally responsible because he recklessly killed the train passengers. Bam had the belief that continuing to drive down Nassau risks train collision. However, had Bam lacked that belief, he still would have driven down Nassau so that he could make his date on time. Bam's wrongful act lacked counterfactual dependence with respect to his belief, but his belief nevertheless grounds his responsibility.

Consider now the following necessary condition for a wrongful act to have arisen from one's belief in the right way:

> *Mental Difference-Maker Condition*: In order for individual *D*'s belief *B* to satisfy the *mens rea* requirement for his wrongful act *A*, the following modal fact must obtain of *B* with respect to *A*. In the closest possible world in which *D* has the desires and self-control of a morally ideal virtuous model as well as belief *B*, *D* abstains from *A*-ing.

35 This example derives from Lewis, "Logic for Equivocators," 436.

Put more colloquially, a belief cannot satisfy the *mens rea* requirement with respect to an offense if the defendant's having morally virtuous desires would not have prevented him from committing the offense. How could we blame and criticize you if you acted just the same as a morally virtuous person would have done in your situation?

What separates Railway_1 from Railway_2 is the violation of this modal proposition. Bam, in Railway_1, thought that meeting his date on time was more important than the risk he imposed on the train passengers. A virtuous person would not have such skewed priorities. If Bam had virtuous desires, he would have thought it more important that he not risk the lives of others and would have consequently changed his course of driving. Dayvid, on the other hand, may already have had the virtuous desire not to harm others. Because Dayvid was acting on his belief that Nassau Street is parallel to the railway, however, those desires would not have made him abstain from driving down Nassau. Therefore, the mental difference-maker condition does not obtain for Dayvid's beliefs, and the *mens rea* requirement cannot be satisfied by those beliefs.

I use Railway_2 to illustrate compartmentalization because it is a canonical example familiar to philosophers, but our story need not be so fancy. Of course our beliefs sometimes fail to make a difference. An obvious case is that of distraction. As I assume is similar for most of us, when I am doing philosophy, I find it very difficult to pay attention to anything else. A few years ago, in the middle of writing an essay on self-defense, I drew a hot glass cup out of my dishwasher and poured in some ice water. The glass immediately shattered. I had known since childhood that putting ice water in a hot glass would break it. And yet I was so focused on doing philosophy that my belief about glass failed to guide my actions. My breaking the glass was an accident because my belief was never accessed or attended to during my decision-making process. My belief that hot glass is disposed to break was not a difference maker.

A Vague Desirer's belief that smoking a thousand cigarettes would cause addiction is also not a mental difference maker. Since he was always making decisions about smoking the next cigarette, not about smoking many cigarettes over the course of a long period of time, his having the virtuous desires of health and concern for his family would not have stopped him from smoking. Recall that this is the principal conclusion of the Vague Desire picture. The conditions of his decision-making were such that, like Dayvid, the belief that is supposed to be the grounds for his *mens rea* never came into play. Therefore, the Vague Desirer's belief cannot form the basis for *mens rea*, and he is not morally responsible for causing his addiction.

For precision, I state explicitly the theoretical role of the mental difference maker condition. There are at least two kinds of phenomena covered by the

condition: (1) knowledge that is never accessed/attended to and (2) knowledge that has no rational bearing on the decisions the actor makes even if the knowledge is accessed. In Railway$_2$, Dayvid is excused because his knowledge that the railway runs north–south and Nassau Street runs east–west was never accessed. If Dayvid did access the knowledge, it would have had rational bearing on his decision to drive down Nassau, since the coming of the train obligated Dayvid to avoid it. For the Vague Desirer, accessing his knowledge that prolonged use of drugs causes addiction is neither here nor there because he is only thinking about whether to take the next hit. As I have argued, the fact that long-term drug use causes addiction has no rational bearing on one's short-term drug-use decisions. Even if one does access that knowledge, it should not influence one's decision to take the next hit. Thus, the mental difference-maker condition provides a unified explanation of these two distinct categories of excuse. The mental difference-maker condition is itself justified by the notion that we cannot blame individuals who act in ways that the morally virtuous would have in that same situation.

We finally return to the original question of this article: are addicted people morally responsible for their wrongful acts? We began with the *prima facie* argument that drug use by addicted people cannot be punished when addiction acted as a compulsion to use drugs. The response of concern tries to trace back moral responsibility to an earlier point in time—addicted people may be acting under compulsion but only because they culpably placed themselves in that situation.

If my arguments are sound, such a response fails. As discussed earlier, a necessary condition of tracing is that the actor was morally responsible for having put himself or herself in the binding situation. Vague Desirers are not responsible for their having become addicted. Strikingly, this is so even if the addicted people knew that repeated drug use causes addiction and that addiction would lead to wrongful behavior—because such knowledge was disconnected from their decision-making. To the extent that addicted people have a compulsion excuse, no tracing negates it.

The question of whether addicted people are responsible for their having become addicted is an important question in its own right. It deepens our understanding of an important fact of many people's lives. And though this article has focused on criminal law and the harms that addicted people may cause others, getting a better understanding of responsibility for addiction may also have similar implications for how we should think about the harms that addicted people may bring on themselves and the proper response to such self-harming by individuals and the government. To give one example, for luck egalitarians who believe that the government should provide aid to correct

only for unchosen disadvantages, it is of the utmost importance whether the disadvantage of being addicted is, in the morally relevant sense, chosen by the addicted people.[36]

4.2. Objections and Replies

In this section, I respond to the following three objections to the Vague Desire framework and its implications for moral responsibility. First, the mental difference-maker condition should be rejected because it cannot account for negligence as a basis for moral responsibility. Second, the tolerance principle is false; people believe that using small amounts of drugs presents a risk (or certainty) of getting addicted. And third, drug users are responsible for becoming addicted because they chose not to quit.

4.2.1. Negligence

The most common counterargument I have received in response to the mental difference-maker condition is that it cannot account for the blameworthiness of negligence.

Here, it is important to distinguish between two kinds of negligence. The first kind of negligence arises when and only when someone inadvertently creates an unreasonable risk of harm. Negligent acts, under this first definition, are a subset of inadvertent acts, and the mental difference-maker condition excuses those who did not advert to the possible risk of harm.[37] This is why Dayvid did not satisfy the mental difference-maker condition when he collided with the train. He knew that the railway runs north–south and Nassau Street runs east–west, but he never attended to this knowledge, so his having virtuous desires would have made no difference to the ultimate outcome of a collision. To the extent that negligent acts are a species of inadvertence—and only to that extent—the mental difference-maker condition entails that negligent acts cannot be the basis for blame.

If the supposed "controversial" implication of the mental difference-maker condition is that it excuses inadvertent behavior, then that is not a controversial implication at all. It is widely agreed that inadvertence excuses.[38] To the extent that the mental difference maker justifies the proposition that inadvertence

36 See Dworkin, *Sovereign Virtue*, 74–78.

37 Moore and Hurd, "Punishing the Awkward, the Stupid, the Weak, and the Selfish," 149–50.

38 "Many people consider it to be a precept just of ordinary common sense that we cannot be held morally responsible for behavior in which we have engaged only inadvertently" (Frankfurt, "Inadvertence and Moral Responsibility," 1). See also King, "The Problem with Negligence," 577.

excuses, it is an advantage of my theory because it shows explanatory power over a widely agreed upon proposition.

Even if one disagrees with the proposition that ordinary intuition is on the side of inadvertence as excuse, one must surely agree that whether negligence suffices for responsibility is heavily disputed. Many explicitly deny that negligent actors are blameworthy.[39] Thus, it would be unreasonable to demand that all ethical theories make room for the blameworthiness of the negligent. Hard cases make bad law.

The second kind of negligence is simply wrongful failure to take a precaution. Negligence under this second definition is not a subset of inadvertence because one can wrongfully fail to take a precaution while fully adverting to the risk that the failure to take the precaution presents. Consider, for example, a ship captain who thinks, "I really should check my boat's seaworthiness before setting sail, but I'm just too lazy and I don't care about my passengers." This alternate kind of negligence is not excused by the mental difference-maker condition, since a captain who has virtuous desires, that of caring for his passengers, would check his ship for seaworthiness based on his belief that doing otherwise presents significant risk of harm to his passengers.

4.2.2. Risk of Addiction

Earlier, I argued on the basis that the tolerance principle is true. If there is some object *a* to which a vague predicate applies and another object *b* that is qualitatively identical to *a* but for a miniscule difference, then the vague predicate will apply also to *b*. I have relied on this principle to argue that smoking an additional cigarette, making only a miniscule difference, cannot cause addiction.

As is famously the case, accepting the tolerance principle leads to a paradox. Namely, if one thinks smoking one additional cigarette does not lead to addiction, then how could smoking many cigarettes, which is merely a combination of smoking one cigarette at a time, lead to addiction?[40] This paradox leads some philosophers to reject the tolerance principle. According to epistemicists, for example, there is some *n*th cigarette that flips the switch between addicted and nonaddicted (though it is epistemically impossible to obtain the value of *n*).[41]

Suppose, then, the tolerance principle is false, and the *n*th cigarette did in fact cause addiction, thereby satisfying the *actus reus* of entering into addiction. Can the drug user be held responsible for smoking the *n*th cigarette, thereby causing his addiction? Not if the drug user is a Vague Desirer. As I earlier

39 E.g., King, "The Problem with Negligence," 577.

40 See Sainsbury and Williamson, "Sorites," 742.

41 Sainsbury and Williamson, "Sorites," 752.

stipulated by definition, a Vague Desirer must believe the tolerance principle. Thus, he did not think of any single cigarette (including the *n*th cigarette) that it would cause addiction. In fact, he was sure of any one cigarette that he smoked that it would not. Someone who wrongfully believes that his action will not cause any untoward consequence lacks the relevant internal reason not to undertake that action and can avail himself of the excuse of ignorance. The drug user believed that smoking the *n*th cigarette would not cause addiction, so he cannot be blamed for his addiction on the basis of his smoking the *n*th cigarette.

Most people believe the tolerance principle—in fact, philosophers sometimes assert it would even be absurd to deny it—but there remains a minority of people who do not believe the principle.[42] Some might believe that addiction begins upon the first use of a drug. Others might be epistemicists, who believe that there is some unknown threshold for becoming addicted and that each use of a drug increases the probability that one has crossed that threshold. Drug users who have such beliefs would not be able to appeal to the Vague Desire excuse.

4.2.3. Choosing Not to Quit

I had earlier discussed the possibility of someone who decides that he will smoke one pack of cigarettes a day for the next year and how unlikely such a possibility was. Although drug users typically do not make long-term plans to use large amounts of drugs, they do consider whether or not they should make a long-term plan to *not* use drugs. Drug users give consideration to the question of whether they should quit using drugs altogether. The key distinction between what the user does and does not consider is illuminated by the following hypothetical decisions. "Should I commit myself to smoke a pack a day for the next year?" is not a question that smokers ask themselves, but "Should I commit myself to quitting smoking for the rest of my life?" is a question that smokers often ask themselves. And many answer the latter sort of question in the negative. They know that if they quit using drugs, they would eliminate the possibility of addiction and the harms that follow from addiction, but they decide not to quit. The drug user had a path by which he could avoid addiction that he chose not to take. One objection to this article's conclusions states that the drug user's choice not to take the safe path is sufficient grounds to hold him to be responsible for his addiction regardless of whether he is a Vague Desirer.

I would like to here consider two variants of this counterargument. The first variant leverages the putative principle that if there is *any* precaution that the

42 Sainsbury and Williamson, "Sorites," 740–41.

agent chose not to take, then he or she is morally responsible for the resulting harm. Quitting the drug was a precaution the user did not take, so he is therefore responsible for the resulting harm.

I disagree with this principle. It cannot be that we are morally responsible for the harms that result from *every* precaution that we choose not to take. Consider those instances in which a driver accidentally hits and kills a pedestrian. In all such instances, the driver had earlier made a decision whether or not to go for a drive. If the driver had decided at the earlier point in time to not get in the car, he or she would have successfully avoided the resultant homicide. Does that mean drivers are morally responsible for *every* pedestrian death, without exception? Surely not. It would be an absurd result to conclude that all instances of vehicular homicide are blameworthy cases.

