JOURNAL *of* ETHICS & SOCIAL PHILOSOPHY

VOLUME XXII · NUMBER 3
September 2022

ARTICLES

DISCUSSIONS

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 Gould School of Law and the Dornsife College of Letters, Arts, and Sciences at the University of Southern California.

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 associate 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. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1192

VICE SIGNALING

Olúfẹ́mi O. Táíwò

> I'm going to say this and I mean—down to my subatomic particles—what I say. And I actually don't care what anyone might think about it:
>
> I don't give a *FUCK* about Justine Damond and what happened to her.
>
> I don't give a fuck because most white people didn't give a fuck when police murdered seven-year-old Aiyana Stanley-Jones as she lay on a couch, sleeping. What most white people—and some *black* people—did was blame Aiyana's family. . . .
>
> Most white people rely on this idea that black people, in situations where white people are in pain, are only ever to be soothing and understanding; only ever to be Mammy or Uncle Remus; only ever to extend condolences; only ever to embody loyalty; only ever to offer the empathy and sympathy that most white people purposely and haughtily deny when the situation is reversed—almost as if most white people still see us as their property.
>
> When the situation is reversed, when we require empathy and sympathy, then suddenly we're all of the opposite things that these once-needy white people previously said we were. When the shoe is on the other foot, then they assess us as immoral, violent, criminal, subhuman, unworthy.
>
> —Son of Baldwin, "Let Them Fucking Die"

FORTY-YEAR-OLD yoga instructor Justine Damond had called police to her Minneapolis suburb to report a sexual assault. Officer Mohamed Noor arrived on the scene and, for unclear reasons, opened fire on Damond, killing her—a tragedy. Yet: Son of Baldwin does not give a fuck about Justine Damond.[1] And neither, apparently, should you.

Son of Baldwin is a writer known for his skillfully crafted and widely circulated pieces about social justice issues in the US, and is known for hot takes on various aspects of white supremacy. His writing has been controversial at times: in particular, Professor Johnny Williams at Trinity College was the target of a coordinated right-wing media campaign and placed on administrative leave for a

1 The original Son of Baldwin post was deleted from Medium. Some of its text is available in Starr, "I Understand Why Some Black People Couldn't Care Less About Justine Damond."

tweet that referenced Son of Baldwin's characteristically provocative piece, "Let Them Fucking Die."[2]

By itself, Damond's death is tragic but unsurprising. We are not quite sure how many people the police kill—for years, the FBI's statistics on police homicides were calculated by voluntary disclosure of police chiefs, which seems to dramatically undercount—but it is probably more common than we realize.[3]

What was surprising, on the other hand, was the response to her death. Legal consequences for police shootings are not terribly common in the US: between 2005 and 2017, only eighty officers were even arrested on charges for shootings on the job, less than half of whom were convicted.[4] Just days after the Damond killing, the police chief resigned at the mayor's public request.

Other differences between this case and other high-profile cases help explain why there were consequences of this severity in this case, and also help explain why Son of Baldwin wrote what he wrote. In several high-profile cases involving Black victims of police violence, major media outlets have released photos or reported information predictably damaging to the perceived character of the victims. A particularly egregious example is the release by CBS media of the arrest record of Alton Sterling, who was shot in the back while fleeing a police officer, in an encounter recorded on video and widely circulated.[5]

But in Justine Damond's case, media targeted the Black police officer. Meanwhile, media venerated the white victim, showing video of Damond saving ducklings from a sewer and asserting that Damond is the "most innocent victim" of a police shooting that the attorney representing her family had ever come across.[6] That last one stings: among the high-profile cases of police violence are Aiyana Stanley Jones, a Black child killed while sleeping in her bed, and Tamir Rice, a Black child killed while playing in the park.

I assume that Son of Baldwin's core audience—the "in-group" for our purposes here—is predominantly Black and other people of color angry about racial injustice. Given the preceding, we have a lot worth being resentful about. But for our purposes, the important part of this assumption about the core audience is that it helps us understand what Son of Baldwin is up to in his polemic.

To signal one's bona fides as a member of the in-group, one can contradict, mock, or otherwise flaunt the moral standards of the out-group. This is what I take

2 Flaherty, "Trinity Suspends Targeted Professor"; Son of Baldwin, "Let Them Fucking Die."

3 Sullivan et al., "Four Years in a Row, Police Nationwide Fatally Shoot Nearly 1,000 People."

4 Stinson, "Police Shootings Data," 29.

5 Media Matters Staff, "CBS Report on Police Shooting of Alton Sterling Inappropriately Highlights Victim's Record."

6 Goyette, "Justine Damond"; Perez, "Bride-to-Be Is 'Most Innocent' Police Shooting Victim."

it that Son of Baldwin is doing when he edgily assures us that he does not care that Damond is dead, presumably either imagining the reproach of white liberals and conservatives with his core audience or ravenously waiting for actual reactions from this peripheral audience. It is also, from a different political vantage point and with very different moral and political implications, what the person who tells racist jokes in mixed company is doing, and what the person who refuses to use a person's stated gender pronouns is doing. This helps explain why such statements earn the label "vice signaling": these statements do what they do by virtue of the fact that some disfavored out-group is taken not to like it.

In April 2015, James Bartholomew wrote a column for *The Spectator* that used the term "virtue signaling," alleging that public indications of one's personal strengths of moral character were on the rise.[7] By October of that same year, Bartholomew declared that this term (that he invented, he hastens to remind us) had "taken over the world," citing its use by authors with large Twitter followings and articles in well-read publications like Breitbart, *The Daily Telegraph*, and *The Independent*.[8]

The following year, Justin Tosi and Brandon Warmke wrote an article preferring the term "moral grandstanding" to virtue signaling.[9] Their initial article, and an associated blog post about it, inspired long-form responses from Eric Schliesser, Liam Kofi Bright, and Justin Weinberg.[10] Tosi and Warmke have continued to investigate the phenomenon empirically, joined by psychologists, and have found preliminary evidence in favor of their explanation of the phenomenon.[11] This piece aims to supplement their account of moral grandstanding by offering a related concept of *vice signaling*, which typically is a special case of virtue signaling or moral grandstanding rather than a different kind of contribution to public discourse altogether. Analyzing how vice signaling works, then, will help us along in understanding both moral grandstanding and public moral discourse more generally.

Tosi and Warmke discuss cases where the speaker intends for the audience to take their expressions as evidence of good moral character. However, another possibility exists that similarly exploits the social communicative architecture. A

7 Bartholomew, "The Awful Rise of 'Virtue Signalling.'"

8 Bartholomew, "I Invented 'Virtue Signalling.'"

9 Tosi and Warmke, "Moral Grandstanding." I use the term virtue signaling to draw out the intended parallel with vice signaling, which is key to the central aim of this paper. Tosi and Warmke express skepticism but stop short of denying that moral grandstanding and virtue signaling refer to the same phenomenon. I will generally use the terms interchangeably unless referring to their work specifically.

10 Weinberg, "A Surprising Instance of Performative Philosophy"; Krishnamurthy, "Featured Philosopher."

11 Grubbs et al., "Moral Grandstanding in Public Discourse."

contribution to public moral discourse may also attempt to strut by purposely failing to meet the evaluative standards of its audience—or, paradigmatically for my purposes, a particular section of its actual or notional audience. Typically, this strutting takes the form of flaunting or violating out-group standards, behaving viciously or injuriously by the lights of an out-group. I call this kind of communication *vice signaling*.

In both an article in *Psychology Today* and in their recently published book on the topic, Tosi and Warmke argue against use of the terms "virtue signaling" and "vice signaling."[12] They maintain that "signaling" language is misleading since many signaling behaviors are unintentional, and moral grandstanding involves deliberate attempts to draw attention to one's self and affect how one is thought about by others.[13] They also anticipate the connection I aim to make here, to vice signaling, but argue that debates about "virtue signaling" versus "vice signaling" would lead to "pointless arguments" about whether an action is best considered virtue signaling or vice signaling depending on "whether they are expressing good or bad values."[14] They do not say why the arguments would be pointless, but advise the reader to notice that either would fall into moral grandstanding as they define it: the combination of wanting to impress others with one's moral qualities ("recognition desire") and the attempt to satisfy this desire by way of "saying something in public moral discourse" ("grandstanding expression").[15]

My discussion here avoids these particular pitfalls. Since I take vice signaling to be, typically, a "special case" of virtue signaling, I agree that there is little to be gained from arguing which cases are which, or whether and to what extent the acts are good or bad. Accordingly, I will treat the terms "virtue signaling" and "moral grandstanding" interchangeably throughout this piece. The contrast between virtue signaling/moral grandstanding and vice signaling is instead used constructively, to build a more full picture of the stakes and dynamics of communication in public moral discourse, rather than to haggle about how to characterize individual cases. Moreover, since much of the discussion to come appeals to social effects and dynamics that are likely outside of the conscious view of vice signalers, the fact that "signaling" encompasses both witting and unwitting forms of communication figures into this discussion as a feature, not as a bug.[16]

But my discussion also makes out the difference between virtue and vice signaling in a different way than Tosi and Warmke anticipate. Whether or not the values

12 Thanks to an anonymous reviewer for pushing me to respond directly to this point.

13 Grubbs et al., "Moral Grandstanding and Virtue Signaling."

14 Tosi and Warmke, *Grandstanding*, 37–40.

15 Tosi and Warmke, *Grandstanding*, 15.

16 Thanks to an anonymous reviewer for calling my attention to this point.

one expresses are "good or bad" full stop is not the difference between virtue signaling and vice signaling. People vice signal by behaving in a way that they expect out-group members to find injurious or vicious, and expect to thereby perform virtue and curry favor in the in-group.

This way of explaining vice signaling leaves open the question of whether or not the behavior is vicious or virtuous full stop in favor of an explanation where the act seems vicious to the out-group, and this very fact helps constitute it as virtuous for the in-group. The actual moral evaluation of the act itself—whether it is virtuous or vicious from the standpoint of morality, or a more cosmopolitan and less partisan perspective—plays no clear role in this aspect of social life. This is, arguably, is what is going on in the Son of Baldwin case: the moral fact about whether it makes any sense to curse a woman after her death is rendered secondary at best to the more salient fact that doing so will infuriate some out-group (presumably, white liberals who are insufficiently permissive of Black rage).

Whether we characterize such communicative acts as simple virtue signaling or also as vice signaling will depend on which sections of the evaluative community we take to be salient. In this paper I attempt to describe these cases, and point out the moral risks and opportunities they present.

1. DESCRIBING VICE SIGNALING

On their face, virtue signaling and vice signaling may seem to be opposites, since the labels imply that they are signaling opposite things. But the Son of Baldwin case helps bring out the important point further suggested by the umbrella term "moral grandstanding": not only is vice signaling not the opposite of virtue signaling, but an important set of cases of vice signaling are in fact also cases of virtue signaling. These are the cases where someone flaunts the standards of an out-group in order to demonstrate solidarity, seriousness, or some other virtue to their in-group. This could help flesh out the connections investigated by Marcus Arvan between group polarization and moral discourse, which is often used to virtue and vice signal.[17]

Tosi and Warmke initially defined moral grandstanding as what one does when "one makes a contribution to public moral discourse that . . . attempts to get others to make certain desired judgments about oneself, namely, that one is worthy of respect or admiration because one has some particular moral quality."[18] Here, "public moral discourse" is "communication that is intended to bring some moral

17 Arvan, "The Dark Side of Morality."

18 Tosi and Warmke, "Moral Grandstanding," 199.

matter to public consciousness," in contrast to private moral discourse that is not intended for a wider audience.[19]

Vice signaling works by exploiting public information, much like more well-studied phenomena like assertions or questions. But, unlike assertion, vice signaling does not characteristically target the subject under discussion (in the case of conversation). Rather, the point of vice signaling is to change the social architecture that provides the scaffolding for conversation. To see how vice signaling works, it will help to revisit fundamental aspects of communication.

When someone communicates, they presuppose things. It is hard to see how interesting conversation could get off the ground if we had to rebuild a shared understanding of the world (including language itself!) from the ground up anew every single time. One aspect of a communicator's presuppositions is that at least some information is treated as public: that is, as available to other communicators for use in reasoning and other acts.[20] Such information makes up the content of what Robert Stalnaker calls the *common ground*.[21] The common ground is the set of background information we treat as mutual knowledge for, at least, the duration of the conversation. This set is neither all of the things that I know about the world nor the set of things that you know, but the set of things that I know that you know that I know that you know, *ad infinitum*. This is also the social architecture targeted by acts of virtue signaling and vice signaling.

Having the common ground as a communicative resource makes the kind of information-rich discussion that makes conversation possible, and, where we are clever and lucky enough, interesting. The common ground, as I analyze it, is not simply a list of things publicly taken to be the case. It also provides the set of expectations against which people guess which uses of public information will be accepted or rejected, valorized or shamed. The common ground thus understood is not simply a resource but also an incentive structure, and thus in a structural sense a causal structure.[22] When one acts communicatively, one updates

19 Tosi and Warmke, "Moral Grandstanding," 197.

20 Stalnaker describes the content of the common ground as "mutual knowledge." But in his more careful moments, Stalnaker admits that we often treat things on the model of mutual knowledge even when we do not mutually know them: for example, when we suppose things for the sake of argument, or, along the lines I prefer to investigate, when we use the reasoning of a higher status person or theory because I do not want to take the social risks of challenging the view. I am indebted to Dan Zeman for this point.

21 Stalnaker, "Common Ground."

22 I discuss these aspects of the common ground under the heading of "agenda setting effects" at greater length in Táíwò, "The Empire Has No Clothes." The sense of structural causation used here is discussed in Malinsky, "Intervening on Structure."

the common ground—that is, one changes what information serves as public practical premises for the parties in conversation.

The paradigm communicative acts are those whose essential purpose is communicative: utterances, speech acts, signs (in sign language), gestures. But other kinds of acts also communicate. Remaining seated when one is expected to get up may communicate disdain and protest (say, if someone is singing the national anthem); a slap may communicate insult; and changing one's behavioral response to a claim communicated by another may not only communicate the like belief in the accepter but also respect for the person making the recommendation. Even though these acts are not speech acts, these acts also can communicate in that they can affect what social information is public—that is, the content of the common ground—through inferences that one makes about the significance of these actions and relies upon others making. When we speak of an action's communicative effects, we could reformulate that question as a question about what changes it caused to the common ground.

To investigate and characterize the communicative effects of an action, it will matter what was already in the common ground. There has been much discussion about how the content of the common ground determines or affects uptake of what is said or communicated, especially when the bare intelligibility of the act depends on particular presuppositions, in the way that "the present king of France is bald" might rely on a presupposition that France presently has a king.[23]

But when we communicate we are not just trying to transfer information, or tell others about what the world is already like. We are often also trying to change that world, or prevent unwelcome changes to it. We may seek to inspire, motivate, or agitate for a variety of ends. We may be trying to align preferences or objectives with others, or remind people of these commitments if they have forgotten or (in our estimation) are failing to live up to them. Some communicative goals may center around concepts or ideas, even those that may not be perceptible at the level of granularity needed to evaluate an utterance's truth value. For example, a sentence explaining the results of a particular experiment may also be an attempt to establish the correctness or usefulness of the larger theory the experiment was designed to help establish, and recognition of that larger goal may be an important part of understanding what is socially at stake in communicating that particular sentence.

One aspect of the world that communicative acts can affect is the standing of things in relevant social categories and hierarchies within, among, and between them, whether those things are explanations, goals, or people. To the extent that

23 See, for example, Potts, "Presupposition and Implicature"; Abbott, "Presuppositions and Common Ground"; Stanley, *How Propaganda Works*.

information about these categories and hierarchies is public, they are also objects of public coordination and thus embedded in the content of the common ground in some sense or other. For example, a person's location in a prestige hierarchy may affect how the common ground updates in response to their speech. A full-fledged medical doctor's claim that a patient has cancer may affect her willingness to undergo chemotherapy in the way that an equivalent claim made by the patient's accountant would not; should she get another opinion, she will likely do so from another doctor rather than an accountant. Also, she may make use of differences in prestige to settle which doctor's claim to treat as a practical premise in the event that the doctors' claims conflict.

The aforementioned helps us more precisely distinguish vice signaling as a specific subset of virtue-signaling cases. Generally, virtue-signaling communicative acts are those that attempt to affect the location of the speaker in the social locations embedded in the common ground in desired ways by way of performing well by the lights of some public set of evaluative standards, paradigmatically those endorsed by the group one views as an in-group. Vice-signaling communicative acts are those virtue-signaling acts that aim to increase the speaker's prestige or standing in a specific way: by performing badly by the lights of a public set of evaluative standards ascribed to a disfavored out-group by the in-group.[24] This fits squarely into Tosi and Warmke's characterization of the root social explanation, which is the effect the speaker aims to have on their standing and prestige in the company of their audience.

This also helps us resist the temptation to view vice signaling and virtue signaling as opposites. Since our public information may allow for a multiplicity of groups, the same speech act may virtue signal when evaluated with respect to one group's preferred evaluative standards and vice signal when evaluated with respect to another group's. In the central cases of virtue-signaling-as-vice-signaling cases, like the Son of Baldwin case given in the introduction, it is *precisely because* an act is thought to vice signal with respect to the out-group's standards that it functions as virtue signaling in the in-group.

Moreover, since intergroup conflict is at the heart of this characterization of vice signaling, the distinction between virtue signaling and vice signaling is of clear interest to philosophers concerned about political polarization and other aspects of the social dynamics and consequences of this behavior, as Tosi and Warmke clearly are.[25] The more antagonistic the relationship between the in-group and the

24 Of course, an individual may simply wish to signal hostility at an audience without wanting to thereby affect some in group, or even without there being an in-group to thereby affect. I do not focus on these cases here.