The second variant of the counterargument gets around my response above by leveraging an alternative principle: if there is a precaution that the agent chose not to take *despite being morally obligated to take that precaution*, then he or she is morally responsible for the resulting harm.[43] The second variant specifies that responsibility attaches only if one had a duty to take the precaution one chose not to take. Quitting drug use is a precaution the drug user did not take despite the moral obligation to do so, so he is responsible for the resulting harm. This variant avoids my earlier response since one presumably has no obligation to forego driving completely, especially when the cost of foregoing driving is high.

On consequentialist grounds, I disagree with the variant's claim that users have a moral obligation to quit drugs as a precaution against addiction. Taking a single hit of a drug would violate the supposed requirement to quit. However, following the tolerance principle, taking a single hit does not cause addiction. Taking a single hit would be a harmless violation of the requirement to quit, demonstrating that one can avoid addiction without following the "requirement." Since violating the rule and not violating the rule have the same consequences when it comes to the harms of addiction, consequentialism entails that there is no moral requirement to quit using drugs as a precaution against addiction. (According to consequentialism, moral properties of actions supervene on the consequences of those actions. Thus, it cannot be that one action is permitted and another action prohibited if the two actions have the same consequences. The relevant consequences, on the counterargument's own terms, is that of harmful addiction.)

43 See generally Moore and Hurd, "Punishing the Awkward, the Stupid, the Weak, and the Selfish," 180–81 (discussing "culpable failures to take precautions" more generally).

Even if the tolerance principle is false, it is hard to see how responsibility could attach so long as the drug user believes the tolerance principle as Vague Desirers do. That is, if the tolerance principle is true, smoking one additional cigarette is a harmless act with respect to the relevant harm, that of addiction. Thus, the Vague Desirer who believes the tolerance principle believes that he or she is engaging in a harmless act. That someone believes he is doing a harmless act is precisely the circumstance under which he ought to be excused from responsibility.

4.3. Moral Responsibility for Causing Harm

The theoretical framework built up in this paper can be generalized to all vague crimes and moral prohibitions. Any crime with a vague *actus reus* can be committed without satisfying the *mens rea* so long as decision-making is done at a gradual level. Take, for instance, homicide, the *actus reus* of which is causing another's death.[44] Causation is a vague polyadic predicate. Given the vagueness of the homicide statute, just as with addiction, one can violate the *actus reus* without ever having the requisite *mens rea*. Suppose Dr. White poisons Nacho Varga one drop at a time. Varga eventually dies. What is the most poison Dr. White can administer to Varga without causing his death? What is the least poison that Dr. White can administer to Varga while still causing his death? The vagueness of causation means that there is no answer to either question.[45] No single drop makes Dr. White's poisoning the cause of Varga's death. This is an instance of *soritical wrongdoing*, wrongdoing in which a series of actions causes some wrongful harm but none of the individual actions caused the wrongful harm.

My analysis thus far illuminates that whether we can hold wrongdoers to be responsible for soritical wrongdoing depends crucially on whether the wrongdoer made a free and knowing decision about the series of actions *qua* series. The key when it comes to soritical wrongdoing is not necessarily the "free and knowing" condition but rather the "decision" condition. Because it is the series that caused the harm rather than any of the individual actions, when a soritical wrongdoer has never made a decision to commit the series of acts, he or she cannot be held to be morally responsible. To use the above example

44 Model Penal Code § 210.1 (Am. L. Inst., Proposed Official Draft 1962).

45 If it is easier to conceptualize, here is a soritical way to think about this. Suppose each drop of poison reduces Varga's lifespan by one second. I take it as obvious that to reduce another's lifespan by one second is not to cause that person's death. It would be absurd if that were sufficient to satisfy the *actus reus* of homicide. Then, applying the tolerance principle, reducing Varga's lifespan by two seconds is not causing his death. Same for three seconds, four seconds, five, etc.

of Varga and Dr. White, suppose further that Dr. White never desired to kill Varga and only ever made decisions about whether to give Varga one more drop of poison. For all of the reasons just explicated about addiction, Dr. White will not be morally responsible for Varga's death. Soritical wrongdoing opens the possibility of someone doing something wrong without ever having decided to do the thing that was wrong.

Contrast this to another, more common case of soritical wrongdoing: environmental pollution. A Manufacturer, as a byproduct of manufacturing a Widget, releases a small amount of chemical *X* into the river that feeds into the local drinking supply. Chemical *X* in trace amounts is harmless but when built up in the body is lethal. One resident in Manufacturer's town was found to have died from a lethal dose of chemical *X*—he had been drinking the water in this town for several years and the build-up of the chemical was too much for his body to take. Here, in contrast to the case of Vague Desirers who make decisions only about using small amounts of drugs, the Manufacturer, as a business, is likely to have made a long-term decision to produce Widgets. Thus, if they knew that long-term production of Widgets would kill local residents, then they are morally responsible for the resident's death. Where there has been soritical wrongdoing, responsibility turns on whether an individual had chosen the entire series of actions rather than choosing only piece by piece. Thus, the manufacturer is responsible while the addicted person is not.

5. Conclusion

I first demonstrated how the problem of Vague Desire would solve the puzzle of addiction and the plausibility with which the Vague Desire picture explains real-world phenomena. Drug users considering the consequences of their next high are right to think, "What's the harm of just one more?" Due to the vagueness of addiction and the tolerance principle, taking another hit of a drug is rationally consistent with their desire not to become addicted.

I argued that Vague Desirers are not morally responsible for their becoming addicted. Although they knew that continued drug use would lead to their addiction, for a belief to satisfy the *mens rea* requirement, it must be a mental difference-maker. Since Vague Desirers never made the decision to do large amounts of drugs, their knowledge that large amounts of drug use would lead to addiction stayed inert.

An analysis of moral responsibility for soritical wrongdoing must principally consider whether the wrongdoer made a free and knowing decision about the series of actions *qua* series. We can attach moral blame only when this condition is satisfied.

For this reason, we cannot trace back moral responsibility for addicted people. Even if Deanna knew that heavy drug use would lead to leaving her child in the car several years in the future, it is highly unlikely that she ever decided to engage in prolonged drug use. Since Deanna is not morally responsible for having become addicted, we cannot trace back responsibility to the start of her drug use to negate the compulsion excuse that addiction provides.[46]

University of Southern California
jnam@law.usc.edu

REFERENCES

Agule, Craig K. "Resisting Tracing's Siren Song." *Journal of Ethics and Social Philosophy* 10, no. 1 (January 2016): 1–24.

Alexander, Larry. "Causing the Conditions of One's Defense: A Theoretical Non-problem." *Criminal Law and Philosophy* 7, no. 3 (October 2013): 623–28.

Andreou, Chrisoula. "Environmental Damage and the Puzzle of the Self-Torturer." *Philosophy and Public Affairs* 34, no. 1 (Winter 2006): 95–108.

Dworkin, Ronald. *Sovereign Virtue: The Theory and Practice of Equality.* Cambridge, MA: Harvard University Press, 2002.

Edgington, Dorothy. "Vagueness by Degrees." In *Vagueness: A Reader*, edited by Rosanna Keefe and Peter Smith, 294–316. Cambridge, MA: MIT Press, 1997.

Elster, Jon, and Ole-Jørgen Skog. "Introduction." In *Getting Hooked: Rationality and Addiction*, edited by Jon Elster and Ole-Jørgen Skog, 1–29. Cambridge: Cambridge University Press, 1999.

Fischer, John Martin, and Neal A. Tognazzini. "The Truth about Tracing." *Noûs* 43, no. 3 (September 2009): 531–56.

Foddy, Bennett, and Julian Savulescu. "A Liberal Account of Addiction." *Philosophy, Psychiatry, and Psychology* 17, no. 1 (March 2010): 1–22.

Frankfurt, Harry G. "Inadvertence and Moral Responsibility." *The Amherst Lectures in Philosophy* 3 (2008): 1–15. https://www.amherstlecture.org/frankfurt2008/.

Heyman, Gene M. *Addiction: A Disorder of Choice*. Cambridge, MA: Harvard University of Press, 2009.

Khoury, Andrew C. "Responsibility, Tracing, and Consequences." *Canadian*

46 Sincere thanks to Scott Altman, Chelsea Chun, Felipe Jiménez, Gregory Keating, Vivian Liu, Michael Moore, Amber Polk, Marcela Prieto, Walter Sinnott-Armstrong, Jack Whiteley, Thomas Verheyen, Yuan Yuan, and participants of the University of Southern California Legal Theory Workshop.

Journal of Philosophy 42, nos. 3–4 (December 2012): 187–207.
King, Matt. "The Problem with Negligence." *Social Theory and Practice* 35, no. 4 (October 2009): 577–95.
———. "Traction without Tracing: A (Partial) Solution for Control-Based Accounts of Moral Responsibility." *European Journal of Philosophy* 22, no. 3 (December 2011): 463–82.
Kirst, Maritt J. "Social Capital and Beyond." *Journal of Drug Issues* 39, no. 3 (July 2009): 653–76.
Lewis, David. "Logic for Equivocators." *Noûs* 16, no. 3 (September 1982): 431–41.
Moore, Michael S. "Addiction." In *The Palgrave Handbook of Applied Ethics and the Criminal Law*, edited by Larry Alexander and Kimberly Kessler Ferzan, 13–44. London: Palgrave Macmillan, 2019.
Moore, Michael S., and Heidi M. Hurd. "Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence." *Criminal Law and Philosophy* 5, no. 2 (June 2011): 147–98.
Pickard, Hanna. "The Puzzle of Addiction." In *The Routledge Handbook of Philosophy and Science of Addiction*, edited by Hanna Pickard and Serge H. Ahmed, 9–22. London: Routledge, 2018.
Sainsbury, Mark, and Timothy Williamson. "Sorites." In *A Companion to the Philosophy of Language*, edited by Crispin Wright and Alexander Miler, 734–64. Chichester: Wiley, 2017.
Schwartz, Stephen P. "Soritic Thinking, Vagueness, and Weakness of Will." *New Ideas in Psychology* 27, no. 1 (April 2009): 18–31.
Shabo, Seth. "More Trouble with Tracing." *Erkenntnis* 80, no. 5 (October 2015): 987–1011.
Sinnott-Armstrong, Walter, and Jesse S. Summers. "Defining Addiction: A Pragmatic Perspective." In *The Routledge Handbook of Philosophy and Science of Addiction*, edited by Hanna Pickard and Serge H. Ahmed, 123–31. London: Routledge, 2018.
Sorensen, Roy. *Vagueness and Contradiction*. Oxford: Oxford University Press, 2001.
Vargas, Manuel R. "The Trouble with Tracing." *Midwest Studies in Philosophy* 29, no. 1 (September 2005): 269–91.
Williamson, Timothy. *Vagueness*. New York: Routledge, 1994.
Wright, Crispin. "On the Coherence of Vague Predicates." *Synthese* 30, nos. 3–4 (September 1975): 325–65.
Yaffe, Gideon. "Compromised Addicts." In *Oxford Studies in Agency and Responsibility*, vol. 5, edited by D. Justin Coates and Neal A. Tognazzini, 191–213. Oxford: Oxford University Press, 2019.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3317

VALUING DIVERSITY

Michael A. Livermore

Domestic and international legal obligations limit the ability of states and individuals to harm endangered species. There are various instrumental, human-centered reasons to preserve many species; these range from the potential discovery of plant-based pharmaceuticals to the recreational value of birdwatching. But many people share the intuition that there are additional, noninstrumental reasons to avoid causing species extinction. One potential foundation for noninstrumental obligations to avoid species extinction is *ecocentrism*, the view that biological aggregates such as species, ecosystems, and landscapes have independent moral value that ought to be respected and given due consideration. The ecocentric view was articulated by Aldo Leopold in his "land ethic" and is reflected in constitutional "rights for nature" and legal rights extended to rivers and mountains in several countries.[1] Several criticisms have been leveled against ecocentrism, including that biological aggregates such as species have only "apparent ends" rather than genuine interests that are worthy of protection.[2]

The claim pursued in this paper is that it is possible to excavate nonecocentric moral intuitions in favor of diversity that, when integrated into a broader welfarist framework, has a range of implications, including providing support for anti-extinction norms. The diversity urged is not biodiversity understood as variety in genomes or phenomes. Rather, the diversity that we have reason to value is diversity of experience. Translating these intuitions to a traditional welfarist framework, I describe a version of welfarism—which I refer to as heteric welfarism—in which diversity takes a place alongside quality (i.e., well-being) and equality of subjective experience.