25 The new book devotes a full chapter to discussion of these: Tosi and Warmke, *Grandstanding*,

out-group, the likelier that inflaming the out-group will be sufficient grounds for one's action being received positively by the in-group.

With this picture of how vice signaling works in individual conversational interactions, I will point out two potentially positive functions of the practice and two potentially negative ones in section 2.

2. EVALUATING VICE SIGNALING

Thucydides provides a helpful early discussion of vice signaling and related problems in his discussion of conflict in Ancient Greece:

> The meanings of words had no longer the same relation to things, but were changed by them as they thought proper. Reckless daring was held to be loyal courage; prudent delay was the excuse of a coward; moderation was the disguise of unmanly weakness; to know everything was to do nothing. Frantic energy was the true quality of a man . . . the lover of violence was always trusted, and his opponent suspected. . . . He who plotted from the first to have nothing to do with plots was a breaker-up of parties and a poltroon who was afraid of the enemy. In a word, he who could outstrip another in a bad action was applauded; and so was he who encouraged to evil one who had no idea of it.
>
> The tie of party was stronger than the tie of blood, because a partisan was more ready to dare without asking why. . . . The seal of good faith was not divine law, but fellowship in crime. If an enemy when he was in the ascendant offered fair words, the opposite party received them not in a generous spirit, but by a jealous watchfulness of his actions. Revenge was dearer than self-preservation. . . . The cause of all these evils was the love of power, originating in avarice and ambition, and the party-spirit which is engendered by them when men are fairly embarked in a contest. . . . An attitude of perfidious antagonism everywhere prevailed; for there was no word binding enough nor oath terrible enough to reconcile enemies.[26]

Many of the observations Thucydides makes about vice signaling correspond to phenomena pessimistically predicted by Tosi and Warmke about moral grandstanding (virtue signaling), of which vice signaling is typically a special case. Much of the passage claims that Hellenes attempted to one-up each other on savagery toward enemies. Similarly, Tosi and Warmke predict "ramping up," where the sig-

ch. 4.

26 Thucydides, *The Peloponnesian War*, bk. III, as quoted in Robertson, *Patriotism and Empire*, 93–94.

naling value of strong moral claims results in a "moral arms race" in which each individual attempts to demonstrate their commitment to justice by making a claim more extreme than the last individual.[27]

Tosi and Warmke note that people want to avoid being seen as cautious or cowardly by members of the in-group. Thucydides, similarly, comments that "reckless daring was held to be loyal courage; prudent delay was the excuse of a coward; moderation was the disguise of unmanly weakness."[28] Tosi and Warmke predict that "excessive outrage" will result from moral grandstanding, where some will exploit the mistaken tendency to judge those with the most outrage about an issue to be the most morally reliable and upstanding people, either with respect to that issue or generally. Thucydides: "Frantic energy was the true quality of a man."[29]

One important difference, however, between Thucydides' analysis and the one offered by Tosi and Warmke is the level of generality for their claims. Tosi and Warmke focus their attention primarily on the effects of virtue signaling on discourse, perhaps corresponding to a strong distinction between discourse and acts in general. But on the view of things advanced in section 1, communication is something that acts can do in general. Language or discourse concerns the sort of action where communication is usually the point, but does not nearly exhaust the domain of action where communicative effects are salient. This thought is at home in Neil Levy's recent rebuttal to Tosi and Warmke, in which Levy points out that "public moral discourse" serves many social functions, thus doing more than just providing a forum for rational deliberation on moral matters (the singular role assigned to public moral discourse by Tosi and Warmke).[30] Thucydides' account provides a telling real-world example of Levy's objection, on the safe assumption that the "plots" and "crimes" he refers to were not merely verbal dressings-down or pronouncements in the town square.

That is: we can and should ask quite generally what the behavioral consequences of both virtue and vice signaling will be. We would then follow Thucydides in investigating social life beyond speech acts or discourse. If the previous section is onto something, then virtue signaling and vice signaling adjust incentive structures not simply for essentially communicative acts but for all acts that communicate, at least where the communicative effects are salient for the overall payoff of the act or otherwise taken into account by actors. Denigrating speech acts

27 Tosi and Warmke, "Moral Grandstanding," 205.

28 Robertson, *Patriotism and Empire*, 93–94. Supplemented with lines added from Thucydides, *The Peloponnesian War*.

29 Thucydides, *The Peloponnesian War*, bk. III, as quoted in Robertson, *Patriotism and Empire*, 93–94.

30 Levy, "Virtue Signalling Is Virtuous."

communicate insult, but rolled eyes, slaps to the face, revenge plots, and ignored invitations do as well. Then, our phenomena of interest will include speech acts, but it will also include many other sorts of actions.

The discussion of the pros and cons of vice signaling in this section will presume this level of generality to the insights about moral grandstanding discussed so far. I take it that vice signaling has many of the same potential benefits and upshots that virtue signaling or grandstanding have generally, as Levy's article explains: vice signaling can express genuinely held moral commitments and contribute to public discussion.[31] But it is nevertheless worth mentioning two benefits that are especially salient for the vice-signaling subset of virtue-signaling actions.

3. POTENTIAL BENEFITS OF VICE SIGNALING

3.1. Vice Signaling Can Serve as a Basis for Solidarity

The example of etiquette in Southern Rhodesia both provides an example of non-speech acts that communicate and signal in the relevant sense, as well as demonstrating some potential benefits of vice signaling as a practice.

Nathan Shamuyarira was a high-ranking member of Zimbabwe's African National Union—Patriotic Front (ZANU-PF, the party of Robert Mugabe, the country's first prime minister and longtime president). Before taking this role, he was a key member of its nationalist struggle against colonial domination while the country was still known as "Southern Rhodesia." In his historical and autobiographical book *Crisis in Rhodesia*, Shamuyarira recounts not only that nationalist leaders deliberately flaunted the prevailing norms of etiquette, wearing hats in the presence of white officials, but that their willingness to do so became a marker of political credibility.[32]

It is not hard to see the wisdom of this. To follow the prescription of (then) Southern Rhodesia that "natives" (Black Africans) were not to wear hats in the presence of white people was to govern one's self by the moral expressive norms of an apartheid regime. Thus, it was not simply the case that each Black person had intrinsic reason to ignore the norm, part and parcel of a racist and oppressive social structure as it was. It was also the case that each person had reason to broadcast their willingness to defect from such norms, and thereby build social awareness that people were willing to stand up to apartheid in at least this small sense. That sense could, and did, build into a larger and more influential form of resistance, culminating in the successful Zimbabwean War of Liberation.

31 Levy, "Virtue Signalling Is Virtuous."

32 Shamuyarira, *Crisis in Rhodesia*.

Shoemaker and Vargas call this signaling role "moral torch fishing" in the case of blame, arguing that signaling one's adherence to moral norms and willingness to enforce adherence in others is an important moral function that helps social systems cement stable cooperation over time.[33] Similarly, Neil Levy points out that the strong feelings involved in acts of virtue signaling—and thus, as this paper has argued, of many cases of vice signaling—are *constitutive* of possession of the moral virtues they exemplify.[34] In the case of Southern Rhodesia, this kind of anti-apartheid signaling proved efficacious (or at the very least, a survivable mistake), as it played a part in a successful revolt against colonial rule. A pro-solidarity effect of moral grandstanding is consistent with Tosi and Warmke's follow-up empirical investigations, which suggested a positive relationship between moral grandstanding and the tendency to grow closer to people of similar moral and political beliefs.[35]

3.2. Vice Signaling Can Restructure Social Relationships

Vice signaling can help publicize and cement opposition to the status quo, and thereby help restructure society by means of subsequent organized political action. This was the story in the previous example of the Zimbabwean War of Liberation. But vice-signaling communicative acts can directly challenge social relationships and thus relations of power and domination.

Social structure consists of both formal and informal elements. Formal elements, like laws and institutions, are easy to recognize and to specify pathways for changing. But informal elements like norms of civility and etiquette are also influential aspects of social structure. Philosopher Chenyang Li goes as far as to suggest that these aspects of social structure are partially constitutive of individual behavior, as the "cultural grammar" that decides whether some individual's behavioral "sentences" are well formed—that is, whether they succeed or fail by the lights of the going interpretive and evaluative norms.[36] Deliberate flaunting of the going norms can call them into question and provoke a wide reconsideration of those norms.

Historian Robin D. G. Kelley and sociologist James C. Scott describe the cultural importance of this kind of broadcasting to various marginalized groups of people, including working class African Americans, and South Asian peasant populations.[37] They credit it with preserving collective self-respect, cultural opposi-

33 Shoemaker and Vargas, "Moral Torch Fishing."

34 Levy, "Virtue Signalling Is Virtuous."

35 See study 5 in Grubbs et al., "Moral Grandstanding in Public Discourse," 16.

36 Li, "Li as Cultural Grammar."

37 Kelley and Scott often emphasize the cases of vice signaling that are inscrutable to the socially

tion to injustice, and persistent material challenge to oppressive power relations.[38] Li's "cultural grammar" view helps make sense of the last claim. If norms of civility and social conduct are an aspect of social structure, and vice signalers flaunt this aspect of social structure in a way that can provoke reconsideration of the attendant norms, it follows that vice signalers can provoke a reconstitution of social structure itself.

4. Potential Drawbacks of Vice Signaling

Though vice signaling has similar benefits to virtue signaling and other grandstanding acts, its differences and unique dangers show up when considering two interrelated drawbacks.

4.1. Vice Signaling Changes the Subject

Andrea Long Chu provides a telling example of how vice signaling changes the subject. In "On Liking Women" she comments on political lesbianism, a movement that advocated for a connection between same-gender relationships between women and the fight against the patriarchy. She writes:

> I take to be the true lesson of political lesbianism as a failed project: that nothing good comes of forcing desire to conform to political principle. . . . Perhaps my consciousness needs raising. I muster a shrug. When the airline loses your luggage, you are not making a principled political statement about the tyranny of private property; you just want your goddamn luggage back.[39]

Her point, as I understand it, is that the demands of this wave of the radical feminist movement for signaling one's commitment to women's liberation in one's personal relationships problematically dominated other reasons and motivations that would otherwise guide members' choices in romantic and sexual partnerships.

I agree with Tosi and Warmke that it is perhaps additionally morally problematic for individuals to use public moral discourse toward their own individual ends. But the effects on the group dynamics as a whole are my primary concern. Vice

dominant groups ("hidden transcript"), but this is not a necessary aspect of vice signaling. Moreover, as social media changes the incentive structures of public communication, I would guess that the hiddenness of the opposition of marginalized groups will decline in political significance. See Kelley, "'We Are Not What We Seem'"; and Scott, "Domination and the Arts of Resistance."

38 Kelley, "'We Are Not What We Seem,'" 78.

39 Chu, "On Liking Women."

signaling can fundamentally change what is being pursued by the group, above and beyond its effects on individual conversations.

One way that vice signaling can change the subject operates through the relationship it can establish between the in-group and out-group. Generally, vice signalers are in constant contact with their group's own moral commitments. These, after all, will decide whether their performance in public space or contribution to public moral discourse succeeds or fails at instantiating virtue as the group defines it. Vice signaling, on the other hand, puts the in-group in a relationship of epistemic dependence to the out-group. For the vice signaler to successfully vice signal, it is the *out-group's* thoughts, moral compass, and evaluative norms that serve as the primarily relevant factors for vice signaling, not the in-group's.

One may object that I have overstated the case here, since I have left out discussion of what role the in-group's moral commitments play.[40] But, if the in-group's moral commitments are relevant at all to these acts—and it is not obvious that they are—they likely factor as a constraint on which violations of out-group morality will be tolerated. But this fact, even if true in the short term, is little consolation. Consider the following conjectures. First, that the higher the level of antagonism between in-group and out-group, the lower the extent to which in-group moral commitments will constrain vice-signaling acts, since inflaming the out-group is more valued when they are more hated. Second, that acts of vice signaling are likely to help create more antagonism between groups, as they involve deliberately inflaming the out-group and then celebrating this fact. Both of these, together, imply that the effective constraint of in-group morality on acts of vice signaling *weakens* as more vice-signaling acts occur. There are then two related dangers: that in-group moral commitments are not an initially effective constraint on vice-signaling acts and that, however effective they might be when vice signaling is rare, they will become increasingly irrelevant as vice signaling proliferates.

The Son of Baldwin case provides a tidy illustration of this possibility. Vice signaling sidelines the in-group's conception of virtue, treating "fuck Justine Damond" as a virtuous expression of righteous Black anger, pearl-clutching white moderates be damned. But vice-signaling acts and the culture built around them thereby treat speech acts like "fuck Justine Damond" as an instance of a general virtuous kind of action—as an "expression of righteous anger"—obscuring the moral evaluation of the specific token act that it is, which is an insult to a homicide victim. Son of Baldwin does not even attempt to argue that Justine Damond herself did anything to merit being spoken about like this, or otherwise justify the specific thing being said. Rather, the expressive act justifies itself by reference to the hated racist political

40 I am indebted to an anonymous reviewer for the importance of this point.

context and its out-group defenders, directing social attention away from the content of what was said and to the people that it involves, except insofar as they can be instrumentalized to express the speaker's and audience's well-deserved anger.

The possibility of the initial or gradual irrelevance of in-group moral commitments is especially hard to square with a version of social justice where the marginalized in-group wants freedom and self-determination. If this strategy is supposed to be how the in-group escapes the influence of the out-group, this result could hardly be worse. It requires in-group members to make constant reference to what the out-group thinks and believes, even though they aim to play contrarian. Groups that vice signal too often and for too long risk forgetting who they are culturally, ideologically, and politically as they subordinate themselves to antagonism for its own sake—and, in so doing, subordinate themselves to the very out-group they may have aimed to liberate themselves from.

A second way that vice signaling can change the subject is by directly affecting the basic character of social interactions around the topic groups are squaring off against each other over. On social media, our speech acts have quantified, measurable reactions from the audience: likes, replies, and retweets. C. Thi Nguyen argues that this can have structuring effects on our agency much like the rules and point systems of games, which structure our behavior by making the full range of practical possibilities quantitatively commensurable and thus making some decisions more "valuable" (often measured in points) than other decisions.[41] This produces "value clarity," an artificially simplified decision-making environment, which is pleasurable in and of itself and a key aspect of the fun of many kinds of games.

When social interaction around real-world issues is gamified in this way, social life is distorted. Nguyen and Bekka Williams use the term "moral outrage porn" to describe one way that discourse can shift people's antecedent relationship to their moral values. They define moral outrage porn as "representations of moral outrage engaged with primarily for the sake of the resulting gratification, freed from the usual costs and consequences of engaging with morally outrageous content."[42] The value clarity provided by Twitter as a platform, when combined with a culture permissive of internet vice signaling, might change how people interact with issues online and offline.

4.2. Vice Signaling Can Undermine In-Group Goals

The changes vice signaling makes to social interactions can have serious, long-term consequences on in-groups' political interests.[43] Today, vice signaling changes

41 Nguyen, *Games*, ch. 9.

42 Nguyen and Williams, "Moral Outrage Porn."

43 A small but growing body of empirical evidence suggests that there may be positive feedback

the subject of discussion in public moral discourse. But this same drawback, considered on a different timescale, could have even deeper consequences: a month, year, or decade from now, vice signaling could change the practical orientation of a whole group of people or the course of a political project.

Take, for example, a progression of values and decisions we could make as organized opponents of mass incarceration. When we first start engaging online about the issue, we are clearly focused on destroying the current carceral system. We view social media instrumentally: we aim to intervene in online public moral discourse to win converts to our cause and proliferate better strategies among those who currently agree with our goals. Over time, our behavior changes, given the susceptibility of our organizing culture to the gamifying effects of social media platforms. Rather than tweeting and organizing about mass incarceration to figure out how to close jails and prisons, we begin tweeting to excite fellow abolitionists and inflame defenders of the carceral status quo and even make organizing decisions for the same reasons. The simpler, social media–inflected version of our values replaces our original values and concerns: we measure how well we are doing by likes and retweets, not by the population of incarcerated people or the closures of jails and prisons.

This subtle shift in goals is what Nguyen calls "value capture": a gradual reorganization of one's goals and values, where things that were initially secondary or even tertiary goals climb the preference-ordering ranks and function as primary goals.[44] Our moral beliefs, the communities we were originally fighting for, and the events we are trying to bring about or prevent can all become instrumental servants to the symbolism of social interactions if signaling behavior goes unchecked. In the case just offered, the instrumental relationship of social media to concrete political goals is entirely reversed by the end of the process. The importance of the fates and lives of the people currently and at risk of being incarcerated falls by the wayside in favor of the group's new selfish and masturbatory ends: they figure in insofar as they enable us to declare victory online, to the extent that they are relevant at all.[45]

---

between number of participants in signaling kinds of moral discourse at a given time and subsequent recruitment of people into similar kinds of moral discourse. Johnen, Jungblut, and Ziegele, "The Digital Outcry"; Pfeffer, Zorbach, and Carley, "Understanding Online Firestorms."

44 Nguyen, *Games*, ch. 9.

45 Nguyen and Williams also point out the pleasure in consuming content that fits a person's moral perspective. I focus on the social aspects of moral outrage porn here for the sake of drawing out the political significance of changing the subject, but self-pleasure is yet another sense in which moral outrage porn and virtue signaling could "change the subject" ("Moral Outrage Porn," 23–26).

Pervasive vice signaling presents dangers, then, because of its long-term political effects: namely, that it might alter the incentive structures of patterns of discourse, political strategy, and behavior in general around the pursuit of ends that are less important or less coherent with our initial values than the ones we would pursue without them. Vice signaling risks a perverse trade between the communicative performance of taking sides in a political contest and the actions that could lead to winning the contest.

The previous point explains how vice signaling could harm political goals through its effect on our attention, and how antagonism can distract us from trying to make actual progress on changing the social world in the way our group wants. Another way vice signaling could undermine political goals is in the way it distorts deliberation about our group's political issues: that is, how we think about our political goals when we *are* paying attention to them.

If in-group members cannot express or act on ideas that smack of agreement or sympathy with the out-group, this might distort group deliberation that otherwise might have converged on some true or effective outlook. Similarly, an idea that would be rejected if evaluated on independent grounds might instead be embraced because it seems combative or militant, its effectiveness or principledness aside. These possibilities present strategic problems for social movements because the epistemic distortions affect the group's understanding of aspects of the world and the political context that are key to the group's success in political campaigns. This corresponds to Thucydides' observed response to vice signaling in Hellas: the "meaning of words no longer had the same relation to things, but were changed by them as thought proper."[46]

Sustained patterns of vice signaling can lead to the kind of conflict for conflict's sake that Thucydides describes, which is a likely result of the "ramping up" and "trumping up" that Tosi and Warmke consider in their discussion of moral grandstanding, that Arvan links to group polarization, and that relate to the short-sightedness diagnosed by Nguyen and Williams's discussion of moral outrage porn.[47] Tosi and Warmke's prediction about moral grandstanding applies just as well to vice signaling: it might generate an arms race to decide who is the most antagonistic to the mutually hated out-group (marginalizing the least antagonistic folks). It also functions as a way for to jockey for higher positions within the in-group hierarchy, threatening to supplant solidarity based on a group's positive goals with a perverse solidarity based on mutual hatred of an out-group or out-groups, bearing no necessary relationship to a positive set of moral and political commitments.

46 Thucydides, *The Peloponnesian War*, bk. III, sec. 3.82.

47 Arvan, "The Dark Side of Morality," 99; Nguyen and Williams, "Moral Outrage Porn"; Tosi and Warmke, *Grandstanding*, 51–57.

Thucydides chronicled ramping-up effects in his history: "He who plotted from the first to have nothing to do with plots was a breaker-up of parties and a poltroon who was afraid of the enemy. In a word, he who could outstrip another in a bad action was applauded; and so was he who encouraged to evil one who had no idea of it."[48] The danger is that maintaining solidarity in an atmosphere where vice signaling reigns will require yet more vice-signaling acts, generating a perverse feedback loop of pointlessly antagonistic actions that might erode the very social institutions that would be needed to address the grievances that kicked off the process in the first place.

All the effort put into resolving the in-group and between-group crises and battles could have been spent on positive projects: reviewing and working toward the in-group's positive commitments. The necessary behaviors for these positive projects (conversations, research tasks, organizing childcare and carpools) risk being distorted or crowded out entirely by the incentive structure that vice signaling often exploits, cements, and propagates.

Finally, it follows from the preceding that patterns of vice signaling also risk undermining the in-group morally. What makes some out-groups worth opposing is their coherence around fundamentally unjust group goals and practices. But the injustice of the dominant out-group does not by itself make the in-group worth joining: if prisons should not exist, then fighting to abolish prisons is a just struggle. But the struggle *against the people who support prisons* bears no such inherent relationship to justice, and is compatible with prisons' continued existence. If the in-group does not organize itself and cohere around just goals and practices—perhaps better yet, the pursuit of justice itself—then it risks cultivating a purely cosmetic relationship to justice.

5. CONCLUSION

In the preceding, I have primarily discussed the possible results of sustained patterns of vice signaling. Both my criticisms and hopes for vice signaling are primarily strategic or tactical. The goodness or badness of instances of vice signaling depends importantly on the moral status of the political project to which they contribute or fail to contribute. But even conceding this much, vice signaling seems to represent an especially intense form of the risks that have been associated with moral grandstanding.[49] In particular, the way that vice signaling incentivizes the irrelevance of one's own in-group moral commitments seems to pose a much more

48 Thucydides, *The Peloponnesian War*, bk. III, as quoted in Robertson, *Patriotism and Empire*, 93–94.

49 I am grateful to an anonymous reviewer for encouraging me to rethink this point.

fundamental risk to public morality than other kinds of grandstanding—perhaps it is no coincidence that Thucydides' discussion of vice signaling is a description of social collapse and endemic conflict.

Interdisciplinary research can help identify the short-, medium-, and long-term risks of vice signaling. Tosi and Warmke are onto something by beginning to study grandstanders empirically, but an investigation of the psychology or goals of individual people who vice signal is of limited value. If the analysis offered in this paper is right, then the basic social dynamics that explain vice signaling are group level and intergroup. Future research should ask fewer questions about what grandstanders are after or whether or not they are hypocrites—these criticisms and preoccupations themselves risk participating in the erosion of the public moral discourse they purport to defend in a manner much like vice signaling itself does, to the extent that they change the subject to whether or not individuals have the standing or conviction to properly express emotions like outrage and away from the circumstances being responded to.

Instead, future research should shed light on how patterns of communication between networks of people manifest in group-level psychological differences (e.g., a group's "affective tone") and patterns of social and political behavior, including political organizing and electoral participation.[50] Psychologists, sociologists, economists, and political scientists would all have much to contribute to a project of this kind.

There is, however, an ethical conviction motivating the arguments that I have pursued here. I believe that the battle for justice will only be won by defeating the current system of injustice if its replacement is just, and we will not figure out what that looks like just by opposing enough specific elements of the status quo, whether its political factions or its values. More importantly, we will not *be* what that replacement looks like merely by way of opposition, and we will not build what that replacement looks like through pure opposition.[51]

Georgetown University
olufemi.taiwo@georgetown.edu

50 George, "Personality, Affect, and Behavior in Groups," 107.

51 Thanks to Meena Krishnamurthy, Liam Kofi Bright, Joel Michael Reynolds, Abigail Higgins, and Shelbi Nahwilet Meissner for their support and comments during the writing of this article.

REFERENCES

Abbott, Barbara. "Presuppositions and Common Ground." *Linguistics and Philosophy* 31, no. 5 (October 2008): 523–38.

Arvan, Marcus. "The Dark Side of Morality: Group Polarization and Moral Epistemology." *Philosophical Forum* 50, no. 1 (Spring 2019): 87–115.

Bartholomew, James. "The Awful Rise of 'Virtue Signalling.'" *The Spectator*, July 7, 2018. https://www.spectator.co.uk/article/the-awful-rise-of-virtue-signalling-.

———. "I Invented 'Virtue Signalling.' Now It's Taking over the World." *The Spectator*, October 10, 2015. https://www.spectator.co.uk/2015/10/i-invented-virtue-signalling-now-its-taking-over-the-world.

Chu, Andrea Long. "On Liking Women." *n+1*, Winter 2018. https://nplusonemag.com/issue-30/essays/on-liking-women.

Flaherty, Colleen "Trinity Suspends Targeted Professor." *Inside Higher Ed*, June 27, 2017. https://www.insidehighered.com/news/2017/06/27/trinity-college-connecticut-puts-johnny-eric-williams-leave-over-controversial.

George, Jennifer M. "Personality, Affect, and Behavior in Groups." *Journal of Applied Psychology* 75, no. 2 (April 1990): 107–16.

Goyette, Jared. "Justine Damond: Video Shows Australian Rescuing Ducklings Near Minneapolis Home." *The Guardian*, July 19, 2017. http://www.theguardian.com/us-news/2017/jul/19/justine-damond-video-shows-australian-rescuing-ducklings-near-minneapolis-home.

Grubbs, Joshua B, Brandon Warmke, Justin Tosi, A. Shanti James, and W. Keith Campbell. "Moral Grandstanding in Public Discourse: Status-Seeking Motives as a Potential Explanatory Mechanism in Predicting Conflict." *PLOS ONE* 14, no. 10 (October 2019).

Johnen, Marius, Marc Jungblut, and Marc Ziegele. "The Digital Outcry: What Incites Participation Behavior in an Online Firestorm?" *New Media and Society* 20, no. 9 (November 29, 2017): 3140–60. https://doi.org/10.1177/1461444817741883.

Kelley, Robin D. G. "'We Are Not What We Seem': Rethinking Black Working-Class Opposition in the Jim Crow South." *Journal of American History* 80, no. 1 (June 1993) 75–112.

Krishnamurthy, Meena. "Featured Philosopher: Liam Kofi Bright." *Philosopher*, January 21, 2017. https://politicalphilosopher.net/2017/01/20/featured-philosopher-liam-kofi-bright.

Levy, Neil. "Virtue Signalling Is Virtuous." *Synthese* 198, no. 10 (October 2021): 9545–62.

Li, Chenyang. "Li as Cultural Grammar: On the Relation between Li and Ren in Confucius' *Analects*." *Philosophy East and West* 57, no. 3 (July 2007): 311–29.

Malinsky, Daniel. "Intervening on Structure." *Synthese* 195, no. 5 (May 2018): 2295–2312.

Media Matters for America. "CBS Report on Police Shooting of Alton Sterling Inappropriately Highlights Victim's Record." July 7, 2016. https://www.mediamatters.org/video/2016/07/07/cbs-report-police-shooting-alton-sterling-inappropriately-highlights-victims-record/211411.

Nguyen, C. Thi. *Games: Agency as Art*. Oxford: Oxford University Press, 2018.

Nguyen, C. Thi, and Bekka Williams. "Moral Outrage Porn." *Journal of Ethics and Social Philosophy* 28, no. 2 (August 2020): 147–72.

Perez, Chris. "Bride-to-Be Is 'Most Innocent' Police Shooting Victim." *New York Post*, July 21, 2017. http://nypost.com/2017/07/21/bride-to-be-is-most-innocent-police-shooting-victim-lawyer.

Pfeffer, Juergen, T. Zorbach, and Kathleen M. Carley. "Understanding Online Firestorms: Negative Word-of-Mouth Dynamics in Social Media Networks." *Journal of Marketing Communications* 20, nos. 1–2 (March 4, 2014): 117–28. https://doi.org/10.1080/13527266.2013.797778.

Potts, Christopher. "Presupposition and Implicature." *The Handbook of Contemporary Semantic Theory*, 2nd ed. Oxford: Wiley-Blackwell, 2013.

Robertson, John Mackinnon. *Patriotism and Empire*. London: Grant Richards, 1899. https://archive.org/details/cu31924021032366.

Scott, James C. *Domination and the Arts of Resistance: Hidden Transcripts*. New Haven, CT: Yale University Press, 1990.

Shamuyarira, Nathan M. *Crisis in Rhodesia*. New York: Transatlantic Arts, 1966.

Shoemaker, David, and Manuel Vargas. "Moral Torch Fishing: A Signaling Theory of Blame." *Noûs* 55, no. 3 (September 2021): 581–602.

Son of Baldwin. "Let Them Fucking Die." Medium, July 23, 2017. https://medium.com/@SonofBaldwin/let-them-fucking-die-c316eee34212.

Stalnaker, Robert. "Common Ground." *Linguistics and Philosophy* 25, no. 5 (December 2002): 701–21.

Stanley, Jason. *How Propaganda Works*. Princeton: Princeton University Press, 2015.

Starr, Terrell Jermaine. "I Understand Why Some Black People Couldn't Care Less about Justine Damond." *The Root*. https://www.theroot.com/i-understand-why-some-black-people-couldn-t-care-less-a-1797189837.

Stinson, Philip M. "Police Shootings Data: What We Know and What We Don't Know." Urban Elected Prosecutors Summit, Atlanta, GA. April 20, 2017.

Sullivan, John, Liz Weber, Julie Tate, and Jennifer Jenkins. "Four Years in a Row, Police Nationwide Fatally Shoot Nearly 1,000 People." *Washington Post*, February 7, 2019. https://www.washingtonpost.com/investigations/four-years-in-a-row-

police-nationwide-fatally-shoot-nearly-1000-people/2019/02/07/0cb3b098-020f-11e9-9122-82e98f91ee6f_story.html.

Táíwò, Olúfémi O. "The Empire Has No Clothes." *Disputatio* 10, no. 51 (December 2018): 305–30.

Thucydides. *The Peloponnesian War*. Translated by Benjamin Jowett. New York, E. P. Dutton. 1910. http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0200.

Tosi, Justin, and Brandon Warmke. *Grandstanding: The Use and Abuse of Moral Talk*. New York: Oxford University Press, 2020.

———. "Moral Grandstanding." *Philosophy and Public Affairs* 44, no. 3 (Summer 2016): 197–217.

Warmke, Brandon, and Justin Tosi. "Moral Grandstanding and Virtue Signaling: The Same Thing?" *Psychology Today*, August 11, 2020. https://www.psychologytoday.com/blog/moral-talk/202008/moral-grandstanding-and-virtue-signaling-the-same-thing.

Weinberg, Justin. "A Surprising Instance of Performative Philosophy." *Daily Nous*, January 19, 2017. http://dailynous.com/surprising-instance-performative-philosophy/.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1730

POSTHUMOUS REPUGNANCY

Benjamin Kultgen

WHAT DOES a life not worth living look like? A life spent in a state of constant and overwhelming physical suffering would not be worth living. A life in which every conscious experience was that of intense emotional anguish would not be a life worth living. But what about a life that was exceptionally good, day in and day out, right up until the moment of death? Could that life wind up being not worth living, not because of any terrible tragedy, but merely because of a great many minor harms? If it is possible to be harmed after death, then yes, that life could wind up being one not worth living. The possibility of posthumous harm entails that one could have an exceptionally good life (by any standard) while one was alive but incur so many small posthumous harms that one actually had a life not worth living. But we should not accept that. Instead, the possibility of posthumous harm should be rejected.[1]

My argument centers on a kind of repugnancy case involving posthumous harm.[2] Supposing the existence of posthumous harm, a person whose well-being was *extremely high* while she was alive could incur small posthumous harms over a long enough period such that it is true of that person that she had a life not worth living.

The overall argument will be that the possibility of Posthumous Repugnan-

1 I am assuming a few other claims about well-being and harm that I take to be uncontroversial—namely: (1) *S* is harmed by *x* only if *x* negatively affects *S*'s well-being in some way, whatever way that is; (2) harm is additive; and (3) if *S*'s well-being is net negative enough, then *S* had a life not worth living. I take assumption 1 to be analytically true and 2 and 3 to be putatively true. Evaluating possible variations on 2 will occupy most of section 6.

2 My case will be structurally similar to that of Derek Parfit's "Repugnant Conclusion." The Repugnant Conclusion is the thesis that compared with the existence of very many people—say, ten billion—all of whom have a very high quality of life, there must be some much larger number of people whose existence, if other things are equal, would be better, even though these people would have lives that are barely worth living. In "Overpopulation and Quality of Life," Parfit imagines a different person-level analogue of the repugnant conclusion. His involves a choice between living a Century of Ecstasy versus a Drab Eternity. I will return to this case specifically in section 5.

cy ought to be rejected, and since the possibility of posthumous harm entails the possibility of Posthumous Repugnancy, we ought to reject the possibility of posthumous harm. After defending the premises from a variety of objections, I conclude that rejecting the possibility of posthumous harm in the face of Posthumous Repugnancy is preferable to all other alternatives. While I may not sway the dug-in, die-hard, posthumous harm proponent, I will have left an acute problem for them to face.

1. THE POSSIBILITY OF POSTHUMOUS HARM

Philosophers of many stripes have found compelling the idea that a subject can be harmed after their death. Endorsements of, or arguments for, the possibility of posthumous harm can be found in Nagel, Feinberg, Levenbook, Pitcher, Parfit, Grover, Sefrani, Luper, Belliotti, Boonin, and even as far back as Aristotle.[3]

The case for the possibility of posthumous harm rests crucially on a particular intuition about desire satisfaction and harm. Nearly all discussions of posthumous harm center on hypothetical cases in which some agent's desires are being frustrated while that agent is completely unaware of the frustration. Intuitively, the agent is being harmed by those frustrations. An oft-cited example comes from Feinberg:

> If someone spreads a libelous description of me among a group whose good opinion I covet and cherish, altogether without my knowledge, I have been injured in virtue of the harm done my interest in a good reputation, even though I never learn what has happened. That is because I have an interest, so I believe, in having a good reputation as such, in addition to my interest in avoiding hurt feelings, embarrassment, and economic injury. And that interest can be seriously harmed without my ever learning of it.[4]

3 Nagel, "Death"; Feinberg, *The Moral Limits of the Criminal Law*; Levenbook, "Harming Someone after His Death"; Pitcher, "The Misfortunes of the Dead"; Parfit, *Reasons and Persons*; Grover, "Posthumous Harm"; Sefrani, "Callahan on Harming the Dead"; Luper, "Posthumous Harm" and "Mortal Harm"; Belliotti, *Posthumous Harm*; Boonin, *Dead Wrong*; and Aristotle, *Nicomachean Ethics*, 1100a15–25. David Boonin's *Dead Wrong* is an excellent resource and a forcefully argued defense of the posthumous harm thesis. I will not discuss Boonin's book at any length because the problem I raise is not one he discusses. Nor is the problem I raise one that can be effectively dealt with by utilizing his various other defenses of the posthumous harm thesis. Where Boonin's discussion and mine most explicitly overlap is their discussion of the problem of non-arbitrarily prioritizing felt harms over unfelt harms. This is addressed in section 6 of this paper, where I will make a brief note regarding the relevance of Boonin's views.