To state this clearly, the claim articulated here is that worlds that have greater diversity of subjective experience—a greater variety in the forms and qualities of experiences—are better *ceteris paribus* than worlds with less diversity of subjective experience. The reason that worlds with greater levels of diversity

1 Leopold, *A Sand County Almanac*. The country of Ecuador was the first to establish constitutional rights for nature. See Constitution of the Republic of Ecuador, October 20, 2008, art. 71.

2 Sandler, *The Ethics of Species*; and Agar, *Life's Intrinsic Value*.

are better is similar to the reason that worlds with more well-being are better and to the reason that worlds with a fairer distribution of well-being over subjects are better. Diversity is a foundational moral-ethical criterion that can and should be used to evaluate consequences, understood as outcomes or alternative possibilities.

This claim has some resonance with but is very distinct from that offered by G. E. Moore.[3] In a critique of Sidgwick, Moore asked the reader to imagine two worlds, one "exceedingly beautiful," the other as ugly as "you could possibly conceive." Both of Moore's worlds are uninhabited, and therefore there is no one to "enjoy the beauty of the one or hate the foulness of the other." Moore believed that we still have a reason to favor the former world, even if there are no effects on any being's subjective experience. Under the view offered here, there is no morally relevant distinction between Moore's two worlds. It is only through subjective experience that worlds take on moral significance.

The bearers of value in this account are entities that are capable of subjective experience. This set includes, at a minimum, human persons but may also (very plausibly) include nonhuman animals and perhaps even other organisms. It does not include species, ecosystems, or other similar biological aggregates. Nor does it include inanimate objects, landscapes, or even entire planets. What matters under the account articulated here is subjective experience. Because subjective experience matters, worlds that have a greater amount of positive subjective experience (i.e., greater aggregate well-being) are better than worlds with less. The distribution of what matters can also matter, and so the fairness of how well-being is distributed over persons matters, as does (under the heteric view) the diversity of distribution over experiences. More fair distributions (over subjects) are better than less fair ones, and more diverse distributions (over experiences) are better than less diverse ones.

Including diversity in the welfare calculus raises a number of questions, including how to define a meaningful diversity metric and how to balance diversity against the aggregate amount and distribution of well-being. But the value of diversity may also help address some longstanding difficulties for welfarism. Diversity provides a reason to be concerned with the extinction of species, entirely apart from their instrumental value for humans. With respect to animal welfare, diversity can help to justify resistance to efforts by humans to reduce animal suffering by interfering with processes such as predation that are common in the natural world. Within the domain of human morality, diversity can provide welfare-based accounts of the value of protecting

3 Moore, *Principia Ethica*, 83–84.

endangered cultures or ways of life and can offer insights into Parfit's Repugnant Conclusion.[4]

The remainder of the discussion unfolds as follows. Section 1 motivates the view that the diversity of subjective experience is valuable and explains how this experience-based approach differs from prior accounts of the value of diversity, mostly drawn from the field of environmental ethics. Section 2 discusses some of the implications of heteric welfarism. These include how it interacts with other normative commitments such as equal respect and the Pareto principle, as well as its consequences for questions such as species preservation, wild-animal suffering, and the Repugnant Conclusion. Section 3 discusses possible ways that heteric welfarism could be formalized, focusing on one promising method grounded in the notion of Weitzman diversity. The final section offers concluding remarks and notes areas where future work may be warranted.

1. MOTIVATION

Heteric welfarism assigns value to the diversity of subjective experiences and is motivated by the intuition that worlds in which experiences are more varied are in some sense better than worlds in which the range of experiences is more limited. Consider the following scenario.

> *Copied Colonies*: An advanced human society is undertaking a plan of space colonization. Colonization will take place through the construction of massive vessels that will travel over long distances over many years to other solar systems, where they will park in orbit around a star, repurposing each vessel as a permanent colony that will support roughly one million people at a time. There is no anticipation that any of the planets will be inhabitable; the vessel colony is intended to serve as the sole habitation in the solar system. The energy resources of the star will be sufficient to support the colony indefinitely. During the transport stage of the plan, the only means to maintain human life will be in the form of frozen embryos. Once the ships arrive at their destinations, the embryos will be thawed and incubated, then raised by robots until adulthood. This founding generation will then live their lives and raise families, and the colony will continue in perpetuity. It turns out to be much easier to produce vessel colonies with identical founding generations. Under the *Identical Plan*, a single set of one million fertilized zygotes will be selected and then split into many sets of identical (monozygotic) twins that will populate the different vessels. A much larger number

4 Parfit, *Reasons and Persons*.

of vessels can be constructed under this approach than the alternative *Unique Plan*, where every vessel contains a genetically unique population in its founding generation.

In the Copied Colonies scenario, the people in question are biological humans who live in large communities where everyone has substantial opportunities for interpersonal relationships. Although there is no reason to believe that the colonists have any less free will than the contemporary human inhabitants of Earth, we can assume that the colonies will remain extremely similar over time.

The question is whether it would be better for the space-faring society to construct a larger number of vessel colonies with identical populations or a smaller number of vessel colonies with unique populations. Under the Identical Plan, more stars will be colonized earlier, leading to a larger aggregate number of people with lives, each of which is individually worth living. But the cost, if there is one, is that each of the colonies will be extremely similar even over long time horizons. Under the Unique Plan, each vessel colony contains a unique population that can be expected to give rise to a larger range of human experiences. There will be a smaller number of colonies, but they will be substantially different from each other.[5]

There are several reasons to favor the Unique Plan. One is the insurance value of having many different populations. There are presumably many unanticipated challenges that the space colonies will face, and having different starting populations may increase the chances that some of the colonies will survive. This possibility is related to portfolio theory in investing and resilience in ecology.[6] A second reason is that the Identical Plan may undermine the value of the projects of the colonists.[7] Assuming that the colonists are aware of the existence of the other colonies and the circumstances of their creation, the existence of a large number of near copies may deflate the colonists' estimations of the worth of their projects.

Both portfolio-theory and project-value justifications for favoring the Unique Plan can be understood in light of their different outcomes for aggregate well-being (broadly understood). However, our question is whether there is value in diversity separate from effects on aggregate well-being. We

5 To be clear, genetic diversity does not on its own matter. In this scenario, genetic diversity is expected to give rise to a greater diversity of experiences. Furthermore, because the planets are not habitable, the different locations are not expected to give rise to different experiences because each star will be functionally identical, even if located at a different point in the galaxy.

6 Markowitz, "Portfolio Selection"; and Elmqvist et al., "Response Diversity, Ecosystem Change, and Resilience."

7 Scheffler, *Death and the Afterlife.*

can sharpen the Copied Colonies scenario to help clarify matters. First, we can assume that the increase in the total number of colonies made possible by the Identical Plan outweighs whatever portfolio theory–based risk reduction could be achieved through the Unique Plan. From an aggregate perspective, in expectation, the Identical Plan will generate a larger number of colonies. We can assume this to be true even with a substantial amount of risk aversion. Second, we can address the project-value problem. We can assume that the colonists do not care about the existence of other nearly identical colonies—by dint of their inclination, education, and experience, the fact that there is a large number of others engaged in nearly identical pursuits simply does not bother them.

It is possible to argue that if the colonists do not care about the existence of near-identical colonies, they nevertheless should. Accordingly, under certain welfarist accounts, the Identical Plan could have lower aggregate well-being even if the colonists themselves were indifferent. I am interested in a different line of argument in which diversity has value even apart from effects on aggregate well-being. So let us stipulate a social welfare function that is either hedonic or based on people's actual rather than idealized preferences. By analogy, we could consider a society in which people were not averse to inequality and in which there was no diminishing marginal utility of consumption. In such a society, the distribution of wealth would not affect aggregate utility. Nevertheless, there still may be reasons to favor more equal distributions. Egalitarians favor reducing relative inequalities, other things being equal, while prioritarians believe that there is greater value in benefiting people who are worse-off.[8] Neither egalitarian nor prioritarian views necessarily depend on the level of inequality aversion within the population or the shape of the utility curve in consumption.

To reiterate, by stipulation, the aggregate level of well-being is greater in the Identical Plan than in the Unique Plan. For the sake of simplicity, we can hold the distribution of well-being over subjects constant in the two scenarios. The question is whether there is any morally relevant sense in which the Identical Plan is worse than the Unique Plan.[9]

One way in which the Unique Plan is different from the Identical Plan is that the lives lived are more varied. Stated another way, there is a greater diversity of subjective experiences. Ben Bramble argues that from the perspective of an

8 Parfit, "Equality and Priority."

9 I set aside questions related to the person-affecting view, which is that for something to be bad, it must be bad for someone. Under some formulations, the person-affecting view would make comparisons between the Unique Plan and the Identical Plan impossible. I assume either a rejection of the person-affecting view or a suitably revised version in which such comparisons are meaningful. See Adler, "Claims across Outcomes and Population Ethics"; and Masny, "On Parfit's Wide Dual Person-Affecting Principle."

individual life, repeated positive experiences contribute nothing to lifetime well-being.[10] A milder view can be expressed at the aggregate level. If we can expect that many of the lives lived in the Identical Plan will be very similar (i.e., will be nearly repeats of each other), that provides some reason to favor the Unique Plan, other things being equal.

Some may reject the claim that the diversity of subjective experiences matters. Under such a view, worlds in which very similar experiences are had a very large number of times are no better or worse than worlds with a great variety of experiences that are less widespread—so long as aggregate well-being (and perhaps the distribution of well-being over subjects) is the same. We can call such views homic theories (based on the Greek *homos*, i.e., same). The alternative views, in which the distribution of well-being over experiences matter, could be called heteric (based on *heteros*, i.e., different).

Heteric welfarism is analogous to aggregate welfarism in population ethics that favors, *ceteris paribus*, larger populations of experiencing subjects over smaller populations. Both aggregate welfarism and heteric welfarism affirm the value of additional experiences, but in different senses of additional. In the population-size context, aggregate welfarism is sensitive to additional experiences in terms of absolute number. In the diversity context, heteric welfarism is sensitive to additional experiences in terms of variety. Both aggregate welfarism and heteric welfarism can be described as extension-sensitive theories of value, in that worlds in which value is more extensive (i.e., there are more experiences or more kinds of experiences) are favored over worlds in which value is less extensive.

By contrast, average welfarism is indifferent to population size, and homic welfarism is indifferent to the variety of experiences. These views can be described as extension insensitive in the relevant senses. Average welfarism is insensitive to the absolute numerosity of experiences; it is attentive only to the quality of experiences. Homic welfarism is insensitive to how many experiences there are in terms of variety. Mixed views are possible. An aggregate homic welfarist would be extension sensitive with respect to the numerosity of experience but extension insensitive with respect to variety.

Can anything be said concerning the relative merits of extension sensitivity? Broadly speaking, a pure extension-insensitive position is that whatever lives are to be lived, it is best that those lives go well; but it is neither here nor there how many lives are lived or in how much variety. The extension-sensitive view is that it is best that lives go well, but it is also good that there be more—more lives, more variety, or both. In this way, extension-sensitive views are affirming

10 Bramble, "A New Defense of Hedonism about Well-Being."

in a way that extension-insensitive views are not. That said, there is a tradeoff, and extension-sensitive views must be willing in some sense to accept lower quality of life in exchange for more of it.[11]

Heteric theories are compatible with value monism; all that may matter is well-being, and well-being levels may be compared across subjects and experiences. But the different characteristics of experiences may nevertheless matter, in that these characteristics give rise to diversity. In this way, the characteristics of experience are both reducible and not reducible to their effects on well-being. They are reducible in the sense that experiences contribute to well-being levels that can be compared. They are irreducible in that there is additional relevant information concerning the distribution of well-being over experiences that would be lost if all experiences were understood only in terms of well-being levels. This characteristic of heteric theories is analogous to differences between utilitarian and prioritarian/egalitarian theories of welfare. Information concerning the distribution of well-being over subjects is irrelevant for basic utilitarians, whereas it is relevant for prioritarians or egalitarians.

The view that this section has sought to motivate is heteric welfarism. Heteric welfarists find the Unique Plan more attractive than the Identical Plan. For a heteric welfarist, a greater diversity of subjective experiences is a reason to favor one world over another. This reason is moral and not merely an expression of a preference. The level of diversity of subjective experiences is, for the heteric welfarist, a morally relevant feature that ought to be given weight when evaluating alternative worlds.

2. DIVERSITY, EXPERIENCES, AND WELFARISM

Diversity is a potential feature of any set. Just as there is diversity over species in an ecosystem, there is diversity over treats in a candy shop. A diverse ecosystem is teeming with many different types of species; a candy shop that sells licorice and nothing else lacks diversity. Heteric welfarism values diversity of subjective experience, while other types of diversity (whether genetic, physiological, or gastronomic) have only instrumental value.

Philosophers and other thinkers have offered several nonwelfarist conceptions of the value of diversity. Peter Miller argued that the "richness" in natural systems, which includes the concepts of variety and unity, helps explain their value. This claim was taken up and expanded by Gregory Mikkelson. Michael Soulé has described "normative postulates" in the field of conservation biology,

11 This is true if extension plays anything other than a tie-breaking role. The Repugnant Conclusion can be understood as a consequence of this feature of extension-sensitive views. See section 3.6 below.

which include the claims that "diversity of organisms is good" and "biotic diversity has intrinsic value." Ben Bradley argues that rare species have value because they contribute to the biological diversity of the system. And Brendan Cline offers an account of environmental ethics in which the "breathtaking designs" found in nature "have a special value" that "merit[s] our evaluative regard."[12] These accounts all share a common emphasis on diversity in forms of life—which is to say variety in form and makeup of arrangements of organic compounds that engage in metabolism and reproduction and are subject to the evolutionary process.[13]

Heteric welfarism does not value variety in forms of life as such but rather the subjective experiences that track (at least some of) those forms of life. In this way, it is analogous to the view offered by Bramble, which addresses the diversity of experiences in the case of individual well-being.[14] Another view that accords with an experience-oriented notion of diversity has been given by Simon James, arguing that at least some of the value of rare or endangered species derives from their "lifeworld value," which is their unique way of experiencing the world.[15]

L. W. Sumner notes that environmental ethics that extend moral standing to all individual organisms or to biological aggregates such as species face the problem of "an indiscriminate distribution of moral standing."[16] Such profusion runs the risk of trivializing moral concern in part because it is difficult to clarify the boundaries of consideration in a nonarbitrary fashion. The nonwelfarist accounts of diversity discussed above face a similar difficulty unless they can explain why biological diversity—but not the many other kinds of diversity that exist in the world—should be accorded moral significance. Lodging moral standing within the limited number of entities that have subjective experience avoids this problem. Limiting moral consideration to entities with subjective experience is also nonarbitrary because moral value is separate from other realms of value (such as aesthetics) exactly due to its concern for others that

12 Miller, "Value as Richness"; Mikkelson, "Weighing Species"; Soulé, "What Is Conservation Biology?" Bradley, "The Value of Endangered Species"; Cline, "Irreplaceable Design."

13 This is a rough definition of life, a concept that is notoriously slippery. The point is that whatever the criteria for what qualifies as life, subjective experience is not on the list.

14 Bramble, "A New Defense of Hedonism about Well-Being."

15 For James, extinction can be bad because it extinguishes a lifeworld. Accordingly, individual members of endangered species are more valuable than similar members of a numerous species because by existing, they stave off this bad event. As discussed below (section 3.1), this view of extinction is an implication of heteric welfarism. James has not articulated or endorsed any broader claims about the diversity of subjective experience. See James, "Rarity and Endangerment."

16 Sumner, *Welfare, Happiness, and Ethics*, 216.

have, at a minimum, an internal perspective—meaning that there is something that it is like to be them.[17]

That said, diversity of experience is not itself a feature of any individual's experience; it is a holistic property that is observable only from a system-wide perspective. In this way, it is analogous to distribution-sensitive welfarist accounts such as egalitarianism or prioritarianism. An egalitarian or prioritarian attends to the system-wide distribution of well-being over subjects; a heteric welfarist attends to the system-wide distribution of well-being over experiences. Both distribution-sensitive and diversity-sensitive accounts take the individual experiencing subject as the fundamental unit of value—but in a way that admits of consideration of relative characteristics.

3. SOME IMPLICATIONS

Heteric welfarists take diversity of subjective experience as having some value, akin to the quantity and equality of well-being. There are some interesting implications of this view.

3.1. Extinction and Conservation

If diversity of subjective experiences is valuable, it will often be the case that there is a reason to disfavor the extinction of a species. This will hold inasmuch as species reflect different ways of experiencing. The more different subjective experiences are from others, the more valuable they are from the perspective of diversity. Costly efforts to conserve rare species could therefore be justified.

The anti-extinction principle derived from the value of diversity as articulated in this paper differs from accounts offered by environmental ethicists. As noted by Ronald Sandler, species mark a "form of life . . . with [a] distinctive way of going about the world, based on its history, ecology, genetics and phenotypic traits."[18] The concept of a species is primarily scientific—and indeed, there are several different scientific concepts of a species that are put to use in different disciplines and for different purposes.[19] From the perspective of heteric welfarism, the normative significance of the species category is incidental:

17 See Sumner, *Welfare, Happiness, and Ethics*, 217. It is possible that moral consideration might appropriately be further limited, for example, to only those entities for which it is the case that their lives matter to them. Subjective experience, as described here, is a minimum criterion to qualify for welfarist consideration. Additional criteria, such as mattering to oneself or being capable of feeling pain and pleasure (or positive and negative sensations), could be layered on top.

18 Sandler, "On the Massness of Mass Extinction."

19 Zachos, *Species Concepts in Biology*.

inasmuch as species distinctions mark differences in subjective experiences, they are useful for purposes of orienting conservation efforts aimed at increasing diversity. But members of different species may have similar subjective experiences, and members of the same species in a different setting may have different subjective experiences. Conservation efforts to preserve ecosystems, landscapes, or other natural systems may all be justified for purposes of preserving diversity of subjective experience.

The orientation toward promoting diversity of subjective experience is different from an orientation toward preserving biodiversity. There are many practical reasons grounded in human well-being to preserve biodiversity. Inasmuch as biodiversity tracks the diversity of subjective experience, then valuing diversity provides another kind of reason to preserve biodiversity. The lines between species often overlap with the kinds of physical differences that might be thought to bear on subjective experiences, such as differences in perceptual apparatus, differences in diet and reproduction, differences in locomotion, or differences in types of social behavior.

Heteric welfarism is not coextensive with a full-fledged conservation ethic. It does not ground concern for endangered species of flora unless they contribute to the diversity of experience of others.[20] Perhaps more important, heteric welfarism lacks the connection to history that is frequently associated with a conservation ethic: it does not necessarily favor the current distribution or arrangement of species or ecosystems or the state of affairs that existed before human interference or that would arise from human-independent processes.[21] Although heteric welfarism will often in practice overlap with a conservation ethic in seeking to protect endangered species, such alliances are not guaranteed.

Heteric welfarism could even endorse affirmative efforts to generate new or even entirely artificial forms of subjective experience.[22] "De-extinction" is the effort to use biotechnologies to reconstruct species that have been lost to extinction.[23] Although it may be impossible to reconstruct the genomes of lost species, new species could be created that are close phenotypical proxies.[24] Research in synthetic biology involves using technology to construct new forms of life, such as ones based on amino acids that are not found in nature.[25] At its current

20 Assuming that flora do not have subjective experiences.

21 Rolston, *Environmental Ethics*; and Callicott, *In Defense of the Land Ethic*.

22 Bradley notes that the creation of new species would increase biological diversity. Bradley, "The Value of Endangered Species."

23 DeFrancesco, "Church to De-extinct Woolly Mammoths."

24 Lin et al., "Probing the Genomic Limits of De-extinction in the Christmas Island Rat."

25 Dvořák et al., "Bioremediation 3.0"; and Zhang et al., "A Semi-synthetic Organism that Stores and Retrieves Increased Genetic Information."

state of development, research into synthetic life is focused on microorganisms, but it is possible that at some future date, more complex, multicellular forms of life could be constructed using these techniques. Research into artificial intelligence has developed extremely sophisticated computable algorithms that are capable of engaging in high-level pattern detection and strategic decision-making.[26] If the physical processes that underlie consciousness are not limited to biological structures, artificial intelligence systems could theoretically be constructed that could instantiate some form of subjective experience.

The subjective experiences that could be enabled by de-extinction, synthetic life, and artificial intelligence technologies could be substantially different from those that are currently experienced by existing life forms. Other things being equal, the value of diversity offers reasons to favor efforts to generate new kinds of ordered physical systems—whether biological, synthetic, or artificial—capable of having new and varied subjective experiences. Of course, there may be substantial risks associated with de-extinction, synthetic life, and artificial intelligence. The value of diversity does not imply that de-extinction, synthetic life, and artificial intelligence technologies should necessarily be pursued, nor does it mean that the benefits of these technologies are greater than their costs in any particular case. Nevertheless, if they have the potential to increase the variety of ways that it is possible to experience the world, the value of diversity provides a reason to favor them, even in the face of at least some risk.

3.2. Ways of Life

Beyond the context of environmental conservation, the value of diversity is also applicable to human cultures and ways of life. The differences in subjective experiences between a bat and a human, or a bat and an elephant, are vast, but there are also significant differences within the human community as well. If diversity of subjective experiences is valuable, the extinction of the Neanderthals, the Denisovans, and other human species was a grave loss. Based on physiological and inferred neurological differences, the members of these species likely experienced the world in substantially different ways than modern humans. The destruction of unique cultures and ways of life also reduced the diversity of human experiences. Lost languages, religions, worldviews, and life practices constitute a loss of ways of experiencing the world. Social and economic trends associated with modernity have led to greater homogenization of culture—for example, by some estimates, 90 percent of the world's languages currently spoken will be extinct or severely endangered within the

26 Jumper et al., "Highly Accurate Protein Structure Prediction with AlphaFold"; and Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search."

next century.[27] From a value-of-diversity perspective, there are reasons to disfavor this cultural homogenization and to support efforts to preserve distinct and threatened ways of life.

3.3. Wild-Animal Suffering

If we take the subjective experiences of animals to matter, the question arises of whether it would be justified for humans to intervene with processes such as predation that cause great amounts of animal suffering.[28] There are practical limits to how much humans can manage nature in such a way as to reduce animal suffering. Nevertheless, it could be possible to intervene in some limited ways to find alternative means of feeding and entertaining predators and managing prey populations in at least some ecosystems. We could imagine a Managed Nature Program that would limit animal populations through the use of humanely administered contraceptives, with predator animals fed a protein-rich but vegetarian diet and entertained by prey-like toys that do not experience any pain.

The benefits of Managed Nature, from the perspective of alleviating animal suffering, could be substantial. There are some costs that might be relevant as well: the opportunity costs of the resources devoted to running the program, the potential for such ecological interventions to result in negative unintended consequences, etc. We can image a Cost-Benefit-Justified Managed Nature Program such that the cost-effectiveness of the program—comparing the reduction in animal suffering to the costs—is greater than for other existing efforts to improve animal well-being. For example, we can imagine a suite of Humane Farming Requirements that delivers a reduction in animal suffering of one unit at a social cost of $1,000. If the Humane Farming Requirements are justified, then the Cost-Benefit-Justified Managed Nature Program would be justified if it delivered the same marginal benefit at less than $1,000. For the sake of simplicity, we can assume that the Humane Farming Requirements and the Cost-Benefit-Justified Managed Nature Program have the same distributional effects.