4 Feinberg, *The Moral Limits of the Criminal Law*, 87.

Many authors have taken this passage from Feinberg as a natural starting point in their discussions of posthumous harm. But it is important to note that the case Feinberg gives is not enough on its own to establish the possibility of posthumous harm. Feinberg is describing a case in which someone is harmed but completely unaware of the events that are harming them. This leaves open the possibility that it is the felt effects of the unknown events that are responsible for those events being harmful and not the unknown frustration of their desires. On that interpretation of the case, it is clearly not analogous to being harmed after death. Establishing that unknown events can harm is not sufficient to establish that posthumous events can harm. To establish the possibility of posthumous harm it must be that one can be harmed but be *completely unaffected* by the harm at any time in the future, and not just unaware of it. To stave off a challenge to Feinberg on these grounds, it is useful to supplement his case with some comments from Nagel's "Death." In that paper Nagel rejects an objection to his position on the grounds that it would also rule out posthumous and unfelt harms:

> [This] type of objection is expressed in general form by the common remark that what you don't know can't hurt you. It means that even if a man is betrayed by his friends, ridiculed behind his back, and despised by people who treat him politely to his face, none of it can be counted as a misfortune for him *so long as he does not suffer as a result*. It means that a man is not injured if his wishes are ignored by the executor of his will, or if, after his death, the belief becomes current that all the literary works on which his fame rests were really written by his brother, who died in Mexico at the age of twenty-eight. It seems to me worth asking what assumptions about good and evil lead to these drastic restrictions.[5]

I will refer to these sorts of cases—cases of unknown desire frustration that in no way affects the one whose desires are being frustrated—as "Nagel-Feinberg cases." I will refer to the intuition that the agent *is harmed* in such cases as the "Nagel-Feinberg intuition."

The Posthumous Harm View is not complicated. It takes our Nagel-Feinberg intuitions about posthumous Nagel-Feinberg cases to be correct. Thus, one can be harmed by the frustration of their desires—the frustration of which has no effect on their experiences. Because a person can desire that certain things happen after their death, they can be posthumously harmed by those things not happening.[6]

5 Nagel, "Death," 76, emphasis added.

6 I realize that my characterization here makes it sound as if it is only via a desire-satisfaction principle that one could argue for the possibility of posthumous harm. That is certainly

For my purposes, I will bracket concerns as to whether the dead have desires, whether posthumous harm requires an untenable backward causation, and whether the sort of desire-satisfaction principle that undergirds the possibility of posthumous harm is defensible in the first place.[7] My task specifically is to bring attention to a previously unidentified and highly implausible result of the possibility of posthumous harm.

I also want to make a note about methodology before going further. Throughout this paper, I follow proponents of the possibility of posthumous harm and take as legitimate a philosophical methodology that relies heavily on hypothetical cases, intuitions, and the weighing of intuitions against one another. I will, like Nagel, Feinberg, and especially Parfit, appeal to considerations of comparative intuitiveness and plausibility. One might very reasonably take issue with such an approach to moral philosophy, but I will not do so here. I am confronting the proponents of posthumous harm on their own methodological turf.

2. POSTHUMOUS REPUGNANCY

Suppose one is posthumously harmed when one's desires are posthumously frustrated. Now, imagine a person named Rosa with what looks like a great life. During her life Rosa saw all her goals realized and all her projects completed to her deep satisfaction. She died peacefully, perfectly contented with how her life had gone at the age of one hundred. Few are as lucky as Rosa. But Rosa had one desire left to be satisfied—she desired that it would always be the case that whenever she was spoken of after her death, only positive things were said of her. It was not a very strong desire of hers, but she desired it nonetheless, and it was the one desire left unfulfilled when Rosa died. In fact, it was the only desire she ever had concerning what would happen after her death.

not the case. What is true, however, is that the possibility of posthumous harm has been defended almost exclusively by appeal to examples involving supposedly harmful posthumous desire frustrations. Further, unrestricted desire-satisfaction views (or sometimes just principles) of well-being are attractive in their own right, and an unrestricted desire-satisfaction principle, in conjunction with a few other widely held theses, *entails* the possibility of posthumous harm.

7 There are many who raise such concerns. For example, Partridge argues against the dead having interests or desires ("Posthumous Interests and Posthumous Respect"). Portmore argues that to be plausible at all any desire-satisfaction theory of well-being will have to restrict which desires can affect one's well-being, and further that those restrictions rule out the possibility of posthumous harm ("Desire Fulfillment and Posthumous Harm"). But it should be said that posthumous harm is possible on a variety of views, and not just a view according to which *all* that is intrinsically good or bad for a person is contingent upon whether or not their desires are satisfied or frustrated.

Unfortunately for Rosa, most everyone quickly forgot about her except for her neighbors who thought she was the Antichrist. The neighbors founded a cult whose central belief was that Rosa was the enemy of all that was good. The cult's daily observances were all centered on speaking ill of Rosa. This happened only among cult members, for no one else had been willing to listen to them for some time.

Rosa is, according to the hypothesis, repeatedly harmed by the repeated frustration of her desire that only positive things were said about her every time she was spoken of after her death.[8] Suppose the cult keeps this up for generations, perhaps thousands of years. Nothing positive is ever said of Rosa, and so, unfortunately, she is never posthumously benefited, only harmed. At some point, enough posthumous harm has been done to Rosa to outweigh all the positive value of her lived life. Eventually, her well-being will be net negative enough that it is true that she had a life not worth living. This is despite the fact that, while she was alive, she had as good a life as anyone could hope for.

Rosa's case is merely an illustration. The particular details do not matter. The example could be amended in whatever way necessary to illustrate the following, which is entailed by the possibility of posthumous harm, and which I call *Posthumous Repugnancy*:

> *Posthumous Repugnancy* (PR): A person whose well-being was *extremely high* while they were alive could incur small posthumous harms over a long enough period such that it is true in the long run that they had a life not worth living.[9]

Objections come to mind immediately. The next several sections are devoted to responding to objections. Section 3 addresses the objection that I have unjustifiably assumed the Time of Desire View of desire satisfaction to be false. Section 4 addresses the objection that Rosa cannot be harmed repeatedly by the repeated frustration of one desire as described. Section 5 addresses the response that, though *prima facie* implausible, we ought to just accept PR, just as many have accepted Parfit's original *Repugnant Conclusion* (RC). Section 6 responds to the objection that the possibility of posthumous harm does *not* entail the possibility

8 The next section is devoted entirely to responding to the worry that a desire cannot be frustrated repeatedly and thus that this claim is false.

9 Here I have formulated PR as if the Time of Object View of desire satisfaction is true. That is the view according to which the satisfaction of a desire benefits me at just those times when the desire's object obtains. In the next section, I will demonstrate how PR can be reformulated to be compatible with the Time of Desire View of desire satisfaction. That is the view according to which a desire's satisfaction benefits me at just those times when I have the desire.

of PR because even if Rosa can be repeatedly harmed by the repeated frustration of this desire she can be harmed *only so much* by the frustration of that desire and thus the posthumous harms never sufficiently aggregate to render her life not worth living.[10]

10 Bramble briefly makes an argument similar to mine ("A New Defense of Hedonism about Well-Being," 89). He points out that Emily Dickinson, Van Gogh, Nick Drake, and others had all-things-considered unfortunate lives; however, they all have enjoyed massive posthumous success. If posthumous benefit is possible, then we have to say that their lives were not that bad after all, but clearly their lives were that bad after all. Therefore, according Bramble, there is no such thing as posthumous benefit (or harm). The only way one might resist his argument, Bramble imagines, is by claiming that posthumous harms and benefits are only ever slight. He dismisses this possibility in a footnote, saying, "But in order to believe this we would need some principled reason to believe that posthumous benefits and harms could only ever be slight. I cannot myself think of what such a reason could be." I agree with Bramble, and though my core argument is similar to his, my overall defense of the impossibility of posthumous harm goes well beyond his.

First, coming up with a problem case—Bramble's Van Gogh et al., or my Rosa—is only the first part of making the case against posthumous harm. As important, and much more arduous, is the task of defending those problem cases against various defeating interpretations. Bramble defends his argument with only what I have quoted—he cannot think of a reason why posthumous benefits and harms could only ever be slight. In contrast, the majority of my paper is spent responding to objections.

Second, according to Bramble, his argument would be thwarted if it could be shown that posthumous harms and benefits are only ever slight ones. My argument would not be similarly thwarted because Rosa's case involves *only* slight posthumous harms. If by "slight harm" Bramble actually means "slight *even in the aggregate*," then I address that exact issue in section 6.

Third, there is a plausible objection to Bramble's argument that he does not address and that does not apply to mine. One could object to Bramble's argument by claiming that Van Gogh et al. primarily desired success *during their lifetimes*. They might have had no desire to be *only* posthumously successful. If that were the case, which seems plausible at least, then one could maintain that while posthumous benefit is possible, these people's lives were nonetheless not improved by their posthumous success since they did not desire to be successful in that way. That response both maintains that there is posthumous benefit but also explains how it is that these people's lives were not made better to any significant extent by their posthumous success. I do not think this objection to Bramble is ultimately successful, but it is plausible that on a clearer understanding of Van Gogh et al.'s desires it can be claimed that their posthumous success was of no great benefit to them despite posthumous benefit being possible. In contrast, Rosa's desires are stipulated. There is thus no way to make a similar objection that on a proper understanding of her desires, Rosa is actually not harmed by all the posthumous slander, despite posthumous harm being possible.

Finally, PR appears to be a nastier problem than the one Bramble raises. It appears far more unintuitive that Rosa's great life could be *not worth living* due to the aggregation of slight posthumous harms than it is unintuitive that Van Gogh's life was at least slightly less bad given stunning, worldwide, multigenerational posthumous success.

3. THE TIME OF DESIRE VIEW AND TIME OF OBJECT VIEW

Assuming I am benefited when my desires are satisfied, there is a question of *when* I am benefited. If I desire right now that there is nice weather for my bike ride this weekend and the weather is nice for my bike ride, when did the satisfaction of that desire benefit me? Was I benefited at just those times when I had the desire (Time of Desire View), or was I benefited at just those times when the object of my desire obtained (Time of Object View), or was I benefited at just those times when I had the desire *and* its object obtained (Concurrentism)?[11]

I have formulated the Rosa example as if the Time of Object View is correct. Rosa's welfare is negatively affected at those times when she is slandered after her death. It is objected that if the Time of Desire View is true, then the Rosa example does not work, and more importantly PR is not possible. The idea is straightforward. If it is true that Rosa is harmed by all the posthumous slanders at the time she has the relevant desire (which is when she is alive), then it is not true that she had exceptionally high well-being while she was alive. So it is not the case that she had an exceptionally good life while alive, only for her to be posthumously harmed enough by the aggregation of many small posthumous harms to have a life not worth living.

In response, if the Time of Desire View is true, then PR must simply be reformulated. What distinguishes the Time of Desire and Time of Object interpretations of Rosa's case is not *whether* she is harmed, or *how much* she is harmed. The views disagree only on *when* it is that she is harmed and thus disagree on when it is that her life is made one not worth living. On the Time of Object View her life is made not worth living once she is slandered enough times after death for the aggregate harm to outweigh the positive well-being she accrued while living. On the Time of Desire View, Rosa's life becomes not worth living *as soon as she forms the desire* to be spoken of only positively after death. The Time of Desire formulation of PR would thus be:

Posthumous Repugnancy TDV (PR2): A person whose well-being is *ex-*

11 For an excellent discussion of these positions and the problems they face, see Lin, "Asymmetrism about Desire Satisfactionism and Time." Concurrentism is thought to be incompatible with posthumous harm. I will thus set aside concurrentism and focus on what most people take the posthumous harm proponents' two options to be—the Time of Desire View or the Time of Object View. Lin defends another option—asymmetrism—according to which the Time of Desire View is true of *past*-directed desires and the Time of Object View is true of *future*-directed desires. Since desires about things after our deaths are always future-directed desires, Lin's position is equivalent, in this discussion, to the Time of Object View.

> *tremely high* could suddenly have a life not worth living *solely* in virtue of forming a weak desire that will be frustrated a vast number of times after their death.

The Time of Desire View formulation is substantially different from PR's initial Time of Object formulation; however, it appears no less repugnant. Both formulations share the essential repugnant feature. Both are cases where a great deal of positive well-being is swamped by a massive number of small posthumous harms. The difference is simply *when* the swamping happens, or rather *when* things get repugnant, but not *whether* things get repugnant. Thus, I conclude that the Time of Desire is not incompatible with PR suitably formulated.

4. DESIRE FRUSTRATION AND REPEATED HARM

Philosophers are surprisingly silent on the issue of whether or not a token desire can be satisfied or frustrated more than once. It is true that many desires, given their objects, can be satisfied or frustrated only once. If I desire that my package be delivered by 3 PM today, then that desire will either be frustrated or satisfied come 3 PM. The package can be delivered only once, and 3 PM today will come around only once. But not all desires are like this. Suppose I desire that my friends be honest with me. *Prima facie*, that sort of desire does not have just one chance of being frustrated or satisfied like my 3 PM package-delivery desire does.

Rosa's PR case presumes that her desire to be spoken of only positively whenever she is spoken of after her death can be frustrated repeatedly. More generally, the view I am presuming is that a single token *desire that x* of an agent *S* can be frustrated or satisfied so long as (1) *S* desires that *x*, and (2) the object of the desire, *x*, is such that the states of affairs that would satisfy or frustrate that particular *desire that x* can repeatedly obtain. Call this view the "Multiple-Frustrations View" for short. The alternative to the Multiple-Frustrations View is that a token desire can be frustrated or satisfied only once. Call this the "One-Frustration View."

The objection I want to address claims that Rosa's desire to be spoken of only positively whenever she is spoken of after death can be frustrated just once. Therefore, she cannot incur repeated posthumous harms that aggregate to the point that renders her life not worth living. The first time someone said something bad about Rosa after her death her desire was frustrated and that was the end of the story. If the One-Frustration View of desire frustration is correct, goes the objection, then PR cases like Rosa's are ruled out, for they require that some desire(s) be repeatedly frustrated.

It turns out that there is no version of a One-Frustration View that the post-

humous harm proponent can reasonably accept and that would make this objection work. Consider Michael and Dwight, who for five years both had equally strong desires that their romantic partners not cheat on them. Over the course of those five years, Michael's partner cheated on him only once, while Dwight's partner cheated on him fifty times. Michael's partner's infidelity was a mere illicit kiss that led to nothing more. Dwight's partner's infidelity started with one illicit kiss, but quickly escalated into a multiyear passionate love affair. Neither Michael nor Dwight ever found out about these infidelities, nor did they experience any effects of their partner's indiscretions. Michael and Dwight's cases are Nagel-Feinberg cases.

Remember that we are supposing in our discussion that the Nagel-Feinberg intuition that agents are genuinely harmed in Nagel-Feinberg cases is correct. We are thus not considering whether or not Michael and Dwight have been harmed *at all.* We are supposing that Michael and Dwight have been harmed. The question is, have they been harmed equally? Obviously not, it seems. The intuition that Dwight has been harmed more seems just as strong as the intuition that they have been harmed in the first place. Even if one denies that there is unfelt harm, they would surely accept the conditional that *if* there is unfelt harm, *then* Dwight was harmed more than Michael in this case.

I have introduced Michael and Dwight's case because there appears no way to explain how Michael and Dwight are harmed unequally while maintaining that Rosa's case is not possible. I endorse a Multiple-Frustrations View and according to it Dwight is harmed more because Dwight's desire was frustrated *more times* than Michael's. If a Multiple-Frustrations View is true, then Rosa's case works as described.

A proponent of a One-Frustration View could get the Michael and Dwight case right by claiming that Dwight is harmed more than Michael because Dwight's desire was frustrated only once but *to a greater degree* than Michael's. However, on the One-Frustration View plus degrees of desire frustration, Rosa's case works once redescribed as a case of her desire being frustrated to an increasing degree over time (and the harmfulness of the frustration increasing commensurately).[12]

A proponent of a One-Frustration View could get the Michael and Dwight case right by claiming that Dwight was harmed more than Michael because Dwight had a constellation of very similar fidelity-related desires and each of his partner's infidelities frustrated a different one. However, on this constellation

12 I should say that I think the correct view is multiple frustration plus degrees of frustration. There is much more to be said about this topic, but I have tried to keep this section brief.

of similar desires view Rosa's case works once redescribed as many of her very similar desires being frustrated over a very long period of time.

Rosa's case would be ruled out on a view according to which her desire that only positive things be said of her after her death is frustrated only once, and further she has no other similar desires that would be frustrated by the posthumous slander. However, on that particular One-Frustration View, Michael and Dwight are harmed *equally*, for they held the same desire at the same strength, which was frustrated for each of them only once. But it is unbelievable that Michael and Dwight would be harmed equally. Michael's partner kissed another person. Dwight's partner had a long-standing love affair with another person. I conclude therefore that if there is a problem with the Rosa case, it is *not* that her desire is frustrated only once and therefore no more harm can come to her after that.

5. THE COSTS OF ACCEPTING POSTHUMOUS REPUGNANCY

Perhaps the posthumous harm proponent ought to bite the bullet and accept PR. Derek Parfit's RC, from which PR takes its name, has been accepted by many philosophers despite its apparent implausibility. Why not do the same with PR? In this section I argue against this strategy.

Hartry Field argues that one reason to reject an epistemicist account of vagueness is that it is unreasonable to fear that noon tomorrow might be the moment you become old.[13] Epistemicism is committed to there being sharp cutoffs in vague cases. So, though it can be vague whether you are old, there is some magic moment in time when it goes from being true that you are not old to being true that you are old. Field argues that since one could not reasonably fear that the cutoff is imminent, we have reason to think it does not exist, and so we have reason to reject epistemicism.

An analogous point can be made here. Suppose Will had a really awesome life, and he knew it. He knows he is a couple of hours from death. He is told that, "You know, some people incur small posthumous harms over a long enough period of time such that even though life was really great for them while they were alive, they in fact had a life not worth living." Could Will at that moment reasonably fear that, contrary to everything he has experienced in his life, he in fact had a life not worth living? I do not think so. This reveals what I call:

13 Field, "This Magic Moment." Epistemicism is the view that vagueness is an epistemic phenomenon; specifically, vagueness consists in a special kind of ignorance. If it is vague whether *p* then it is either true that *p* or true that not *p*, however it is unknowable which it is. The *locus classicus* defense of epistemicism is Williamson, *Vagueness*. See also Sorenson, *Blindspots* and *Vagueness and Contradictions*.