There may be some reasons to favor the Humane Farming Requirements while disfavoring the Cost-Benefit-Justified Managed Nature Program. One possibility is that humans have special obligations to domestic animals compared to animals in a natural habitat. Humans have taken on this greater responsibility by breeding and training domestic animals to be helpless without human support. In contrast, animals in a natural habitat arguably exist outside the scope of human activity; although, as noted by Dale Jamieson, predation

27 Krauss, "The World's Languages in Crisis," 7.

28 Cowen, "Policing Nature."

is often influenced by humans in some way.[29] In the case of truly natural conditions, the argument runs, humans are not obligated to take steps to benefit animals in natural habitats.[30] Even more strongly, an appropriate respect for animals might require us to leave them free to lead their own lives, without having human morality imposed on them.[31]

The value of diversity can provide an alternative reason to favor the Humane Farming Requirements while disfavoring the Cost-Benefit-Justified Managed Nature Program. Animals in nature have a great diversity of subjective experiences, which are vastly more varied than the lives of domesticated animals. Were a Managed Nature Program to result in a semi-domestication of wild animals, the richness and variety of experiences in the natural world may be lessened.[32] These diversity effects could be a reason to disfavor a Managed Nature Program, even if it delivered animal suffering benefits at greater (non-diversity-related) cost-effectiveness than Humane Farming Requirements.

3.4. Equal Respect and Consideration

One concern with valuing diversity is that it appears to place greater weight on the well-being of some subjects in proportion to the degree of distance of their subjective experiences from others. This may conflict with treating people with equal respect and consideration. For example, imagine the case of a Lifesaving Medicine in which a choice must be made between saving the lives of the members of a small group with a rare culture versus an equivalent number of lives of members of a large but homogenous culture. If diversity of subjective experience is valuable, that provides a reason to direct the medicine toward the members of the small group rather than the large group.

From the *ex ante* perspective, before life outcomes are known, valuing diversity acts as a kind of insurance for people with rare experiences. If people are risk averse with respect to whether they have rare experiences, analogous to being risk averse with respect to being less well-off, then such insurance could be justified.[33] But unlike the case of well-being, it is unclear why people would want insurance that compensates them when their experiences are rare. Perhaps one might be concerned that people with rare experiences will be politically

29 Jamieson, "The Rights of Animals and the Demands of Nature."

30 Simmons, "Animals, Predators, the Right to Life, and the Duty to Save Lives."

31 Hursthouse, "Virtue Ethics and the Treatment of Animals."

32 The experience of being preyed upon is obviously a negative experience that does not contribute to well-being. As discussed below (section 3.7), there are different ways that heteric welfarism could treat such negative experiences.

33 Harsanyi, "Can the Maximin Principle Serve as a Basis for Morality?"; and Rawls, *A Theory of Justice*.

isolated or face discrimination and so are deserving of special consideration. But this would simply be a case of them being less well-off: people with rare experiences could also be wealthy and politically powerful. Insurance-based arguments do not seem to justify the distributional outcome in the Lifesaving Medicine scenario.

One reason that valuing diversity could arise from the *ex ante* perspective is that people may prefer to live in worlds where there is more diversity, perhaps because those worlds are more interesting, or there is a greater range of lifestyles that are potentially available.[34] But satisfaction of this kind of preference would be an input into well-being rather than an independent consideration. It is also possible that diversity is a natural result of the kinds of good societies that people would favor from the *ex ante* perspective. If good societies are those in which people are free and have a broad scope to determine their life paths, a diversity of experiences may come about from the choices that people make in those societies. In this case, diversity would not be valued for its own sake but would simply be a sign that a society is good in other ways.

The Lifesaving Medicine scenario addresses the question of how people are treated once they are alive. Imagine an alternative Fertility Treatment scenario, in which a scarce fertility treatment must be allocated. As with the case of Lifesaving Medicine, the value of diversity would provide a reason to direct the treatment toward members of the small group with relatively rare experiences rather than toward members of the larger group. The Fertility Treatment scenario involves the question of who will be brought into being, in addition to the question of how people who have already been brought into being (the potential parents) are treated. Although the *ex ante* perspective can be invoked for the questions concerning treatment, it may not give much traction on questions related to who will be brought into being. It places some strain on the *ex ante* perspective to attempt to make recourse to it to address population dynamics, such as whether worlds with larger populations are better than worlds with smaller populations.[35] The value of diversity may be similar in that respect.

Perhaps it is best to say that the independent value of diversity is a different kind of consideration than those that would be important for individuals deliberating from the *ex ante* perspective. It would not be surprising then that the kinds of norms and values that the *ex ante* perspective gives rise to, such as equal respect and consideration, might clash with valuing diversity. Such clashes are not logically necessary, but they cannot be excluded.

34 Dowding and van Hees, "Freedom of Choice."

35 De Lazari-Radek and Singer, *The Point of View of the Universe*, 363–64.

When there are conflicts between diversity and principles such as equal respect and consideration, either could potentially win out. Perhaps the principle of equal respect and consideration is sufficiently weighty in the Lifesaving Medicine scenario that the medicine should be distributed in a way that does not account for the consequences for diversity. But other situations, such as the one described in the Fertility Treatment scenario, could have milder effects on current people but large effects on who is brought into being. Such situations might be a domain in which the value of diversity could win out against conflicting equal treatment principles.

3.5. Pareto Principles

Standard welfarist views satisfy the Pareto principles of Indifference and Ordering. Under the Indifference principle, if everyone in two worlds has the same well-being, the worlds are equally good. Under the Ordering principle, a world is better if at least someone is better-off, and no one is worse-off. Heteric welfare conflicts with these principles. With respect to Indifference, under heteric welfarism, two worlds in which everyone has the same level of well-being are not equally good if, in one of them, the diversity of subjective experiences is greater. For diversity considerations to have any force, heteric welfarism must violate the Indifference principle. With respect to Ordering, a heteric welfarist would not favor a world in which one person is better-off to an arbitrarily small degree (and everyone else if at least as well-off) if, in that world, the diversity of subjective experience is reduced in some sufficient amount; otherwise, diversity is reduced to a mere tiebreaking role.

Notwithstanding these violations of the Pareto principles, there is still a substantial role for well-being in heteric welfarism. Holding diversity constant, worlds are equally good if everyone has the same level of well-being; and worlds are better if someone is better-off, and others are no worse-off. Tracking Matthew Adler, we can refer to these as diversity-modified Pareto principles.[36] Furthermore, under heteric welfarism, worlds are equally good if everyone has the same well-being relevant experiences; this again highlights the well-being orientation of heteric welfarism.

3.6. Repugnant Conclusion

As first articulated by Derek Parfit, if decisions affect not only the well-being of persons brought into being but also the identity and number of persons who are brought into being, then the maximization of aggregate well-being criteria can

36 Adler, "Prioritarianism"

lead to the Repugnant Conclusion.[37] For any given population of persons with high levels of well-being, there is a larger population with lives just above the level that is worth living that will have larger aggregate well-being. This scenario starkly places the extension sensitivity of aggregate welfarism in opposition to the concern with the quality of life that is common to all forms of welfarism.

In the standard formulation of the Repugnant Conclusion scenario, the diversity of subjective experiences is not considered. This may lead to confusing intuitions if one imagines that lives that are barely worth living are also not very different from each other. For example, in contemplating the Repugnant Conclusion, David Heyd has asked, "What is the good in a world swarming with people having lives barely worth living?"[38] The choice of the word "swarm" here may be illuminating. The word literally applies to a "body of bees" or allusively (and, according to the *Oxford English Dictionary*, often contemptuously) a "crowd, throng, [or] multitude" of persons. The word invokes sameness, similarity, and a lack of individuality. At least part of the aversion to the Repugnant Conclusion expressed by Heyd may be that the world envisioned is not only one in which the quality of each individual life is low but also one in which the lives lived are very similar to each other. It is never explicitly stated that lives barely worth living are similar to each other or that lives that are more fulfilling are more varied. But this may be a natural feature of how we imagine the Repugnant Conclusion picture. If this is the case, intuitions concerning the value of the variety of experiences may become confused with intuitions concerning the quality of experiences.

It is possible to offer a slightly reformulated version of the Repugnant Conclusion scenario within heteric welfarism that would avoid this confusion. For aggregate, heteric welfarism, there are two extensive margins: the number of experiences and the variety of experiences. For any world with arbitrary levels on the extensive margins (i.e., a given population with a given diversity) where everyone has a high quality of life, there are alternative worlds that are better because the extensive margins are sufficiently higher (i.e., a larger population; more diversity of experiences; or both) even though quality of life decreases to the point where everyone's lives are barely worth living. This reformulated version of the Repugnant Conclusion separates population size, quality of experiences, and diversity of experiences.

As noted by Stéphane Zuber and coauthors, intuitions concerning very large populations may be unreliable, which makes reasoning about the

37 Parfit, *Reasons and Persons*.

38 Heyd, *Genethics*, 57.

Repugnant Conclusion scenario difficult.[39] Accounting for the diversity of subjective experiences does not necessarily make things easier, because there is an additional parameter of concern (i.e., diversity), and intuitions concerning hyperdiverse worlds may be as unreliable as intuitions concerning worlds with explosively large populations. But the reformulated version does allow consideration of new hypotheticals that were not contemplated in the original formulation of the scenario.

For example, imagine holding the diversity of experiences constant while increasing population size and decreasing quality of life. Begin with Small World, which has a population of some graspable number, say one hundred thousand, living extremely fulfilling and varied lives. Now consider the alternative Big World, which has a larger population of people with lives at a level of well-being just high enough to be worth living but whose experiences are as varied as in Small World. Every person in Big World is as much an individual—no more part of a swarm—as the persons in Small World, but their lives are more difficult, with fewer happy moments and a larger number of setbacks. When the sum of their struggles and satisfactions is taken, there is more total well-being in the larger population. It is perhaps not altogether obvious that favoring Big World over Small World is a truly repugnant conclusion.

Or consider Big Boring World, which has an even larger population, with less variety of experience. If tradeoffs can be made on the extensive margins, Big Boring World may be better than Big World. But there may be declining marginal value, such that as population increases and diversity decreases, ever larger populations would be needed to make up for declines in diversity. There may also be a minimum threshold for diversity, such that no amount of additional population could make up for reductions in diversity. The constraints imposed by diversity may be such that even if there is a Big Boring World that is better than Big World (which is better than Small World), lives are sufficiently varied (and there are sufficiently many of them) in Big Boring World that favoring it is not obviously repugnant.

Finally, consider Small Wild World, with the original population of one hundred thousand, now with lives that are extraordinarily varied but only barely worth living. It is possible that a version of heteric welfarism would favor Small Wild World over Small World—for some, this conclusion may seem repugnant. But diversity may not be entirely population independent. At the limit, there must be at least two people for there to be diversity at all, and small population size would seem to place limits on diversity. If this is the case, then there may not be a Small Wild World that is preferable to Small World. An alternative may be

39 Zuber et al., "What Should We Agree on about the Repugnant Conclusion?"

Larger Wild World, which has a large enough population to support a sufficient diversity of experiences that it offsets the fact that everyone's life is barely worth living. Intuitions may vary concerning whether the conclusion that Larger Wild World should be preferred to Small World is repugnant.

The Repugnant Conclusion scenario is illuminating because it starkly contrasts the extension sensitivity of aggregate welfarism with quality of life. Heteric welfarism complicates the picture by adding another extensive parameter. For some, accounting for diversity may be enough to avoid any repugnant conclusions. For others, because heteric still favors worlds that are very extensive in value but where value is spread very thinly over persons, it too may lead to conclusions of varying degrees of repugnance. But at the very least, accounting for diversity helps separate intuitions concerning the undesirability of sameness (the swarm) from the undesirability of low quality of life.

3.7. Negative Experiences

The value of diversity can provide a reason to favor worlds in which experiences are more varied over worlds in which experiences are more homogenous. One question that naturally arises is how to account for negative experiences.

Under an experience diversity formulation, the only feature of an experience that would be relevant when considering the diversity of experiences is how rare it is, not whether the experience contributed to well-being. This pure experience diversity formulation appears to run afoul of the anti-sadism constraint. A world in which one person lives a good but common life could be disfavored over one in which that person's experiences are all negative, so long as those experiences are sufficiently distinctive.

An alternative formulation would be a value of well-being diversity, in which the diversity that matters arises from positive experiences that are different from other positive experiences. Under this approach, a world in which a person has a good but common life would be preferred to one in which that person suffers in unusual ways. The negative experiences that person has in the second world would not contribute to well-being diversity at all, and there would also be a reduction in aggregate welfare compared to the first world as well.