> *No Reasonable Fear of* PR *Intuition*: Any person whose well-being was *extremely high* while they were alive could not, right before their death, reasonably fear that enough small posthumous harms might add up such that they in fact, and contrary to everything they have experienced, had a life not worth living.

Contrast this with someone who is thirty and knows they have probably seventy more years of life left. They have no idea how those seventy years are going to go. They could reasonably fear that enough harm will befall them in those seventy years such that in the end they will have had a life not worth living. They know there is plenty of time left for them to be harmed that much. But the same is not true of Will. Employing Field's strategy, because it would be unreasonable to fear PR, i.e., because of the No Reasonable Fear of PR Intuition, we have strong reason to doubt the possibility of posthumous harm. Admittedly, the No Reasonable Fear of PR Intuition does not constitute a *decisive* reason to reject PR. It is just an intuition. But remember that intuition plays a central role in justifying the possibility of posthumous harm in the first place. The Posthumous Harm View is supported largely on the basis of Nagel-Feinberg intuitions that persons can be harmed while their experiences are unaffected by those harms.

The No Reasonable Fear of PR Intuition concerns what attitudes it would be reasonable to have toward PR. A related intuition is worth mentioning as well. The *No Reasonable Preventive Suicide Intuition* concerns what actions it would be reasonable to take in light of PR. Suppose that I am told that tomorrow I will be kidnapped and tortured ceaselessly, but kept alive, for decades. Taken as a whole, my life will have been so bad as to not have been worth living. I could intervene however. I could kill myself today, before I am kidnapped. This would ensure that the events that would render my life not worth living—the decades of torture—would never come to pass. I will have died having had a life worth living.

Under these conditions, it is reasonable to entertain preventive suicide. It is plausible that, for any person who knows that their life will truly end up not being worth living because of *x*, it would be reasonable for that person to choose to end their life to prevent *x*.

But what about for Rosa? Remember Rosa had an amazing life but incurred enough small harms after death such that she had a life worth living. Imagine you saw Rosa on her thirtieth birthday and told her the bad news:

> Look, Rosa, I'm sorry, but you are going to wind up with a life not worth living. Sure, the next seventy years up until your death will be downright great, but so many small harms will befall you after your death that it will be true that you had a life not worth living. Luckily, you have some op-

> tions. You could kill yourself today. Sadly, you would miss out on the next seventy years of great life, but it will ensure that you are not posthumously harmed such that you end up having had a life not worth living. You have to act now and end your life or else suffer the terrible fate of Posthumous Repugnancy.

Would it be reasonable for Rosa to choose to end her life? Intuitively, definitely not. It seems absurd that she would kill herself and miss out on seventy more great years just to avoid the aggregation of many small posthumous harms. This is the:

> *No Reasonable Preventive Suicide Option Intuition*: Any person who knows they have decades of high-quality life ahead of them could not reasonably choose to commit suicide and forgo those years *merely* to prevent a large enough number of small posthumous harms.

Just as before, the intuition that it would be unreasonable to choose preventive suicide under such conditions is not decisive against the possibility of posthumous harm. But again, intuition is absolutely central to the defense of the possibility of posthumous harm in the first place.

Discussion of these intuitions helps make clear the high intuitive costs of biting the bullet and accepting PR. In accepting PR as true, one commits to it being reasonable to fear that, despite having lived an amazing life right up until death, one actually has a life not worth living. And one commits to it being reasonable to commit suicide and forgo decades of great life solely to avoid a large number of small posthumous harms.

All that being said, one could still accept PR despite its great implausibility. It is true that many philosophers accept Parfit's RC despite its initial implausibility. So why not take the same route with PR?

The disanalogies between Parfit's RC and our PR seriously undercut such a strategy. Most importantly, the primary motivation for accepting Parfit's RC is that, however implausible RC may seem, it is not as implausible as denying any one of the claims from which it follows—that *better than* is transitive, that adding a life worth living does not make a world worse *ceteris paribus*, and that increasing both the average and the total utility of a world makes that world better all other things being equal.[14]

But no such thing can be said of PR. To reject PR, we need only reject the possibility of posthumous harm, and the posthumous harm thesis is controversial to begin with. RC is so hard to avoid because to do so we need to give up what look

14 Here I am following the characterization of Huemer, "In Defence of Repugnance."

to be obvious moral truths. This is not analogous to PR. We can avoid PR merely by denying a controversial thesis about harm.

6. ON THE SUPPOSED LIMITS OF HARM

Rosa had a life of extremely high well-being while she was alive, but, if posthumous harm is possible, she repeatedly incurred small posthumous harms over a long enough period of time that she had a life not worth living. The initial response to the case usually is along the following lines: "Can't Rosa be posthumously harmed *only so much*, or *up to a point*? Posthumous harm is possible, but there's just no way that posthumous harm, however long it goes on, can render an otherwise good life not worth living." Whatever the details, the response is that for some reason the posthumous harms just cannot outweigh the positive value of Rosa's lived life.

Parfit expresses something like this view when he compares two possible futures for himself—a Century of Ecstasy versus a Drab Eternity.

> Suppose that I can choose between two futures. I could live for another 100 years, all of an extremely high quality. Call this the Century of Ecstasy. I could instead live forever, with a life that would always be barely worth living. Though there would be nothing bad in this life, the only good things would be muzak and potatoes. Call this the Drab Eternity.
>
> I believe that, of these two, the Century of Ecstasy would give me a better future. And this is the future that I would prefer. Many people would have the same belief, and preference.
>
> On one view about what makes our lives go best, we would be making a mistake. On this view, though the Century of Ecstasy would have great value for me, this value would be finite, or have an upper limit. In contrast, since each day in the Drab Eternity would have the same small value for me, there would be no limit to the total value for me of this second life. This value must, in the end, be greater than the limited value of the Century of Ecstasy.
>
> I reject this view. I claim that, though each day of the Drab Eternity would be worth living, the Century of Ecstasy would give me a better life.... The Century of Ecstasy would be better for me in an essentially qualitative way. Though each day of the Drab Eternity would have some

> value for me, no amount of this value could be as good for me as the Century of Ecstasy.[15]

Parfit's view is that there is a lexical priority in the values being compared in the Century of Ecstasy (CE) versus the Drab Eternity (DE).[16] Parfit claims that the value of CE would be better in an "essentially qualitative way" and that "no amount" of the value of DE could be as good as the value of CE. Lexical priority is the only way that DE could have value and yet have an *infinite* amount of that value *not* outweigh the *finite* value of CE.

Applied to PR, such a response would say that the value of life pre-death is lexically prior to the value involved in posthumous harm and benefit. Thus, posthumous harm has nonzero disvalue, yet no amount of posthumous harm will ever outweigh the positive value one's life accrued before death. Put otherwise, PR is not possible even if the amount of posthumous harm is infinite. This is just how no amount of the good from DE can outweigh the good of CE.

Parfit's discussion of CE and DE is brief, only a few paragraphs. He gives no full-fledged argument in defense of his position. He only draws a comparison to John Stuart Mill's qualitative distinction between "higher" and "lower" pleasures, and notes that many share his beliefs and preferences in such cases.[17]

It is well known that lexical priority views like Parfit's are problematic.[18] Parfit's view entails that *no amount* of drab (but still positive) value would be better than *any amount* of ecstasy value. Thus, if I have to choose between two futures, an ecstasy future, no matter how short, will be better than the drab but still good future, no matter how long. Three seconds of ecstasy followed by death would be a better future for me than twenty drab but still good years before I die. It would also be the case that a brief future of intense suffering—the several seconds after stubbing a toe—would be worse for me than an eternal mild hell. I doubt many would share a preference for a mild hell over a stubbed toe.

When applied to posthumous harm in particular, lexical priority renders it trivial. Imagine an extremely small pre-death harm, x. Say x harmed me in the following way: incurring x brought me from a state of maximal euphoria to a state that was 99.999… percent of maximal euphoria. Now take an infinite amount of posthumous harm y. According to this Parfit-inspired response, x would be

15 Parfit, "Overpopulation and the Quality of Life," 17–18.

16 The lexical priority claim is often expressed by saying that there is a *discontinuity* in the values involved in the CE versus DE.

17 Mill, *Utilitarianism*, ch. 2; Parfit, "Overpopulation and the Quality of Life," 17–19.

18 For discussions of the problems arising out of lexical-priority views, see Lemos, "Higher Goods and the Myth of Tithonus"; and Huemer, "Lexical Priority and the Problem of Risk."

worse for me than *y*. If posthumous harm is such that an infinite amount of it is less harmful than this puny pre-death harm, one might wonder whether posthumous harm is worth caring about at all. On this view, if I could spare you the slightest pre-death harm you can imagine, or could spare you an infinite series of the worst posthumous harms you can imagine, I ought to spare you the slight pre-death harm. Being spared the pre-death harm is what would be better for you.

On such a view the effect on a person's well-being of an individual posthumous harm (however large) is utterly trivial, perhaps infinitesimal. This does not square well with the initial Nagel-Feinberg intuitions with which our discussion started. The intuition in Nagel-Feinberg cases—e.g., your spouse is cheating on you, but you do not know it—is that you are significantly harmed. The intuition in Nagel-Feinberg cases *is not* that you are infinitesimally harmed, and that an infinite number of Nagel-Feinberg case harms would not be as bad for you as the smallest possible amount of felt harm. Reflecting on these considerations, the lexically priority response looks like a dead end.[19]

The posthumous harm proponent might try at this point to pivot to the claim that posthumous harm and pre-death harm are *incommensurable*—that they cannot measured on the same scale, or otherwise compared in quantity or magnitude. Parfit's claim that there is an essential "qualitative" difference between CE and DE does have the ring of incommensurability. Suppose one takes the incommensurability route. Immediately it is asked, "Though incommensurable with pre-death harm, does posthumous harm negatively affect one's well-being nonetheless? Put otherwise, when one is posthumously harmed, is their life made worse?"

The posthumous harm proponent cannot answer no to this question. If being posthumously harmed does not negatively affects one's well-being, then posthumous harm is not actually *any kind* of harm at all, for it is analytic that harm makes one worse off in *some way*. But to answer, "Yes, posthumous harm negatively affects one's well-being," one must explain how it is that posthumous harm negatively affects one's well-being and pre-death harm negatively affects one's

19 Ironically, were the posthumous-harm proponent to go the lexical priority route they would be saddled with the position that we ought to respect the wishes of the dead (or the well-being of the dead) *much less* than we do now. Why execute the will of the deceased when a failure to do so would be infinitely less bad than the slight inconvenience that is done to you by signing some paperwork? Why refrain from posthumously framing your rival for crimes against humanity when the harm done to him will pale in comparison to the pain you would incur having to resist framing him? Obviously, someone rejecting the possibility of posthumous harm must ultimately answer these difficult questions, but it is very strange to be a proponent of posthumous harm *and still* have to answer these questions. The lexical-priority proponent winds up having to do the double duty of explaining how there is posthumous harm and yet nearly all our intuitions about it are wrong.

well-being, and yet they still cannot be compared at all. Here one would have to claim that there are different ways to negatively affect one's well-being. Further, the way posthumous harm negatively affects one's well-being is incommensurable with the way pre-death harm negatively affects one's well-being. Even if we accept that picture, the fact that posthumous harm negatively affects one's well-being *at all* leaves open that it can negatively affect *enough* to make their life not worth living. Even if we set aside as incommensurable pre-death harm and benefit, posthumous harm still aggregates.[20]

To stop the aggregation of posthumous harm on this incommensurability picture, the posthumous harm proponent should not appeal to the lexical priority claim that no amount of posthumous harm can outweigh the value of a lived life, and that no amount of posthumous harm *is worse* than any amount of pre-death harm. It was the failure of lexical priority strategy that motivated the move to the incommensurability strategy.

The posthumous harm proponent ought to go looking for better. What is needed is something to prevent posthumous harms from sufficiently aggregating to outweigh the large positive amount of pre-death well-being, yet not render posthumous harms trivial or their effect on well-being infinitesimal. If the posthumous harm proponent does not appeal to lexical priority, or incommensurability, what then is left?

The sufficient aggregation of posthumous harm might be blocked by either a diminishing marginal value effect on posthumous harm such that posthumous harms become less and less harmful, or by a limit on the amount of posthumous harm a person can incur no matter what. Suitably formulated, either could prevent posthumous harms from rendering an otherwise great life not worth living.

How could posthumous harms diminish in harmfulness, or cease to be harmful at some limit, when the natural basis of the repeated posthumous harms—the strength of the desire, its content, and the degree to which it is frustrated—

20 Boonin confronts a related problem and winds up in the same place as we do (*Dead Wrong*, 178–79). He is concerned with how to compare on a single scale the harmfulness of unfelt harms to felt harms in a way that is non-arbitrary but also does not make unfelt harms lexically prior in harmfulness. In brief, felt harms and unfelt harms are weighted according to how much one would want to avoid them. If *S* prefers to avoid an unfelt harm *h* twice as much as a felt harm *f*, then on Boonin's view *h* would be twice as harmful for *S* than *f*. Notice that this is just what we have been assuming as our starting point—that the harmfulness of a desire frustration is a function of the strength of the desire. Whether or not the desire frustration leads to pre-death or posthumous harm does not matter. Since that view is perfectly compatible with PR, we have tried examining the alternatives—incommensurability and lexical priority—but those turned out to be too problematic. Boonin starts with incommensurability and lexical priority, finds them too problematic, and lands on a view that just so happens to be the one we started with, and one that is perfectly compatible with PR.

remains fixed? The strength and content of one's desires do not change after their death. The only explanation for posthumous harm diminishing in harmfulness, or else having a hard limit, would be that posthumous harm has either of these properties *essentially*. But this is problematic.

Take a series of temporally successive qualitatively identical posthumous desire frustrations, F_1–$F_{\ldots}$. The desire frustrations differ *only* in their location in the series (only in when they happened), everything else has been fixed by death. On the diminishing posthumous harm proposal, *how harmful* any frustration F_n is will be a function of F_n's location in the series. The harmfulness of each F diminishes as the series goes on, but all members of the series are otherwise identical. Thus, I could know everything there is to know about a desire frustration, F_n, other than where F_n occurs in the series, and yet not be able to tell you how harmful F_n is. If F_n is at the beginning of the series it could very harmful, but if F_n is much further on in the series it could be barely harmful at all. I could know everything there is to know about a pair of desire frustrations F_n and F_r and I will not be able to tell you which is more harmful if I do not know the location in the series of both F_n and F_r. More concretely, I could know (a) that when a person's desire is frustrated they are harmed, (b) that Dwight strongly desired his partner not cheat on him, (c) that Dwight's partner cheated on him, and (d) the first time it happened Dwight was very harmed by this—but I would not be able to tell you on that basis whether another identical frustration was for Dwight similarly very harmful or barely harmful at all, unless I knew when in the temporal series it appears.

On the posthumous-harm-is-limited view, whether or not some desire frustration F_x is *harmful at all* is determined by F_x's location in the series F_1–$F_{\ldots}$. On this view there is some n such that desire frustrations F_1 through F_n are equally harmful, but every frustration from F_{n+1} on is not harmful at all. Thus, I could know everything there is to know about that desire frustration, F_x, other than where F_x is in the series and not be able to tell you whether F_x is very harmful or not harmful. If F_x is at the beginning of the series it could very harmful, but if it is late enough in the series F_x could be not harmful at all even though none of its other properties would change with a change in its location in the series. And a further difficulty with this view is that, even if I knew where in the series F_x was, I still would not know whether or not it was harmful because I would need to know where the limit is. Knowing everything about the *F*s in the series will not tell me which number of frustrations, n, is the magic number where the qualitatively identical frustrations after F_n cease to be harmful. More concretely, I could know that the twentieth time Dwight was cheated on behind his back was very harmful (given his desire that it not happen), and be totally unable to tell you

whether the twenty-first time it would be harmful to him *at all*, even though the twentieth and twenty-first instances were qualitatively identical.

It would be strange indeed if knowing everything about a desire frustration other than where it appears in a series of identical frustrations would not be enough to have any guess as to the extent of the harmfulness of the frustration. I would have no clue whether the frustration is very harmful, barely harmful, not at all harmful, or anywhere in between. Knowing everything about the frustration other than its location in the series would not even be enough to know whether it is more or less *likely* that the frustration is very harmful or not at all harmful. Put otherwise, unless you know where a posthumous desire frustration lies in a series of identical frustrations, you cannot know *anything* about that frustration's level of harmfulness. On the limit view, you could even know the desire frustration's location in series and still not know whether it is harmful or not, for you would have to know where the limit is as well.

It will be helpful to make the case more concrete. Let us return to the hypothetical Nagel-Feinberg cases with which our discussion of posthumous harm began. Reflecting on such cases I think we will see that in pre-death cases of desire frustration where the strength and content of a desire remains fixed, it does not appear that the harmfulness of the desire's frustration diminishes *merely* in virtue of repetition. Nor does it appear to reach some limit all on its own. Imagine that someone is spreading libelous rumors about you but you never find out about it, nor are you otherwise affected at all. Surely, Nagel and Feinberg think, you will judge that you have been harmed.

Let us iterate this Nagel-Feinberg case. Suppose you are a traveling salesperson. Every three months you move to a new region, make new short-term friends, and then move again. You enjoy your job, and you are good at it, and you have been doing it for thirty years. However, unbeknownst to you, you have a stealth slanderer and he is quite a persistent fellow. Perhaps he felt slighted by you in high school and has been on a mission to stealthily slander you as long as you live. He has followed you as you have moved around, slandering you behind your back in each new venue. Assume that throughout your adult life your desire not to be slandered and to enjoy a good reputation has remained constant—same object and same strength. If Nagel and Feinberg are right and one instance of being slandered behind your back is harmful, then what reason would there be to think that more instances of the very same harm would become less and less harmful in virtue of repetition alone when you do not know about them nor experience any of their effects? What reason would there be to think that at some point during the stealth slandering the slander would just cease to be harmful? Put otherwise, the facts have stayed the same—same harm, same de-

sire, same strength of desiring, your complete lack of experiencing effects of the harm. The only thing that changes over time is how many times you have been slandered before. It does not appear that the harmfulness of the slander will diminish merely in virtue of how many times you are slandered if all other facts remain constant. Nor does it appear that the stealth slander would just stop being harmful all on its own. If that were the case, then there would be some *n* number of slanders such that slanders 1 through *n* were harmful but every slander from *n*+1 on was not harmful at all even though the only difference between slanders *n* and *n*+1 is simply how many slanders preceded them.

If the harmfulness of the slander in the iterated Nagel-Feinberg case above does not diminish nor does it reach a limit, then what reason is there to think that posthumous harm has either property? Remember that the Posthumous Harm View is motivated fundamentally by the intuition in the Nagel-Feinberg cases that one can be harmed without one's experience ever being affected by what has harmed them.

7. CONCLUSION

The case for posthumous harm rests crucially on the Nagel-Feinberg intuition that an agent is harmed when their desires are frustrated, even if they in no way experience effects of the frustration. We must now reassess that intuition given where we have ended up in our discussion of PR. Which of the following fares best?

1. *Bite the bullet and accept* PR: Trust the intuition that we are harmed in posthumous Nagel-Feinberg cases. Posthumous harm is possible, which entails PR. Distrust the No Reasonable Fear of PR Intuition and the No Rational Preventive Suicide Option Intuition.
2. *Bite other bullets*: Trust the intuition that we are harmed in posthumous Nagel-Feinberg cases. Posthumous harm is possible, but it does not entail PR. That is because it is an essential property of posthumous harm that an individual can be posthumously harmed only so much or else posthumous harm is marginally diminishing in harmfulness. These properties are not, however, properties of the unfelt harm involved in iterated non-posthumous Nagel-Feinberg cases.
3. *Reject the possibility of posthumous harm*: Posthumous harm is not possible and therefore PR is not possible. Trust the No Reasonable Fear of PR Intuition and the No Rational Preventive Suicide Option Intuition.

Distrust the intuition that we are harmed in posthumous Nagel-Feinberg cases.

Weighing 1, 2, and 3, it seems more plausible that our Nagel-Feinberg intuitions are in error about posthumous Nagel-Feinberg cases than that PR is possible, or that PR is not possible because posthumous harm has either of the essential properties necessary to block PR (i.e., a built-in limit or diminishing marginal harmfulness), neither of which is a property of the unfelt harm in non-posthumous Nagel-Feinberg cases.

On balance, 3 fares best from among the options, and so we ought to accept it over 1 or 2. We ought to reject the possibility of posthumous harm.[21]

University of Colorado Boulder
benjamin.kultgen@colorado.edu

REFERENCES

Aristotle. *Nicomachean Ethics*. Translated by W. D. Ross. Oxford: Oxford University Press, 2009.

Belliotti, Raymond A. *Posthumous Harm: Why the Dead Are Still Vulnerable*. Washington, DC: Lexington Books, 2011.

Boonin, David. *Dead Wrong: The Ethics of Posthumous Harm*. Oxford: Oxford University Press, 2019.

Bramble, Ben. "A New Defense of Hedonism about Well-Being." *Ergo* 3, no. 4 (2016).

Feinberg, Joel. *The Moral Limits of the Criminal Law*. Vol. 1, *Harm to Others*. New York: Oxford University Press, 1987.

Field, Hartry. "This Magic Moment: Horwich on the Boundary of Vague Terms." In *Cuts and Clouds: Vagueness, Its Nature, and Its Logic*, edited by Richard Dietz and Sebastiano Moruzzi, 200–8. Oxford: Oxford University Press, 2010.

Grover, Dorothy. "Posthumous Harm." *Philosophical Quarterly* 39, no. 156 (July 1989): 334–53.

Huemer, Michael. "In Defence of Repugnance." *Mind* 117, no. 468 (October 2008): 899–933.

21 For all their help I have to thank Mike Huemer, David Boonin, Graham Oddie, an audience at the University of Colorado Center for Values and Social Policy, and, especially, Chris Heathwood.

———. "Lexical Priority and the Problem of Risk." *Pacific Philosophical Quarterly* 91, no. 3 (September 2010): 332–51.

Lemos, Noah M. "Higher Goods and the Myth of Tithonus." *Journal of Philosophy* 60, no. 9 (September 1993): 482–96.

Levenbook, Barbara Baum. "Harming Someone after His Death." *Ethics* 94, no. 3 (April 1984): 407–19.

Lin, Eden. "Asymmetrism about Desire Satisfactionism and Time." In *Oxford Studies in Normative Ethics*, vol. 7, edited by Mark Timmons, 161–83. Oxford, UK: Oxford University Press, 2017.

Luper, Steven. "Mortal Harm." *Philosophical Quarterly* 57, no. 227 (April 2007): 239–51.

———. "Posthumous Harm." *American Philosophical Quarterly* 41, no. 1 (January 2004): 63–72.

Mill, John Stuart. *Utilitarianism*. 1863. Rutland, VT: Tuttle, 1993.

Nagel, Thomas. "Death." *Noûs* 4, no. 1 (February 1970): 73–80.

Parfit, Derek. "Overpopulation and the Quality of Life." In *The Repugnant Conclusion*, edited by Jesper Ryberg and Torbjörn Tännsjö, 7–22. Amsterdam: Kluwer Academic Publishers, 2004.

———. *Reasons and Persons*. Oxford: Oxford University Press, 1984.

Partridge, Ernest. "Posthumous Interests and Posthumous Respect." *Ethics* 91, no. 2 (January 1981): 243–64.

Pitcher, George. "The Misfortunes of the Dead." *American Philosophical Quarterly* 21, no. 2 (April 1984): 183–88.

Portmore, Douglas W. "Desire Fulfillment and Posthumous Harm." *American Philosophical Quarterly* 44, no. 1 (January 2007): 27–38.

Serafini, Anthony. "Callahan on Harming the Dead." *Journal of Philosophical Research* 15 (1990): 329–339.

Sorensen, Roy A. *Blindspots*. Oxford: Oxford University Press, 1988.

———. *Vagueness and Contradiction*. Oxford: Oxford University Press, 2001.

Williamson, Timothy. *Vagueness*. New York: Routledge, 1994.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1266

EVOLUTION, UTILITARIANISM, AND NORMATIVE UNCERTAINTY

THE PRACTICAL SIGNIFICANCE OF DEBUNKING ARGUMENTS

Andreas L. Mogensen and William MacAskill

MANY PHILOSOPHERS believe that evolutionary considerations debunk whatever ethical beliefs they explain, drawing on the assumption that natural selection does not "track the truth" when it comes to ethics. If some evaluative disposition has been favored by selection—so the thought goes—then the truth value of any associated ethical belief is entirely irrelevant in explaining the fitness advantages associated with that disposition. Only by a coincidence could it turn out that these beliefs are true, and such a coincidence cannot reasonably be expected.[1]

Some philosophers who regard evolutionary explanations as debunking hold, in addition, that whereas evolutionary considerations provide discrediting explanations for the acceptance of many normative theories, they nonetheless cannot explain why utilitarians accept utilitarianism. Belief in utilitarianism seemingly transcends our evolved biases. Evolutionary considerations are thus thought to tip the balance in favor of utilitarianism by selectively debunking its competitors.[2]

The claim that natural selection cannot explain belief in utilitarianism is *prima facie* plausible. Utilitarianism asks us to attach equal value to the well-being of all individuals and act so as to maximally promote the general welfare. Given its complete impartiality and extreme demandingness, belief in utilitarianism would seem to represent a serious cost to an organism's inclusive fitness. Belief

1 See Joyce, *The Evolution of Morality*; Ruse, *Taking Darwin Seriously*; Street, "A Darwinian Dilemma for Realist Theories of Value." Strictly speaking, Street argues that natural selection explanations are debunking iff we assume meta-ethical realism.

2 Lazari-Radek and Singer, *The Point of View of the Universe*; Singer, *The Expanding Circle* and "Ethics and Intuitions"; Greene, "The Secret Joke of Kant's Soul"; Wiegman, "The Evolution of Retribution."

in utilitarianism may therefore be thought to have emerged *in spite of* the selection pressures shaping human moral psychology.

Our concern in this paper is with the possibility that evolutionary considerations still pose a serious problem for utilitarians. One particular concern, highlighted by Kahane, goes as follows.[3] Utilitarianism tells us to do whatever maximizes well-being. This prescription is empty unless we specify the nature of well-being. However, standard beliefs about well-being are prime candidates for evolutionary debunking. It is easy to see how natural selection would have led us to believe that pleasure is good for us and pain is bad for us. It is also easy to see how it could have led us to value desire satisfaction, or the characteristic ingredients in objective theories of well-being.[4] Since it looks like the beliefs we happen to hold about well-being will be debunked if any evaluative beliefs are, utilitarianism seems to be left without any practical content, even if the utilitarian principle is not itself undermined by evolutionary considerations.

We will argue that this is not the case. In sections 1 and 2, we show that successful debunking arguments targeting standard beliefs about well-being do not undermine the practical significance of utilitarianism, provided that we understand the requirements of practical rationality as sensitive to normative uncertainty.[5]

A different way in which evolutionary considerations may be thought to pose a serious problem for utilitarians is via the claim that belief in utilitarianism turns out to be debunked via the provision of a suitable evolutionary explanation for certain commonsense moral beliefs, since utilitarianism represents the reasoned extension of those beliefs, and so belief in utilitarianism is ultimately traceable to discredited starting points.[6] In section 3, we argue that evolutionary considerations may still increase the practical significance of utilitarianism even if belief in utilitarianism is debunked by evolutionary considerations, so long as belief in competing moral theories is undermined to an even greater extent.

3 Kahane, "Evolutionary Debunking Arguments" and "Evolution and Impartiality."

4 See Crisp, *Reasons and the Good*, 121–22.

5 Our response to Kahane therefore differs importantly from recent replies due to Bramble ("Evolutionary Debunking Arguments and Our Shared Hatred of Pain") and Jaquet ("Evolution and Utilitarianism"), who both try to resist the claim that relevant commonsense beliefs about well-being are debunked. We mean to show that the practical significance of utilitarianism is not undermined even granting that these beliefs are undermined. Obviously, this claim is compatible with the view that these beliefs are not in fact debunked.

6 Tersman, "The Reliability of Moral Intuitions"; Kahane, "Evolutionary Debunking Arguments."

1. DEBUNKING ARGUMENTS AND NORMATIVE UNCERTAINTY

To make our case, we will begin by clarifying how to conceptualize the damage done by evolutionary debunking arguments.

1.1. What Does It Mean for a Theory to Be Debunked?

Typically, the notion of debunking is characterized in terms of *categorical belief*: a theory is debunked iff belief in that theory is subject to an (undefeated) defeater.[7] But we could also characterize the notion of debunking in terms of *graded belief*.[8] We would then say that successful debunking arguments require us to (significantly) reduce our credence in various normative theories.

Plausibly, a debunking argument never requires us to reduce our confidence in some ethical theory to zero. To assign credence zero to some proposition is to be certain that one could never gain evidence that would raise one's credence above zero. But it would be extreme to suppose that debunking arguments could be so forceful as to render it impossible for any future evidence to support the normative theories we currently believe. Debunking arguments do not salt the earth.

Furthermore, we should not be certain of the soundness of any evolutionary debunking argument. Critics have alleged that these arguments rest on faulty epistemological principles, disputable meta-ethical presuppositions, and even mistakes about the nature of evolutionary explanations.[9] Thus, even if you are confident that some debunking argument is sound, you ought to assign non-negligible credence to the possibility that it is not.

1.2. Rational Decision-Making under Normative Uncertainty

It is plausible that we should never be completely certain of anything in ethics. Any reasonable person should acknowledge that their values could be mistaken and assign some degree of confidence to a range of ethical views. Since these different views will often diverge in what they tell us to do, we may wonder how we are to decide what to do, given our normative uncertainty.

7 Kahane, "Evolutionary Debunking Arguments"; Joyce, *The Evolution of Morality*.

8 As noted by Nichols, "Process Debunking and Ethics," 731.

9 For epistemological objections see White, "You Just Believe That Because …"; and Vavova, "Debunking Evolutionary Debunking." For meta-ethical objections, see Kahane, "Evolutionary Debunking Arguments." For objections from the philosophy of biology, see Mogensen, "Evolutionary Debunking Arguments and the Proximate/Ultimate Distinction" and "Do Evolutionary Debunking Arguments Rest on a Mistake about Evolutionary Explanations?"; and Hanson, "The Real Problem with Evolutionary Debunking Arguments."

One possible view is that we should be guided by the theory in which we are most confident.[10] In the literature, this view is known as *My Favorite Theory*. As it turns out, My Favorite Theory is beset with problems, the most troubling of which is that its recommendations are sensitive to arbitrary choices about theory individuation.[11] In recent years, a number of philosophers have argued that in cases of normative uncertainty we ought instead to act so as to *maximize expected choice-worthiness*.[12] This view is analogous with the orthodox decision-theoretic principle of maximizing expected utility.

Here is the basic idea. In a decision situation, an agent confronts a set of options. The agent's credence function assigns a probability to each member in a finite set of first-order normative theories, corresponding to the agent's confidence in the theory. A theory ranks the agent's options in terms of their choice-worthiness. We assume (for now) that choice-worthiness is interval-scale measurable and intertheoretically comparable. Roughly, this means that each theory tells us how much more (or less) choice-worthy one option is as compared to another and each theory can be represented as ranking the options according to the same scale of choice-worthiness. The expected choice-worthiness of some action is the sum of its choice-worthiness according to each of the theories in the set, weighted according to their probability. The most appropriate option is that which maximizes expected choice-worthiness.

Consider a stylized example. Suppose *S* is 70 percent confident that some form of rights-based deontology is true. According to this theory, it is wrong to intentionally harm one person in order to prevent two others from being harmed in the same way. *S* assigns the remainder of her confidence to utilitarianism.[13] An evil mastermind offers *S* the option to electrocute *A* in order to stop

10 Gracely, "On the Noncomparability of Judgments Made by Different Ethical Theories"; and Gustafsson and Torpman, "In Defence of My Favourite Theory."

11 See MacAskill and Ord, "Why Maximise Expected Choiceworthiness?" 332–35.

12 Lockhart, *Moral Uncertainty and Its Consequences*; MacAskill, *Normative Uncertainty*; Sepielli, "What to Do When You Don't Know What to Do." For objections, see Gustafsson and Torpman, "In Defence of My Favourite Theory"; Harman, "The Irrelevance of Moral Uncertainty"; and Weatherson, "Running Risks Morally." Our argument proceeds on the assumption that maximizing expected choice-worthiness accounts are at least approximately correct, at least in contexts where the different theories in which the decision maker is confident yield choice-worthiness values that are interval-scale measurable and intertheoretically comparable. For a recent, comprehensive defense of these assumptions, see MacAskill, Bykvist, and Ord, *Moral Uncertainty*.

13 There are obviously many different varieties of utilitarianism depending on what theory of welfare is adopted and how positive and negative welfare are weighted relative to one another. We note that since the choices in the example relate only to the minimization of suffering, classical and negative utilitarianism agree in their evaluation of this case. Throughout this

B and *C* from being electrocuted by the evil mastermind. Alternatively, she can refuse and allow *B* and *C* to be electrocuted. Her decision situation might then be represented as follows:

Matrix 1

	Deontology 70%	Utilitarianism 30%
Electrocute	5	25
Don't Electrocute	25	5

The numerical values in the cells represent the choice-worthiness scores of the different actions under the two moral theories. The deontological theory ranks Don't Electrocute as most choice-worthy. The utilitarian theory ranks Electrocute as equally choice-worthy. For simplicity, we assume that utilitarianism ranks Don't Electrocute as worse than Electrocute to the same extent that the deontological theory ranks Electrocute as worse than Don't Electrocute. Given these stipulations, the expected choice-worthiness of Electrocute is 11 and the expected choice-worthiness of Don't Electrocute is 19. Therefore, the most appropriate option in light of *S*'s confidence in the two moral theories is Don't Electrocute.

Decision matrix 1 assumed that electrocution harms a person, since it causes pain. *S* might not be totally certain that pain is intrinsically prudentially bad. To take account of this, we might think of *S* as distributing her credence over four different normative theories, each representing the conjunction of a moral theory and theory of well-being.[14] Assume that *S*'s confidence in utilitarianism remains at 30 percent and her confidence in deontology at 70 percent. Suppose, in addition, that she is 99 percent confident that pain is bad and 1 percent confident that pain is indifferent. Assuming for simplicity that the probability that pain is bad or indifferent is independent of which moral theory is true, the decision matrix might then look like this:

paper, we focus principally on cases like this, since the badness of suffering is focal in Kahane's discussion. *Mutatis mutandis*, our arguments can easily be transposed to deal with other putative sources of intrinsic subjective (dis)value, belief in which may be thought subject to evolutionary debunking arguments

14 Some readers may find it strange to think that utilitarianism can be combined with the view that pain is not bad, as utilitarianism may be understood by some to include certain beliefs about the nature of well-being, or to at least exclude views that treat pain as good or indifferent. Here we understand utilitarianism simply as the view that we ought to maximize aggregate well-being, and hence as compatible in principle with any theory of well-being.

Matrix 2

	Deontology Pain is bad 69.3%	*Utilitarianism* Pain is bad 29.7%	*Deontology* Pain is indifferent 0.7%	*Utilitarianism* Pain is indifferent 0.3%
Electrocute	5	25	15	15
Don't Electrocute	25	5	15	15

The right-hand side of decision matrix 2 looks as it does because we assume that if pain is neutral then either choice is equally permissible according to either theory. The side constraint against intentional harm has no force, since *A* is not harmed by electrocution. And there would be no reason to ensure that a smaller number of people are electrocuted on utilitarianism, since being electrocuted makes no difference to a person's well-being. Whatever *S* chooses will be equally unobjectionable, whichever moral theory happens to be true.