Even well-being diversity has some undesirable characteristics, including violating the Pareto Ordering principle. A world in which one person lives a good but common life could be disfavored over a world in which that person's experiences are less good (but still positive) and less common. This means that improving one person's lot while leaving everyone else the same would not be preferred if that improvement came with a sufficiently large reduction in diversity. There is some similarity between this result and the leveling down objection to egalitarianism—a concern with the world-level distribution of

well-being (over subjects or experiences) can conflict with more person-centered concerns.

Under the well-being diversity formulation, there are alternative ways to treat suffering. One possibility would be for suffering diversity to be disvalued, such that worlds with similar negative experiences would be favored over worlds with varied negative experiences. The alternatives would be to favor suffering diversity (which would collapse into experience diversity) or ignore suffering diversity and treat all negative experiences the same.

Disvaluing suffering diversity results in a kind of symmetrical treatment of positive and negative experiences. If the world is more full of positive experiences when they are diffused over many types of experiences, one could say that the world is less full of negative experiences when they are concentrated over fewer types of experiences. However, there are reasons to reject this symmetrical treatment of positive and negative experiences. The harm of suffering could be entirely independent of whether the associated negative experiences are common or rare. Indifference over the concentration of suffering over experience would imply that a larger population of suffering subjects is always worse than a smaller suffering population, even if the larger population is made up of the same life lived many times. This is a mild extension of the anti-sadism principle.

Accordingly, the most plausible formulation of the value of diversity focuses on well-being diversity while treating suffering the same, regardless of whether the associated experiences are common or rare.[40]

4. FORMULATIONS

Nothing in heteric welfarism demands (or precludes) any particular degree of analytic formality: it could be realized as a fully articulated quantitative

40 In a critique of Bramble, Timmerman and Pereira criticize what they see as a potential "indefensible ad hoc asymmetry" between his treatment of pleasure and pain if, with respect to an individual human life, "purely repeated pleasures *do not* alter the value of one's life considered as a whole," but "purely repeated pains *do* alter the value of one's life considered as a whole" (Timmerman and Pereira, "Non-repeatable Hedonism Is False," 702). One might extend a similar critique *mutatis mutandis* to the formulation offered above in which the value of well-being is (at least partially) diversity contingent, while the disvalue of suffering is diversity independent. However, there is no obvious reason why asymmetrical treatment must be justified while symmetrical treatment need not be. Well-being and suffering are very different kinds of things, and just as we seek to maximize the former and minimize the latter, we might be concerned with the diversity of well-being and not of suffering. At the very least, the anti-sadism principle provides a reason for asymmetrical treatment.

social welfare function, or it could take the form of qualitative moral evaluations.[41] Nevertheless, it may be useful for even a qualitative analysis to have a clear definition of how diversity could be estimated.[42] In many approaches to measuring diversity, some concept of a *type* is deployed, such that metrics of variation can be estimated within type, and metrics concerning diversity can be used to describe distribution across types. In biological and ecological sciences, species are a common type. Variation is an estimate of the degree of differences found within species for some characteristic, such as height. The relative balance of different species in an ecosystem can be captured with a measure of entropy. Within the field of ecology, different diversity indexes have been proposed that balance the number of types within an ecosystem with the evenness of the distribution across types.[43] Generally, diversity is greater when there are more types with more evenly balanced populations.

Type-based metrics may apply to the diversity of subjective experience. Species are an obvious candidate, in that species boundaries mark out relatively stable differences between organisms. As mentioned above, the genetic or phenotypic distinctions between species do not have foundational moral significance, but they may track differences in how organisms experience the world, which is of moral significance for heteric welfarists. It is possible that type-based metrics of diversity could be made applicable to humans as well, perhaps tracking different sociological categories that deeply influence how people experience the world.

There are, however, several problems with applying type-based measures to the diversity of subjective experience. Most obviously in the case of humans, people do not fall neatly into groups, and efforts to force them into those groups have resulted in many profound harms. Other organisms also create at least some challenges of categorization. More profoundly, type-based metrics treat all types as being similarly different, so that two ant species are counted as two types in the same way that one ant species and one primate species are counted as two types. As a way of understanding the diversity of subjective experience, this characteristic of type-based metrics is a serious limitation.

An alternative approach to measuring diversity, introduced by Martin Weitzman, is based on a measure of distance.[44] The Weizman diversity index can be calculated for any distance measure for which a non-negative, symmetrical distance can be calculated between any two elements of a set. The diversity

41 Regarding the former, compare Adler, *Well-Being and Fair Distribution*.

42 Page, *Diversity and Complexity*.

43 Tuomisto, "A Consistent Terminology for Quantifying Species Diversity?"

44 Weitzman, "On Diversity."

of a set is calculated algorithmically by summing a set of pair-wise comparisons between each member of the set. The calculation begins by selecting a random member of the set (defined as a), then identifying the member with the smallest distance to a (defined as b); the next member is identified with the smallest distance to the set $[a, b]$, meaning the smallest distance to either a or b (defined as c). Then, the next member is identified based on the smallest distance to the set $[a, b, c]$, and so on until all of the members have been identified. At each step, the distance of the member being identified is kept track of; the sum of these distances is the diversity measure.

The major advantage of the Weitzman diversity index compared to other measures is the notion of distance: an ecosystem with two species of ants is treated as less diverse than an ecosystem with an ant species and a primate species. For biodiversity, one natural candidate for distance between two species would be time back to a common ancestor. The common ancestor metric takes advantage of the tree-like phylogenetic structure of biological evolution. Genetic differences can be used to recover evolutionary information based on various biological assumptions or could be used directly to determine distances through some other formalism.

The primary challenge for transposing a concept such as the Weitzman diversity index to the context of subjective experience is in deriving an appropriate notion of distance. At a coarse level, at least some ordinal distance judgments should be relatively uncontroversial: the distance between the subjective experiences of two horses is less than between those of a horse and a cat, which is less than the distance between the experiences of a horse and an octopus. Two undergraduate students at US universities have more similar experiences to each other than one of those students does to a Sumerian farmer in the year 2500 BCE.

Formalized, we might imagine a high but finite dimensional space of subjective experience in which all subjective experiences could be located. In such a space, a measure such as (symmetrized) Kullback–Leibler (KL) divergence could be used to calculate a distance. KL divergence is a flexible metric that can be interpreted as the mutual information between two distributions. If any given experience is understood as a distribution over the possible features of any experience, then KL divergence measures how much any two experiences are alike, where alikeness is understood in the sense that one experience carries a great deal of information about the other.

Even if more formal estimates such as the Weitzman index cannot be calculated in practice, they can help structure the rough judgments that can be made. The upshot is that worlds become more diverse when new subjective experiences are added that are very different from the existing set of experiences. From the perspective of heteric welfare, such worlds are better, *ceteris paribus*.

A final question worth posing about how heteric welfarism is formulated is whether the diversity that matters is diversity of individual experiences or diversity of experiences collected as lives lived. Imagine two different worlds, Left and Right, with experiences drawn from [*X*, *Y*, *Z*] for three individuals [Abe, Betty, Caleb] over three time periods:

Table 1. Left and Right World

	Left world			Right world		
	Abe	Betty	Caleb	Abe	Betty	Caleb
$Time_1$	*X*	*X*	*X*	*X*	*X*	*X*
$Time_2$	*Y*	*Y*	*Z*	*Y*	*Y*	*Y*
$Time_3$	*Z*	*Y*	*Z*	*Z*	*Z*	*Z*

In both worlds, experience *X* is had three times, experience *Y* is had three times, and experience *Z* is had three times. In Left World, those experiences are clustered by person, whereas in Right World, the experiences are spread evenly over persons. Under a life-level understanding of the diversity of subjective experiences, Left World is more diverse because Abe's life is different from Betty's life, and both of their lives are different from Caleb's life. In Right World, each of the inhabitants has the same life. At an experience-level understanding, the worlds have the same level of diversity because the individual experiences in the two worlds are the same.

From the level of the individual, Right World seems superior, if change or richness of experience is valuable.[45] To consider diversity in isolation, we should assume that aggregate well-being is the same in the two worlds (implying that the second experience of *Z* for Caleb is no less valuable than the first), and we can imagine that all of the experiences are of the same quality. Under these conditions, Right World has the problem that each person's life is a copy of the others. The aggregation of experiences over lives can transform the set of experiences [*X*, *X*, *X*, *Y*, *Y*, *Y*, *Z*, *Z*, *Z*] either into a set of copies or into a set of unique lives lived.

Using the Weitzman index, assuming that the distances between *X*, *Y*, and *Z* are the same (imagine an equilateral triangle), the index for both Left World and Right World is the same (the sum of the distances *AB* and *AC*) if an experience-level view of the diversity of experiences is taken. But if the metric is calculated on lives, then the diversity index is zero for Right World (i.e., the distance between identical lives) and is some positive number for Left World.

45 As discussed above, Bramble argues that "purely repeated pleasures . . . add nothing in and of themselves to [a person's] lifetime well-being" ("A New Defense of Hedonism about Well-Being," 98).

The same style of reasoning could be applied at another level of organization. Abe, Betty, and Caleb could be the names of towns rather than people, for example. We can consider whether Right World, which has towns with the same collection of experiences, is less diverse than Left World, which has greater variety at the town level.

Although a diversity measure could be calculated at different levels of aggregation, the individual level has the most obvious appeal. Counting against the experience level would be the difficulty in defining an indivisible experience that could be considered in isolation. Given the rich connection between subjective experiences over the course of an individual's life, this may be impossible. If so, the level of lives, which marks off a clear boundary between entities that experience distinct streams of subjective experience, is the most natural unit.

An individual life can be understood as a distribution over experiences, where experiences are themselves represented as vectors in a high dimensional space of subjective experiences. A Weitzman index would be calculated based on the individual-life distributions, using a metric such as symmetrized KL divergence. The resulting estimate could be incorporated into reasoning that compares the relative goodness of alternative worlds.

5. CONCLUSION

The discussion above provided an overview of heteric welfarism, the view that the diversity of subjective experiences has foundational value within a welfarist framework. This discussion was general and necessarily sacrificed fine-grained detail. The goal was to introduce and describe the view, discuss how it could be formulated, and explore some of its implications. Many of the issues discussed above could bear further scrutiny.

There are many potentially thorny issues that the value of diversity raises for welfarism. In terms of formulations, open questions include whether diversity is considered qualitatively or quantitatively, how best to estimate diversity, and how to trade off diversity against other values. There is likely more to be said about how diversity interacts with other normative criteria, such as the Pareto principle and norms of equal treatment, and with debates within welfarism concerning the Repugnant Conclusion and wild-animal suffering. Different formulations of heteric welfarism likely have consequences for how it fits in with other moral intuitions and criteria.

These questions are worth exploring if there is some minimal plausibility to the view that worlds with a greater diversity of experiences are in some sense better than worlds with less variety of experiences. The Copied Colonies scenario is intended to motivate that plausibility—for those who have the intuition

that the Unique Plan is in some sense better than the Identical Plan, heteric welfarism can justify that appeal. In addition, there is some analogy between heteric welfarism and aggregate welfarism, in that both are extension-sensitive views in which the amount of value (in terms of numerosity or in terms of variety) matters. Extension sensitivity is in this sense affirming in a way that extension insensitivity is not. This may provide another reason to find heteric welfarism at least sufficiently plausible that it is worth further consideration.

A final advantage of heteric welfarism is that it provides a means of understanding the motivation behind certain kinds of nature conservation and a perspective from which to balance concerns with diversity against other important considerations. It vindicates the intuition that diversity matters while remaining grounded in the view that subjective experience is the foundation of moral consideration. In this way, it offers a bridge between welfarism and environmental ethics—fields of moral theory that have for the most part proceeded along separate tracks.

University of Virginia
mlivermore@virginia.edu

REFERENCES

Adler, Matthew D. "Claims across Outcomes and Population Ethics." In *The Oxford Handbook of Population Ethics*, edited by Gustaf Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns, 320–49. Oxford: Oxford University Press, 2022.

———. "Prioritarianism: Room for Desert?" *Utilitas* 30, no. 2 (June 2018): 172–97.

———. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford: Oxford University Press, 2012.

Agar, Nicholas. *Life's Intrinsic Value*. New York: Columbia University Press, 2001.

Bradley, Ben. "The Value of Endangered Species." *Journal of Value Inquiry* 35, no. 1 (March 2001): 43–58.

Bramble, Ben. "A New Defense of Hedonism about Well-Being." *Ergo* 3, no. 4 (2016): 85–112.

Callicott, J. Baird. *In Defense of the Land Ethic*. Albany: State University of New York Press, 1989.

Cline, Brendan. "Irreplaceable Design: On the Non-instrumental Value of Biological Variation." *Ethics and the Environment* 25, no. 2 (Fall 2020): 45–72.

Cowen, Tyler. "Policing Nature." *Environmental Ethics* 25, no. 2 (Summer 2003): 169–82.

DeFrancesco, Laura. "Church to De-extinct Woolly Mammoths." *Nature Biotechnology* 39, no. 10 (2021): 1171.

de Lazari-Radek, Katarzyna, and Peter Singer. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press, 2016.

Dowding, Keith, and Martin van Hees. "Freedom of Choice." In *The Handbook of Rational and Social Choice*, edited by Paul Anand, Prasanta Pattanaik, and Clemens Puppe, 374–92. Oxford: Oxford University Press, 2009.

Dvořák, Pavel, Pablo I. Nikel, Jiří Damborský, and Victor de Lorenzo. "Bioremediation 3.0: Engineering Pollutant-Removing Bacteria in the Times of Systemic Biology." *Biotechnology Advances* 35, no. 7 (November 2017): 845–66.

Elmqvist, Thomas, Carl Folke, Magnus Nyström, Garry Peterson, Jan Bengtsson, Brian Walker, and Jon Norberg. "Response Diversity, Ecosystem Change, and Resilience." *Frontiers in Ecology and the Environment* 1, no. 9 (November 2003): 488–94.

Harsanyi, John C. "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory." *American Political Science Review* 69, no. 2 (June 1975): 594–606.

Heyd, David. *Genethics: Moral Issues in the Creation of People*. Berkeley: University of California Press, 1992.

Hursthouse, Rosalind. "Virtue Ethics and the Treatment of Animals." In *The Oxford Handbook of Animal Ethics*, edited by Tom L. Beauchamp and Raymond G. Frey, 119–43. Oxford: Oxford University Press, 2011.

James, Simon P. "Rarity and Endangerment: Why Do They Matter?" *Environmental Values* 33, no. 3 (June 2024): 296–310.

Jamieson, Dale. "The Rights of Animals and the Demands of Nature." *Environmental Values* 17, no. 2 (May 2008): 181–99.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (2021): 582–89.

Krauss, Michael. "The World's Languages in Crisis." *Language* 68, no. 1 (March

1992): 4–10.
Leopold, Aldo. *A Sand County Almanac.* New York: Oxford University Press, 1949.
Lin, Jianquin, David Duche, Christian Carøe, Oliver Smith, Marta Maria Ciucani, Jonas Niemann, Douglas Richmond, Alex D. Greenwood, Ross MacPhee, Guojie Zhang, Shyam Gopalakrishnan, and M. Thomas P. Gilbert. "Probing the Genomic Limits of De-extinction in the Christmas Island Rat." *Current Biology* 32, no. 7 (April 2022): 1650–56.
Markowitz, Harry. "Portfolio Selection." *Journal of Finance* 7, no. 1 (March 1952): 77–91.
Masny, Michal. "On Parfit's Wide Dual Person-Affecting Principle." *Philosophical Quarterly* 70, no. 278 (January 2020): 114–39.
Mikkelson, Gregory M. "Weighing Species." *Environmental Ethics* 33, no. 2 (Summer 2011): 185–96.
Miller, Peter. "Value as Richness: Toward a Value Theory for the Expanded Naturalism in Environmental Ethics." *Environmental Ethics* 4 no. 2 (Summer 1982): 101–14.
Moore, G. E. *Principia Ethica.* New York: Cambridge University Press, 1903.
Page, Scott E. *Diversity and Complexity*. Princeton: Princeton University Press, 2011.
Parfit, Derek. "Equality and Priority." *Ratio* 10, no. 3 (December 1997): 202–21.
———. *Reasons and Persons.* Oxford: Oxford University Press, 1984.
Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.
Rolston, Holmes, III. *Environmental Ethics: Duties to and Values in the Natural World*. Philadelphia: Temple University Press, 1988.
Sandler, Ronald. *The Ethics of Species: An Introduction*. Cambridge: Cambridge University Press, 2012.
———. "On the Massness of Mass Extinction." *Philosophia* 50, no. 5 (November 2022): 2205–20.
Scheffler, Samuel. *Death and the Afterlife*. Oxford: Oxford University Press, 2016.
Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (2017): 484–89.
Simmons, Aaron. "Animals, Predators, the Right to Life, and the Duty to Save Lives." *Ethics and the Environment* 14, no. 1 (Spring 2009): 15–27.

Soulé, Michael E. "What Is Conservation Biology?" *BioScience* 35, no. 11 (December 1985): 727–34.

Sumner, L. W. *Welfare, Happiness, and Ethics*. Oxford: Oxford University Press, 1996.

Timmerman, Travis, and Felipe Pereira. "Non-repeatable Hedonism Is False." *Ergo* 6, no. 25 (2019): 697–705.

Tuomisto, Hanna. "A Consistent Terminology for Quantifying Species Diversity? Yes, It Does Exist." *Oecologia* 164, no. 4 (December 2010): 853–60.

Weitzman, Martin L. "On Diversity." *Quarterly Journal of Economics* 107, no. 2 (May 1992): 363–405.

Zachos, Frank E. *Species Concepts in Biology: Historical Development, Theoretical Foundations, and Practical Relevance*. Cham: Springer Nature, 2016.

Zhang, Yorke, Jerod L. Ptacin, Emil C. Fischer, Hans R. Aerni, Carolina E. Caffaro, Kristine San Jose, Aaron W. Feldman, Court R. Turner, and Floyd E. Romesberg. "A Semi-synthetic Organism that Stores and Retrieves Increased Genetic Information." *Nature* 551 (2017): 644–47.

Zuber, Stéphane, Nikhil Venkatesh, Torbjörn Tännsjö, Christian Tarsney, H. Orri Stefánsson, Katie Steele, Dean Spears, Jeff Sebo, Marcus Pivato, Toby Ord. Yew-Kwang Ng, Michal Masny, William MacAskill, Nicholas Lawson, Kevin Kuruc, Michelle Hutchinson, Johan E. Gustafsson, Hilary Greaves, Lisa Forsberg, Marc Fleurbaey, Diane Coffey, Susumu Cato, Clinton Castro, Tim Campbell, Mark Budolfson, John Broome, Alexander Berger, Nick Beckstead, and Geir B. Asheim. "What Should We Agree On about the Repugnant Conclusion?" *Utilitas* 33, no. 4 (December 2021): 379–83.

Journal of Ethics and Social Philosophy
VOL. 28, NO. 2 · AUGUST 2024
https://doi.org/10.26556/jesp.v28i2.3426

MORE ON THE HYBRID ACCOUNT OF HARM

Charlotte Franziska Unruh

THE HYBRID ACCOUNT of harm combines a temporal account and a noncomparative account of harm.[1] According to the temporal account, agent *A* suffers a harm if and only if *A* is worse-off at time t_1 than *A* was at an earlier time t_0. According to the noncomparative account, agent *A* suffers a harm if and only if *A* suffers negative well-being. Both temporal and noncomparative accounts face counterexamples:

> *Bad Start*: Celia takes a medication before she gets pregnant. As a result, her child Dylan is born with a painful condition.
>
> *Decline*: Fanny is exceptionally athletic. She takes a drug that lowers her athletic ability significantly. However, her athletic skills remain well above average.

The temporal account implausibly implies that Dylan does not suffer harm. The noncomparative account implausibly implies that Fanny does not suffer harm.[2]

The novelty of the hybrid account lies in combining the temporal and noncomparative accounts:

> *Hybrid Account*: Agent *A* suffers a harm if and only if *A* is worse-off at time t_1 than *A* was at an earlier time t_0 or if *A* suffers negative well-being.

By combining the two accounts, the hybrid account avoids the counterexamples. Since Dylan suffers noncomparative harm and Fanny suffers temporal harm, the hybrid account gives the right result in the cases of Bad Start and Decline.

However, Erik Carlson, Jens Johansson, and Olle Risberg have criticized the hybrid account on two counts. First, they argue that the hybrid account fails to correctly classify temporary benefits. Second, they argue that the hybrid account fails to identify death as a harm. In what follows, I defend the hybrid account against both criticisms.

1 This section closely follows my earlier argument in Unruh, "A Hybrid Account of Harm."

2 These cases are inspired by Thomson's "gene paraplegia" case and Hanser's "Nobel Prize winner" case. See Thomson, "More on the Metaphysics of Harm," 445–46; and Hanser, "The Metaphysics of Harm," 432.

1. TEMPORARY BENEFITS

Carlson, Johansson, and Risberg argue that the hybrid account finds harm where there is none.[3] More specifically, they suggest that the hybrid account wrongly finds harm in cases of temporary benefits:

> *Beneficial Pill*: Cesar's well-being level is zero at time t_0 and will stay at zero, unless Cesar takes a pill. Taking the pill would cause Cesar's well-being level to rise to ten at t_1 and leave Cesar with a well-being level of one from t_2 onwards.[4]

According to the hybrid account—more precisely, according to its temporal component—Cesar suffers a harm at t_2, since he is worse-off at t_2 than he was at t_1. But Carlson, Johansson, and Risberg argue that this is implausible, since "there is nothing negative to say about his taking the pill and what this action brings about."[5] According to Carlson, Johansson, and Risberg, first, Cesar does not suffer harm in this scenario, and second, taking the pill does not harm Cesar.

Pace Carlson, Johansson, and Risberg, I argue that Cesar does suffer a harm, and the hybrid account correctly identifies that harm. That being said, I agree with Carlson, Johansson, and Risberg that taking the pill does not harm Cesar. However, the hybrid account does not imply otherwise.

I suggest that the intuition that there is no harm in the Beneficial Pill scenario depends significantly on the presumed effect of the pill:

> *Harmful Pill*: Cesar's well-being level is zero at t_0 and will rise to ten at t_1 and stay at ten, unless Cesar takes a pill. Taking the pill would cause Cesar's well-being level to drop to one at t_2 and stay at one.

I submit that Cesar suffers harm and that taking the pill harms Cesar. But note that Cesar's well-being levels are exactly the same in the cases of Beneficial Pill and Harmful Pill. What differs is how taking the pill affects Cesar's well-being. So what drives our intuition is not the state that Cesar is in but rather the precise effect that the pill has.

Since our focus is on whether Cesar suffers harm, consider a case that does not involve pills:

> *No Pill*: Cesar's well-being level is zero at t_0. It rises to ten at t_1 before it drops to one at t_2 and stays there.

3 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm."

4 This is a simplified version of the "welfare boost" case in Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 4.

5 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 4.

According to the temporal component of the hybrid account, Cesar suffers a harm at t_2, since he is worse-off at t_2 than he was at t_1.

I submit that the fact that Cesar's well-being drops at t_2 is bad for Cesar. (Carlson, Johansson, and Risberg might agree: Johansson and Risberg put forward an account of harm according to which harm consists in adverse effects on welfare.[6] A drop in well-being constitutes an adverse effect on Cesar's welfare.) Moreover, for classifying the adverse effect as a harm, it does not matter whether it is caused naturally. I conclude that Cesar plausibly suffers harm at t_2 in all three cases.

I now turn to the question of whether taking the pill in Beneficial Pill *harms* Cesar. In laying out the hybrid account, I emphasize that the hybrid account is an account of what it is to *suffer harm*; it is not an account of what it is to harm someone.[7] So the hybrid account does not attempt to answer the question of whether taking the pill *harms* Cesar. Taking the pill harms Cesar only if it stands in the right causal relation to the harm that Cesar suffers.