The prescription to maximize expected choice-worthiness still tells *S* not to electrocute. Its expected choice-worthiness is 18.96, compared to 11.04 for the alternative. Having some slight worry that pain is indifferent makes no difference to what is most appropriate for *S* to do in this context.

1.3. The Significance of Debunking Arguments

Suppose *S* becomes aware of a plausible evolutionary debunking argument that considerably reduces her confidence in deontology, but not in utilitarianism. Since utilitarianism has always seemed plausible to *S* apart from the fact that it conflicts with certain entrenched deontological intuitions, she becomes a lot more confident in utilitarianism. Suppose *S* now assigns 30 percent confidence to deontology and 70 percent confidence to utilitarianism. In that case, the expected choice-worthiness of Electrocute is 18.96, while the expected choice-worthiness of Don't Electrocute is 11.04. In that case, Electrocute is the most appropriate choice under normative uncertainty.

What if *S* is also made aware of a debunking argument targeting her belief that pain is bad? Well, if she loses all confidence in the badness of pain, this would mean that Electrocute and Don't Electrocute are equal in terms of expected choice-worthiness. In that case, the fact that she is also quite confident that utilitarianism is the correct moral theory would be genuinely irrelevant.

However, we have already ruled out the idea that debunking arguments require us to reduce our confidence to zero. Suppose, more realistically, that *S* ends up only 30 percent confident that pain is bad. In that case, the expected choice-worthiness of Electrocute is 16.2 and the expected choice-worthiness of Don't Electrocute is 13.8. Electrocute remains the most appropriate choice.

In fact, it should be straightforward to see that so long as *S* retains some con-

fidence in the badness of pain, reducing her confidence in this proposition to any arbitrary degree ultimately makes no difference to what would be most appropriate, given her relative confidence in utilitarianism vis-à-vis deontology. If pain is indifferent, then either action is equally choice-worthy no matter which moral theory is true. The normative theories represented in the right-hand side of the second decision matrix make no difference to the relative expected choice-worthiness of the two options. The question of which action is most choice-worthy in expectation is decided entirely by how *S* distributes her confidence across those normative theories on which pain is bad, represented in the left-hand side of the decision matrix. Therefore, so long as her relative confidence in utilitarianism is significantly greater, Electrocute remains the most appropriate option.[15]

Therefore, the availability of a debunking argument targeting the belief that pain is bad turns out to be without practical significance. As we recall, the debunking argument targeting *S*'s deontological moral intuitions did make a significant difference. In light of that argument, Electrocute became the most appropriate choice. And the fact that *S* is significantly more confident of utilitarianism ensures that this remains so regardless of the extent to which she reduces her confidence that pain is bad, so long as it remains above zero.

2. WHAT FOLLOWS?

Our discussion in the previous section focused on a stylized example, constructed using a number of simplifying assumptions. What does this case really tell us about our actual practical predicament?

2.1. Beyond Expected Choice-Worthiness

The example presumed that the normative theories to which *S* assigns credence yield choice-worthiness rankings that are interval-scale measurable and intertheoretically comparable. This might seem unrealistic.[16] Where these assumptions do not hold, we cannot act so as to maximize expected choice-worthiness. We have to apply some other rule.

Fortunately, this makes no difference to the key point for which we have argued. On any plausible principle for decision-making under normative uncertainty, the most appropriate option will be determined purely by *S*'s credence in those normative theories that assume the badness of pain. Her credence in those

15 Compare Ross, "Rejecting Ethical Deflationism," on the irrelevance of "uniform ethical theories" given normative uncertainty.

16 Gracely, "On the Noncomparability of Judgments Made by Different Ethical Theories"; and Ross, "Rejecting Ethical Deflationism."

theories that treat pain as indifferent will be irrelevant, since they treat her choice as indifferent. Only those theories that assume pain's badness can tip the balance.

By way of illustration, consider a principle that works for purely ordinal theories: the *Borda rule*.[17] According to the Borda rule, one option is more appropriate than another iff it receives a higher *credence-weighted Borda score*. An option's Borda score according to some theory is the number of options to which it is superior, minus the number of options to which it is inferior. Its credence-weighted Borda score is the sum of its Borda score under each theory multiplied by one's credence in the theory.

Suppose that deontology and utilitarianism did provide only an ordinal ranking of *S*'s options in terms of choice-worthiness. Given the previously stipulated confidence levels assigned by *S* to deontology, utilitarianism, pain's badness, and pain's indifference, her credence-weighted Borda-score for Electrocute is 0.12. For Don't Electrocute, it is −0.12. Electrocute is still most appropriate.

Furthermore, it is relatively easy to work out that the relative ranking of *S*'s options in terms of their credence-weighted Borda score is insensitive to her credence in pain's badness vis-à-vis its indifference, in that neither normative theory on which pain is indifferent contributes to the credence-weighted Borda score of either option. In this respect the Borda rule behaves just like the principle of maximizing expected choice-worthiness. And any other plausible principle should behave similarly.

2.2. Beyond Harm

Another respect in which the decision situation we have considered might be thought unrepresentative is that only the avoidance of harm was assumed to have normative significance.

However, a deontological theory might well posit that a rights violation occurs when one person electrocutes another without their consent, even if doing so is harmless. In that case, the deontological theory favors Don't Electrocute even on the assumption that pain is indifferent. *S*'s choice situation might then look like this:

Matrix 3

	Deontology Pain is bad 9%	*Utilitarianism* Pain is bad 21%	*Deontology* Pain is indifferent 21%	*Utilitarianism* Pain is indifferent 49%
Electrocute	5	25	10	15
Don't Electrocute	25	5	20	15

17 MacAskill, "Normative Uncertainty as a Voting Problem."

Here, the expected choice-worthiness of Electrocute remains highest. However, this can change if *S* becomes even more confident that pain is indifferent. Suppose she is 90 percent confident that pain is indifferent. Then the expected choice-worthiness of Electrocute becomes 14.05. The expected choice-worthiness of Don't Electrocute becomes 15.95. Don't Electrocute would then be most appropriate.

The reason for this should be clear. The utilitarian theory on which pain is indifferent does not tell for or against Electrocute. By contrast, the deontological theory on which pain is indifferent tells against. The more confident *S* becomes that pain is indifferent, the more weight she gives to these theories in deciding what to do. Since the utilitarian theory is indifferent on this point whereas the deontological theory is not, increasing her confidence that pain is indifferent strengthens her reasons for choosing Don't Electrocute.

It does not follow that the combined effect of a successful debunking argument targeting *S*'s deontological intuitions and another targeting her belief in the badness of pain will generally leave everything as it was before. This will hold true in some decision situations, but not in others. Whether things are left unchanged in any given case will be highly sensitive to the confidence *S* actually assigns to utilitarianism vis-à-vis deontology and to the badness of pain vis-à-vis its indifference. It will also be highly sensitive to the particular choice-worthiness ordering generated by each theory. This is easy to see by tinkering with the credences and rankings we used above. Slight adjustments can easily tip the balance.

It would be an astonishing coincidence if our credences and choice-worthiness rankings were calibrated so that reducing our confidence in deontology and in our beliefs about well-being never made any difference to which option was most appropriate in cases that potentially involve violation of side constraints. Furthermore, side constraints are just one point of contention between deontology and utilitarianism. Many of the remaining contrasts are purely a matter of how to weigh harms and benefits befalling different people. For example, deontological theories typically posit *agent-centered permissions*, in light of which each person is entitled to attach added weight to her own well-being. Deontological theories may also posit *irrelevant utilities*: a non-consequentialist might think it is more important to save a single individual from some terrible harm than provide a trivial benefit to each person in an arbitrarily large group of people.[18] The aggregative character of utilitarianism rules out this possibility.

In choice situations where agent-centered permissions or irrelevant utilities lead deontological theories to issue prescriptions that run against the implications of utilitarianism due to intertheoretic disagreement about the weighting

18 Kamm, *Morality, Mortality*, vol. 1.

of harms and benefits, reducing one's confidence in deontology will make an important practical difference, whereas reducing one's confidence that one's actions will make any difference to people's well-being will make no difference.

2.3. What about Really Bizarre Views?

A final worry centers on the possibility that debunking arguments require us to increase our credence in bizarre views about the nature of well-being. For example, we should perhaps increase our credence in the view that pain is intrinsically good for us and pleasure intrinsically bad, as we can be confident that this view would not have been selected for. But we have so far ignored this possibility.

In a similar vein, Kahane notes that certain highly counterintuitive beliefs about well-being will resist evolutionary explanation: "These would include the views that the good life consists of ascetic contemplation of deep philosophical truths, or celibate spiritual communion with God, or a kind of Nietzschean perfectionist aestheticism (which might even revel in pain), and so forth."[19] In combination with such theories, he notes, utilitarianism might retain its practical significance. However, its implications would be utterly repugnant: few people would be able to accept these implications. Is our argument vulnerable to this sort of worry? Does the ability of bizarre moral views to escape debunking mean that they are likely to end up playing a substantial role in determining what is most appropriate in light of our normative uncertainty?

That would be the case if evolutionary debunking arguments pushed our confidence in commonsense views about well-being down so far that it was not appreciably higher than our confidence in these wildly counterintuitive theories. We could end up in this position if debunking arguments required us to reduce our confidence in commonsense intuitions very close to zero. But the effect of encountering these arguments will not be so catastrophic. Debunking arguments may seem convincing, but it is far from certain that they are sound. For this reason, we ought to retain significant credence in commonsense views about well-being of which we were extremely confident prior to encountering these arguments. In the examples we considered earlier, we set *S*'s posterior credence in pain's badness at 30 percent or 10 percent. Given *S*'s antecedent confidence and the controversy surrounding the soundness of debunking arguments, even this might be too low.

If she is like the authors, *S* would have assigned a much, much lower prior probability to the view that pain is good or that celibate spiritual communion with God is the key determinant of well-being. Her posterior confidence in com-

19 Kahane, "Evolution and Impartiality," 334.

monsense views could therefore be orders of magnitude greater than her credence in wildly counterintuitive theories of this kind. The practical significance of these views would therefore be negligible.[20]

Of course, this would *not* be the case if her confidence in these counterintuitive theories should increase significantly upon encountering debunking arguments. That would be the case if one of these theories of well-being was like utilitarianism in that it seems plausible apart from the fact that it conflicts with certain entrenched commonsense intuitions that now get debunked, provided that the plausibility of the theory itself remains intact in the face of debunking arguments.

However, the theories considered here do not seem to fit that description. The view that pain is intrinsically good is not the sort of view that seems somewhat plausible, except for the fact that it conflicts with intuition. As we see it, it has basically zero inherent plausibility. The view that the good life is centered on celibacy, meditation, and prayer strikes us as false principally because it attaches value to things that seem valueless owing to our confidence that God does not exist. Debunking arguments will not change that fact.[21] We are more attracted to the view that contemplation of philosophical truths or the realization of aesthetic value can be intrinsic sources of well-being. Theories that count such goods as the primary or only determinants of well-being seem weird to us principally because they attach too little value to other things, such as pleasure or desire satisfaction. Nonetheless, these theories do not fit the criterion we specified above. To the extent that such theories have plausibility in light of the intuitive value of knowledge and aesthetic excellence, they will lose plausibility in the face of debunking arguments. After all, it is easy to see why natural selection should lead human beings to value knowledge: we are informavores by design.[22] There is also good reason to expect that natural selection has played a significant role in shaping our aesthetic responses.[23]

20 For the view that pain is good and pleasure is bad, there is a further argument for discounting its practical significance. When combined with utilitarianism, this view has exactly opposite recommendations to classical utilitarianism. Therefore, under normative uncertainty this theory simply "cancels out" part of one's credence in classical utilitarianism. For example, with 60 percent credence in deontology, 38 percent credence in classical utilitarianism, and 2 percent credence in pain-is-good utilitarianism, a rational decision maker will take the same actions as if she had 60 percent credence in deontology, 36 percent credence in classical utilitarianism, and 4 percent credence in a view that was indifferent between all options.

21 Except perhaps to increase our confidence in atheism; see Wilkins and Griffiths, "Evolutionary Debunking Arguments in Three Domains."

22 Dennett, *Consciousness Explained*, 176–82.

23 Dutton, *The Art Instinct.*

It might be argued that our confidence in the verdict that, say, pain is good should rise significantly once we are made aware of relevant evolutionary debunking arguments, simply because the belief that it is not the case that pain is intrinsically good has an evolutionary explanation. To the extent that evolutionary debunking arguments are sound, this ethical belief ought therefore to end up being debunked. In order to maintain probabilistic coherence, our credence that pain is good must rise accordingly, and so must rise significantly.[24]

We are not convinced by this line of argument. To see why, let us start by asking in what sense the belief that it is not the case that pain is intrinsically good can be said to have an evolutionary explanation. There are many things of which we are confident that they are not intrinsically good, such as having an odd number of hairs on one's left shin. In some sense, this confidence is explained in terms of evolution by natural selection, since "all phenotypes are to some extent the products of the process of evolution by natural selection."[25] Nonetheless, it is highly implausible to suppose that there existed some specific selection pressure that accounts for our confidence that it is not intrinsically good to have an odd number of hairs on one's left shin. Rather, we can presume that our confidence in this hypothesis is explained by considerations of parsimony, given the absence of any perceived reason to accept any contrary hypothesis. A similar story presumably accounts for our confidence that it is not the case that it is intrinsically bad to have an odd number of hairs on one's left shin. We are confident of these things on roughly the same grounds that we are confident that there is no luminiferous aether—because it is the simpler hypothesis.

On its face, beliefs such as that it is not intrinsically good (or bad) to have an odd number of hairs on one's left shin are not within the scope of evolutionary debunking arguments, precisely because they are not explained in terms of specific selection pressures yielding particular ethical intuitions and can instead be explained at a proximate level in terms of the application of a domain-general principle of parsimony.

We note, then, that we also find it implausible to suppose that there existed any specific selection pressure that accounts for our confidence that it is not the case that pain is intrinsically good—over and above whatever selection pressures account for the judgment that pain is intrinsically bad. With respect to the confidence previously assigned to that hypothesis, it is possible to redistribute that confidence over two alternative hypotheses: namely, that pain is intrinsically neutral and that pain is intrinsically good. The same principles of parsimony

24 We are grateful to an anonymous referee for this suggestion.

25 Brandon, *Adaptation and Environment*, 41.

that should lead us to be confident that it is neither intrinsically good nor intrinsically bad to have an odd number of hairs on one's left shin should presumably lead us to redistribute probability mass from the proposition that pain is intrinsically bad to the proposition that pain is intrinsically neutral, leaving our credence in the hypothesis that pain is intrinsically good effectively unchanged, remaining anchored at a very low prior.

3. UTILITARIANISM DEBUNKED?

Throughout sections 1 and 2, we have operated under the assumption that, whereas evolutionary considerations provide discrediting explanations for the acceptance of many normative theories, they nonetheless cannot explain why utilitarians accept utilitarianism. As a result, we have assumed that belief in utilitarianism is not debunked by evolutionary considerations. We have focused our attention on the worry that utilitarianism may nonetheless be robbed of its practical significance, given that our ordinary beliefs about the nature of well-being seem vulnerable to debunking arguments. In this final section, we briefly outline how our conclusions may nonetheless go through—and for roughly the same reasons—even if we grant that belief in utilitarianism can also be debunked.

We argued earlier that since belief in utilitarianism seems to represent a significant cost to an organism's inclusive fitness, belief in utilitarianism may be thought to have emerged in spite of—and not because of—the selection pressures shaping human moral psychology. A standard response to this suggestion is that we can explain belief in utilitarianism as the reasoned extension of the more restricted forms of benevolence and impartiality that we expect natural selection to have favored in the environment of evolutionary adaptedness, placing belief in utilitarianism within the scope of discrediting evolutionary explanations after all. As Kahane puts it: "If a disposition to partial altruism was itself selected by evolution, then the epistemic status of its reasoned extension should also be suspect."[26]

Let us grant that utilitarianism represents the reasoned extension of more fundamental evolved evaluative judgments, such as that it is morally right to help one's kin and the members of one's community. Presumably, standard non-consequentialist theories also derive from the same evolved evaluative judgments. Furthermore, it seems plausible that standard non-consequentialist theories hew more closely to these evolved evaluative judgments than do utilitarian moral theories. If this is the case, then, it seems plausible that, to the extent that these

26 Kahane, "Evolutionary Debunking Arguments," 119.

beliefs are debunked, we ought to end up increasing our relative confidence in utilitarianism vis-à-vis other standard moral theories. In other words, we ought to reduce our confidence in standard non-consequentialist theories to a greater extent than we ought to reduce our confidence in utilitarianism, since standard non-consequentialist theories stick closer to the evaluative judgments that end up being debunked. If, in addition, the confidence that we lose in these standard normative theories is redistributed to the hypothesis that nothing matters and so all options are equally choice-worthy, then, since any hypothesis that entails that all options are equally choice-worthy cuts no ice with respect to the appropriateness of the different options available to us under conditions of moral uncertainty, for the reasons explained previously in this paper, it will end up being the case that evolutionary considerations shift our decision making in the direction of utilitarianism by virtue of increasing our confidence in utilitarianism vis-à-vis its standard competitors, even granting that we ought to significantly reduce our confidence in utilitarianism.[27]

4. Conclusion

Assuming that we ought to take normative uncertainty into account, debunking arguments that selectively undermine non-utilitarian theories have genuine practical significance, even if we are also aware of debunking explanations targeting our beliefs about well-being. The latter do not rob utilitarianism of its practical significance. Given the resulting credence distribution over different moral theories and theories of well-being, the most appropriate action will in many cases accord with the action required by utilitarianism in combination with commonsense theories about well-being. Furthermore, the effect of debunking arguments may be similar even if we ought to significantly reduce our confi-

27 It may be objected that it is a mistake to assume that probability mass that is shifted from utilitarianism, deontology, and other normative theories with roots in our evolved moral intuitions should be redistributed to the hypothesis that all options are equally choice-worthy. To the extent that we lose confidence in these different moral theories, it may be argued that we ought instead to gain confidence in *nihilism*, interpreted as the view that all options are *incomparable* in respect of choice-worthiness, as opposed to equally choice-worthy. This need not undermine our argument. Ross argues that nihilism, so understood, can also be ignored under conditions of moral uncertainty ("Rejecting Ethical Deflationism"). MacAskill ("The Infectiousness of Nihilism") raises a number of objections to Ross, but MacAskill, Bykvist, and Ord (*Moral Uncertainty*) go on to outline an improved theory of rational decision-making under moral uncertainty that also allows us to treat all but full confidence in nihilism as practically irrelevant.

dence in utilitarianism in light of evolutionary debunking arguments, so long as other moral theories end up being undermined to an even greater extent.[28]

University of Oxford
andreas.mogensen@philosophy.ox.ac.uk
william.macaskill@philosophy.ox.ac.uk

REFERENCES

Bramble, Ben. "Evolutionary Debunking Arguments and Our Shared Hatred of Pain." *Journal of Ethics and Social Philosophy* 12, no. 1 (September 2017): 94–101.