I suggest that taking the pill in Beneficial Pill does not stand in the right relation to the harm to count as harming, on any plausible account of causing harm. The pill causes a temporary benefit: it causes Cesar's well-being to rise for a short amount of time. However, the pill does not cause Cesar's well-being to drop. The beneficial effect of the pill simply wears off after some time.

To give an analogous example, taking a painkiller does not cause the headache that resurfaces after the effect of the painkiller has worn off, and so while the headache constitutes a harm, taking the painkiller does not harm the agent. I support this suggestion with the following case:

> *Two-Way Pill*: Cesar's well-being level at t_0 is zero. Cesar takes a pill that contains two active ingredients. The first ingredient takes effect at t_1 and raises Cesar's well-being level to ten. The second ingredient takes effect at t_2 and lowers Cesar's well-being level to one. Without the second ingredient added to the pill, Cesar's well-being level would have remained at ten.

My claim is that Beneficial Pill is like taking only the first ingredient in Two-Way Pill. Taking the beneficial ingredient benefits Cesar. The processes that make the effects of the pill wear off in Beneficial Pill are like the second ingredient in Two-Way Pill. They harm Cesar by causing his welfare to drop. However, this harm is normal and expected, and suffering it does not wrong Cesar. This arguably limits the moral significance of the harm that Cesar suffers in Beneficial Pill.

6 Johansson and Risberg, "A Simple Analysis of Harm."

7 Unruh, "A Hybrid Account of Harm," 891.

In contrast, Cesar arguably suffers unexpected and wrongful harm when he is given the pill in Harmful Pill. This explains the difference in moral significance between these cases.

Carlson, Johansson, and Risberg make another point regarding Beneficial Pill.[8] According to my version of the hybrid account, the magnitude and duration of welfare loss can influence the severity of a harm.[9] It seems to follow that the extent of the harm Cesar suffers exceeds the benefit he enjoys, since the loss persists for longer. But this, Carlson, Johansson, and Risberg argue, is implausible.

There are two ways to understand this objection. First, it seems implausible that taking the pill harms Cesar more than it benefits him. But a proponent of the hybrid account can agree with this, since the hybrid account does not imply that taking the pill harms Cesar. Second, it seems implausible that the temporal harm Cesar suffers is greater than the temporal benefit he enjoys. (When considering whether to take the beneficial pill, it would be odd for Cesar to think, "I'll get some benefit from it, but I'll suffer a much greater harm once the effect of the pill wears off, so is it worth it?")

A proponent of the hybrid account might offer the following response. Temporal harm is always relative to some earlier time. Carlson, Johansson, and Risberg consider the extent to which Cesar is worse-off at t_2 *relative to* t_1. And Carlson, Johansson, and Risberg compare this temporal harm to the extent to which Cesar is better-off at t_1 *relative to* t_0. This is why they claim that Cesar suffers more harm than he enjoys benefits. In most contexts, the earlier time relative to which temporal harm is defined is plausibly the time just before the agent enters the harmful state. But I suggest that in some contexts, a different temporal baseline is more appropriate. Beneficial Pill is such a case.

At time t_2, Cesar is worse-off than he was at t_1. But Cesar is better-off at t_2 than he was at t_0. I submit that t_0 is the appropriate comparison for Cesar when he is contemplating whether to take the beneficial pill. Cesar is interested in the *effects* of the beneficial pill. The pill causes Cesar's well-being at t_2 to be higher than it was at t_0, but it does not cause Cesar's well-being at t_2 to be lower than it was at t_1. (Cesar might think, "The pill will cause a temporary boost in well-being at t_1 and then a small permanent boost from t_2 onwards.")

In sum, Cesar suffers a temporal harm at t_2 relative to t_1. But Cesar also enjoys a temporal benefit at t_2 relative to t_0. This temporal benefit, together with the fact that Cesar enjoys noncomparative benefits throughout, can explain why Cesar should take the pill, prudentially speaking.

8 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 4–5.

9 Unruh, "A Hybrid Account of Harm," 900–2.

2. DEATH

Carlson, Johansson, and Risberg offer a further criticism of the hybrid account. They argue that it undergenerates harm by failing to classify death as a harm. An agent is *noncomparatively* badly off when the agent has negative well-being. However, a dead person is not at any well-being level, and so their well-being level cannot be negative. For similar reasons, death is not a harm on the temporal account: since the dead person is not at any well-being level, it cannot be lower than before. Since neither the noncomparative nor the temporal component of the hybrid account classifies death as a harm, the hybrid account implies that death is not a harm. But this, Carlson, Johansson, and Risberg claim, is implausible: clearly, assassins harm their victim by causing the victim's death.[10]

A first reply is to concede that death is not a harm, but this blow is softened by the fact that the hybrid account enables us to view death as the prevention of benefits—the benefits of a continued life.

According to the prominent deprivationist view on the badness of death, death is bad for the person who dies because it deprives them of the rest of their life, thereby making their life worse than it would have been, considered in its entirety. I argue that the hybrid account is compatible with and even lends support to the deprivationist view. At the heart of the deprivationist argument as I understand it is the view that what is bad about death is not what it brings to the person's life but what it takes away or prevents. It is in line with this view, I propose, to view death as *the prevention of a benefit*. I claim that benefits are states that are noncomparatively good (i.e., positive well-being) or temporally good (i.e., better than before).[11] Death prevents a person from obtaining benefits they could otherwise have had.

Saying that death harms a person then would be speaking loosely. But our tendency to speak of harming in cases of benefit preventions should not surprise us, for this tendency is apparent not only in cases of death. Consider, for example, a case in which Ann has sent Bob a birthday gift, but Celia intercepts the parcel and keeps it for herself. In this case, strictly speaking, Celia has prevented Bob from receiving a benefit, and yet it seems tempting to say that Celia has harmed Bob.

I think there are two reasons that explain why we often see benefit preventions as harmful. The first reason is that benefit preventions are often wrongs. In intercepting the parcel, Celia wrongs Bob (and perhaps Ann). More obviously,

10 This summarizes the argument in Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 3.

11 Unruh, "A Hybrid Account of Harm," 898.

an assassin who kills a victim commits a grave moral wrong, violating the victim's right to life. Given the severity and significance of this wrong and given how closely harms and wrongs tend to be linked, it is plausible to understand death as a harm to the victim.

The second reason is more speculative. I suggest that it is often the case that people are morally entitled to the benefits that they are wrongfully prevented from receiving, and this makes it seem suitable to classify benefit preventions as harms. One might think that since Ann's gift was meant for Bob, Bob should have it: it is already his from a moral point of view. When Celia intercepts the parcel, the moral status of that interception is similar to taking away what is already in Bob's possession. Perhaps a similar point can be made about death: people are entitled to their continued life, and taking away these future benefits is taking away present entitlements, which is harmful.

These remarks point to a second, less concessive reply, which draws on Thomson's point that "one's current chances of good or ill matter to whether one is currently well or ill off."[12] Death *is* a harm to an agent who dies, because that agent loses the prospect of a continued life, thereby making the agent's life worse than it was before. A flourishing life that is about to end is worse than a flourishing life that will continue.

A third reply might be to categorize the harm of death as a purely noncomparative harm, following Harman, whose list of noncomparative harms includes "disease, deformity, disability, or death."[13] (Note that this would not commit the proponent of the hybrid account to claiming that there is posthumous harm. The proponent of the hybrid view might claim that *death,* i.e., the loss of one's status as a welfare subject, is temporally limited, unlike the state of *being dead*.)

In sum, I suggest that the hybrid account has more resources to account for the harm of death than might be apparent at first sight. (To clarify, this is not a point about the moral wrongness of killing, which does not lie only in its effects on the victim's well-being. But this is a separate question.)

3. THE PRIORITY OF HARM

Interestingly, the intuitions that Carlson, Johansson, and Risberg appeal to in the case of temporary harms and in the case of death are intuitions not about whether the victim is harmed but rather about whether the agent does harm. Carlson, Johansson, and Risberg argue that it is "highly counterintuitive" that

12 Thomson, "More on the Metaphysics of Harm," 445.

13 Harman, "Harming as Causing Harm," 139.

an assassin does not harm his victim.[14] They also argue that the implication that taking the pill harms Cesar in Beneficial Pill cases is "very unappealing."[15] What seems to be underlying these criticisms is a third, more fundamental point, which Carlson, Johansson, and Risberg make only briefly in their concluding section: their objection to the claim that harm is more fundamental than harming.[16]

As mentioned above, on my view, the hybrid account explains what it is for an agent to suffer a harm. We should keep accounts of harm distinct from accounts of harming, which explain what it is for an event to harm an agent.[17] On my view, the relation between harm and harming is as follows. *A* can harm *B* only if *B* suffers harm and *A* stands in the right relation to this harm to count as harming. *A* cannot harm *B* if *B* suffers no harm:

> For a behavior (such as Ann's throwing the stone) to count as harming, the behavior needs to be related, in an appropriate way, to an outcome that counts as a harm (such as Bob's broken nose).[18]

Carlson, Johansson, and Risberg argue that such a tight connection between harm and harming is not plausible:

> If one wants to understand what it is to kick someone, one would presumably not start by theorising about the notion of a "state of being kicked," on the purported ground that any account of kicking needs to presuppose an account of that state. A better idea is to focus directly on the verbal notion; that is, on what it is to kick someone.[19]

However, *pace* Carlson, Johansson, and Risberg, I do not claim that an analysis of harming must begin with an analysis of harm. Rather, an analysis of harm should be conducted separately from an analysis of harming. Carlson, Johansson, and Risberg themselves offer a good explanation for why this claim is true: if we wanted to understand what it is to suffer a *harm*, we also would not start by offering an analysis of *harming*, on the basis that an account of harm needs to presuppose an account of the event that leads to that state. We would rather proceed by investigating both questions separately.

14 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 3.

15 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 5.

16 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 6.

17 Unruh, "A Hybrid Account of Harm," 891.

18 Unruh, "A Hybrid Account of Harm," 891.

19 Carlson, Johansson, and Risberg, "Unruh's Hybrid Account of Harm," 6–7.

Of course, the questions are related. I offered an explanation of how they are related by pointing out that if *A* suffers no harm, then *B* has not harmed *A*.[20] This seems plausible, for the same could be said in relevantly similar cases. For example, *B* cannot have injured *A* if *A* does not have an injury; *B* cannot have blinded *A* if *A* is not blind; and *B* cannot have kicked *A* if *A* has not been kicked. However, of course, *A* can be injured, blinded, or kicked without *B* having injured, blinded, or kicked *A*. (*C* might have done all those things.)

In conclusion, the hybrid account of harm is not mistaken to find harm in cases of temporal benefits. Moreover, the hybrid account can explain why death is bad for the person who dies. What seems to be underlying Carlson, Johansson, and Risberg's criticisms is the view that a philosophy of harm should begin with the analysis of events that harm agents. If I am correct, however, a philosophy of harm should proceed by investigating *both* which states constitute harms and which events constitute harming in order to provide a complete metaphysics of harm.[21]

University of Southampton
c.unruh@soton.ac.uk

REFERENCES

Carlson, Erik, Jens Johansson, and Olle Risberg. "Unruh's Hybrid Account of Harm." *Theoria* 89, no. 5 (August 2023): 1–7.

Hanser, Matthew. "The Metaphysics of Harm." *Philosophy and Phenomenological Research* 77, no. 2 (September 2008): 421–50.

Harman, Elizabeth. "Harming as Causing Harm." In *Harming Future Persons: Ethics, Genetics and the Nonidentity Problem*, edited by Melinda A. Roberts and David T. Wasserman, 137–54. Dordrecht: Springer Netherlands, 2009.

Johansson, Jens, and Olle Risberg. "A Simple Analysis of Harm." *Ergo* 9, no. 19 (March 2023): 509–36.

Thomson, Judith Jarvis. "More on the Metaphysics of Harm." *Philosophy and Phenomenological Research* 82, no. 2 (March 2011): 436–58.

Unruh, Charlotte Franziska. "A Hybrid Account of Harm." *Australasian Journal of Philosophy* 101, no. 4 (October 2023): 890–903.

20 Unruh, "A Hybrid Account of Harm," 891.

21 I am very grateful to two reviewers and the editors of *JESP* for very helpful comments; and to Tony Zhou for feedback on a previous version of this note.