Brandon, Robert. *Adaptation and Environment*. Princeton: Princeton University Press, 1989.

Crisp, Roger. *Reasons and the Good*. Oxford: Oxford University Press, 2006.

de Lazari-Radek, Katarzyna, and Peter Singer. *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford: Oxford University Press, 2014.

Dennett, Daniel. *Consciousness Explained*. London: Penguin, 1991.

Dutton, Denis. *The Art Instinct: Beauty, Pleasure, and Human Evolution*. Oxford: Oxford University Press, 2009.

Gracely, Edward J. "On the Noncomparability of Judgments Made by Different Ethical Theories." *Metaphilosophy* 27, no. 3 (July 1996): 327–22.

Greene, Joshua. "The Secret Joke of Kant's Soul." In *The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 35–80. Vol. 3 of *Moral Psychology*. Cambridge, MA: MIT Press, 2008.

Gustafsson, Johan E., and Tom Torpman. "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly* 95, no. 2 (March 2014): 159–74.

Hanson, Louise. "The Real Problem with Evolutionary Debunking Arguments." *Philosophical Quarterly* 67, no. 268 (July 2017): 508–33.

Harman, Elizabeth. "The Irrelevance of Moral Uncertainty." In *Oxford Studies in Metaethics*, vol. 10, edited by Russ Shafer-Landau, 53–79. Oxford: Oxford University Press, 2015.

Jaquet, Francois. "Evolution and Utilitarianism." *Ethical Theory and Moral Practice* 21, no. 5 (November 2018): 1151–61.

Joyce, Richard. *The Evolution of Morality*. Cambridge, MA: MIT Press, 2006.

28 We wish to thank Guy Kahane for comments on an early draft of this paper, as well as the anonymous referees who offered insightful comments and criticisms during the review process.

Kahane, Guy. "Evolution and Impartiality." *Ethics* 124, no. 2 (January 2014): 327–41.

———. "Evolutionary Debunking Arguments." *Noûs* 45, no. 1 (March 2011): 103–25.

Kamm, Frances M. *Morality, Mortality*. Vol. 1, *Death and Whom to Save from It*. Oxford: Oxford University Press, 1993.

Lockhart, Ted. *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press, 2000.

MacAskill, William. "The Infectiousness of Nihilism." *Ethics* 123 no. 3 (April 2013): 508–20.

———. *Normative Uncertainty*. Doctoral thesis, University of Oxford, 2014.

———. "Normative Uncertainty as a Voting Problem." *Mind* 125, no. 500 (October 2016): 967–1004.

MacAskill, William, Krister Bykvist, and Toby Ord. *Moral Uncertainty*. Oxford: Oxford University Press, 2020.

MacAskill, William, and Toby Ord. "Why Maximise Expected Choiceworthiness?" *Noûs* 54, no. 2 (June 2020): 327–53.

Mogensen, Andreas L. "Do Evolutionary Debunking Arguments Rest on a Mistake about Evolutionary Explanations?" *Philosophical Studies* 173, no. 7 (July 2016): 1799–1817.

———. "Evolutionary Debunking Arguments and the Proximate/Ultimate Distinction." *Analysis* 75, no. 2 (April 2015): 196–203.

Nichols, Shaun. "Process Debunking and Ethics." *Ethics* 124, no. 4 (July 2014): 727–49.

Ross, Jacob. "Rejecting Ethical Deflationism." *Ethics* 116, no. 4 (July 2006): 742–68.

Ruse, Michael. *Taking Darwin Seriously: A Naturalistic Approach to Philosophy*. Oxford: Blackwell, 1986.

Sepielli, Andrew. "What to Do When You Don't Know What to Do." In *Oxford Studies in Metaethics*, vol. 4, edited by Russ Shafer-Landau, 5–28. Oxford: Oxford University Press, 2009.

Singer, Peter. "Ethics and Intuitions." *Journal of Ethics* 9, nos. 3–4 (October 2005): 331–52.

———. *The Expanding Circle: Ethics and Sociobiology*. Oxford: Clarendon Press, 1981.

Street, Sharon. "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, no. 1 (January 2006): 109–66.

Tersman, Folke. "The Reliability of Moral Intuitions: A Challenge from Neu-

roscience." *Australasian Journal of Philosophy* 86, no. 3 (September 2008): 389–405.

Vavova, Katia. "Debunking Evolutionary Debunking." In *Oxford Studies in Metaethics*, vol. 9, edited by Russ Shafer-Landau, 76–101. Oxford: Oxford University Press, 2014.

Weatherson, Brian. "Running Risks Morally." *Philosophical Studies* 167, no. 1 (January 2014): 141–63.

White, Roger. "You Just Believe That Because …" *Philosophical Perspectives* 24, no. 1 (December 2010): 573–615.

Wiegman, Isaac. "The Evolution of Retribution: Intuitions Undermined." *Pacific Philosophical Quarterly* 98, no. 2 (June 2017): 193–218.

Wilkins, John S., and Paul E. Griffiths. "Evolutionary Debunking Arguments in Three Domains: Fact, Value, and Religion." In *A New Science of Religion*, edited by Greg Dawes and James Maclaurin, 133–46. London: Routledge, 2013.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1768

THE VALUE OF A LIFE-YEAR AND THE INTUITION OF UNIVERSALITY

Marc Fleurbaey and Gregory Ponthiere

As highlighted by Sen in his early criticism of national accounts statistics, the measurement of the achievements of a society can hardly abstract from how long the members of that society live.[1] Individual lifetime is a key dimension to be taken into account in the measurement of social achievements when one wants to measure the degree of development of a society or social welfare. That intuition motivated the inclusion of indicators of human lifetime—such as period life expectancy at birth—in the construction of indexes of development, such as the Human Development Index (HDI), and in the construction of inclusive indexes of well-being, such as equivalent income.[2] Moreover, since various public policies modify the production and the distribution of life-years within a population (e.g., health care programs, transportation policies, and environmental policies), the lifetime dimension can hardly be ignored when considering the design of policies.[3]

Taking human lifetime into account requires the social evaluator or the government to find a way to weight the quantity of life-years against other relevant dimensions of life. In other terms, the social evaluator needs to find a way to assign a value to life-years. Assigning a value to life-years is necessary not only for the measurement of human development or well-being, but also to be able to solve policy dilemmas involving various implications in terms of the production and distribution of life-years within a population. This necessity to assign a value to life-years has given rise in economics to the increasingly large literature on the value of a statistical life.[4] Empirical estimates of the value of a statistical life

1 See Sen, "On the Development of Basic Income Indicators to Supplement GNP Measures."

2 See United Nations Development Programme, *Human Development Report 1990* and *Human Development Report 2010*. On the equivalent income index, see Fleurbaey and Ponthiere, "Prevention against Equality?"

3 See Sen, "Mortality as an Indicator of Economic Success and Failure"; Broome, *Ethics out of Economics* and *Weighing Lives*.

4 See Jones-Lee, *The Economics of Safety and Physical Risk*. The value of a statistical life is

have become key parameters in cost-benefit analyses, as well as in the analysis of optimal policies, in particular in the context of climate change.

The empirical literature on the value of a statistical life has a purely positive nature: it aims at quantifying how individuals tend, in real life, to solve trade-offs involving risk about the duration of life, or in other words, how individuals are willing to exchange money against variations of the risk of death. Empirical studies of various kinds (wage-risk studies, contingent valuation methods, etc.) show that the value of a statistical life varies with several factors, such as age and occupation, and increases with income. For instance, in his meta-analysis of empirical studies, Miller provides a rule of thumb for the valuation of a life, which is a simple linear transformation of the gross domestic product (GDP) per capita.[5]

Obviously, it does not logically follow from those empirical studies that, when measuring social achievements or designing policy objectives, the social evaluator should make the valuation of life or the valuation of a life-year dependent on income or on any of those factors. Empirical studies on the value of life may well show that the value of a life-year is increasing with income, but such a positive premise cannot, if taken separately, lead to any normative corollary. Drawing such a conclusion would be nothing other than an occurrence of a naturalistic fallacy. There is thus a need to examine, at the normative level, how a social evaluator should value an extra life-year—that is, the principles that should govern such a valuation of life-years.

From a normative perspective, the valuation of a life-year leads to two conflicting intuitions: on the one hand, the intuition of universality, and on the other hand, the intuition of complementarity.

According to the intuition of universality, the value of a life-year should be universal. The value of a life-year should be the same whatever contexts are considered; in particular, a life-year should have exactly the same value when we consider a poor or rich country. Thus, from that perspective, the fact that life-years are more valued in richer countries than in poorer countries should be irrelevant when considering the social valuation of those life-years: universalism requires life-years to be valued in the same way, independently from the context, and, in particular, independently from the associated quality of life.

Such an intuition of universality concerning life-year valuations was defended, among others, by Anand, who criticized the new HDI—which is based on a geometric average rather than an arithmetic average across income, lifetime, and

defined as the value that x individuals assign to a reduction of the risk of death from $1/x$ to 0, leading to saving one life.

5 See Miller, "Variations between Countries in Values of Statistical Life." According to Miller, the value of a statistical life lies between 120 and 180 times the GDP per capita.

education dimensions—on the grounds that it violates this intuition of universal valuation of life-years.[6] According to Anand, the value of an extra life-year should be the same, whatever one considers a rich or poor country. That intuition of universality is satisfied by the standard HDI, but not by the multiplicative HDI.

This intuition of universality conflicts with another intuition, which can be called the intuition of complementarity. According to that intuition, the quantity of life cannot be valued independently from the quality of life. The reason lies in the singular nature of lifetime in comparison to other goods. Lifetime is not a good like a standard good, which could be enjoyed on its own. On the contrary, lifetime is like a "container," whose value depends on what it will allow for—that is, on what lies "within the container" (life events, activities, projects, etc.). As a consequence, the valuation of life-years cannot be made independently from the associated quality of life. A corollary of this is that, when the quality of life varies, the value of the quantity of life cannot remain the same, and, hence, cannot be universal, in opposition to what the intuition of universality recommends.

The intuition of complementary can take two forms: a weak version, according to which the value of a life-year depends, among other things, on the quality of that life-year, and a strong version, according to which the value of a life-year depends only on the quality of that life-year. Although the latter version is much stronger than the former, it finds some support in several simple thought experiments. For instance, if one could artificially slow down life without modifying the number of life events that take place in that life, then one could hardly regard this lengthening of life as valuable: enlarging the size of the "container" without modifying its contents could hardly make a life better. Alternatively, consider another thought experiment, where one can shut down life during one hour, and shorten all lives by that amount, without anyone being aware of that shutdown. As long as this temporary shutdown was unnoticed, and did not affect events among humans, it is hard to see how this reduction of the size of the "container" could reduce the value of life. Thought experiments provide some support for the strong version of the intuition of complementarity. But it is important to stress that even the weak version of that intuition is in conflict with the intuition of universality.

In order to better present the differences between the intuition of universality and the intuition of complementarity, it can also be useful to refer to the concept of intrinsic value of life, that is, the value of life *per se*, independently from the characteristics of that life. According to the intuition of universality, the value of a life-year is composed exclusively of its intrinsic value, and thus

6 See Anand, "Recasting Human Development Measures." The treatment of lifetime within the multiplicative HDI is also criticized by Ravallion; see "Troubling Trade-offs in the Human Development Index."

does not depend on the quality of that life-year (only the "container" has value). On the contrary, the weak version of the intuition of complementarity states that the value of a life-year may or may not have an intrinsic component, but for sure includes a component that is related to the quality of that life-year (what is inside the "container" matters, and possibly the "container" as well). Finally, the strong version of the intuition of complementarity states that the value of a life-year only includes a component that is related to its quality, and includes no intrinsic value component (only what is inside the "container" is valued, not the "container" itself).

The incompatibility between the intuition of universality and the intuition of complementarity raises deep challenges for the valuation of life-years. Although this is not often explicitly acknowledged, a large part of the literature on the valuation of life relies on the intuition of complementarity, and as such violates the intuition of universality.[7] Those violations are problematic only to the extent that the intuition of universality is worth being pursued. But the ethical appeal of the intuition of universality is hard to evaluate, simply because no precise account of that intuition has been given so far. As a consequence, it is difficult to have an idea of the precise implications of adopting such an intuition.

The goal of this paper is precisely to provide a more accurate account of the intuition of universality, in order to better discuss its implications for the valuation of life-years. In order to better understand what the intuition of universality is, we propose to study its implications for the valuation of life-years when that intuition is true. For that purpose, this paper will proceed in two stages. In a first stage, we provide three distinct definitions of the intuition of universality, in terms of the constraints this intuition imposes on the form of value functions aimed at valuing lives. That first approach informs us about the formal constraints that the intuition of universality imposes on the valuation of the quantity and quality of life, but does not inform us about the priority to be given in allocation problems. Then, in a second stage, we consider a more general approach, in terms of social preferences, and reformulate those three accounts of the intuition of universality, in order to explore their consequences in terms of priority when considering problems of life-year allocations.

Our main results are twofold. First, we show that the three distinct accounts

7 This is true for indicators of well-being relying on a life-cycle perspective, such as the equivalent income approach. See Costa and Steckel, "Long-Term Trends in Health, Welfare, and Economic Growth in the United States"; Nordhaus, "The Health of Nations"; Becker, Philipson, and Soares, "The Quantity and Quality of Life and the Evolution of World Inequality"; Fleurbaey and Blanchet, *Beyond GDP*. This is also the case for normative studies on compensation for unequal lifetime, such as Fleurbaey and Ponthiere, "Prevention against Equality?"

of the intuition of universality lead to quite counterintuitive implications from a normative perspective. One of these is shown to be in conflict with a basic property of monotonicity, whereas two other accounts of the intuition of universality lead to indifference with respect to how life-years are distributed within the population, which is also quite counterintuitive. Thus adopting a universal perspective on life-years valuations—and, thus, abstracting from the associated quality of those life-years—leads to quite questionable consequences. Those negative results support the abandonment of the intuition of universality. However, we show that abandoning the intuition of universality, and adopting instead the intuition of complementarity, does not prevent a social evaluator from giving priority, when allocating life-years, to individuals with the lowest quality of life.

This paper is related to several branches of the literature. First of all, it is related to the literature on the measurement of human development, such as Ravallion and Anand, who criticize the geometric HDI on the grounds of an intuition that is strongly related to the intuition of universality as explored in this paper.[8] Our paper is also related to multidimensional indicators of well-being inclusive of the lifetime dimension, such as the equivalent income.[9] Second, this paper is also related to the literature on the normative foundations of the valuation of life.[10] Actually, there is a formal similarity between some arguments developed by Broome against the intuition of neutrality in the context of valuing the life of a person, and some of our arguments against the intuition of universality in the context of valuing a life-year.[11] Third, this paper is also related to the normative literature on fairness in the context of life and death.[12] The design of optimal policies is not independent from how lives are valued. This makes the distinction between the intuition of universality and the intuition of complementarity most relevant for policy purposes.

The rest of the paper is organized as follows. Section 1 presents three distinct accounts of the intuition of universality and explores their consequences on the structural form of value functions. Section 2 develops an approach based

8 See Ravallion, "Troubling Trade-offs in the Human Development Index"; and Anand, "Recasting Human Development Measures."

9 See Costa and Steckel, "Long-Term Trends in Health, Welfare, and Economic Growth in the United States." See also Nordhaus, "The Health of Nations"; Becker, Philipson, and Soares, "The Quantity and Quality of Life and the Evolution of World Inequality"; Fleurbaey and Blanchet, *Beyond GDP*.

10 See Broome, *Ethics out of Economics* and *Weighing Lives*.

11 See Broome, *Weighing Lives*.

12 See Fleurbaey and Ponthiere, "Prevention against Equality?" See also Adler, Hammitt, and Treich, "The Social Value of Mortality Risk Reduction."

on social preference orderings, reformulates the three accounts of the intuition of universality in that framework, and explores their consequences for the social ranking of distributions along the quantity and quality of life dimensions. Section 3 proposes simple ways to reconcile the abandonment of the intuition of universality with the priority given to individuals with low qualities of life in problems of life-years allocation. Concluding remarks are in section 4.

1. THREE ACCOUNTS OF THE INTUITION OF UNIVERSALITY: AN AXIOLOGICAL APPROACH

According to the intuition of universality, the valuation of a life-year is independent from the context under study, and thus independent from the associated quality of life. That intuition is in need of a more exact formulation, since the independence to which it refers may take various distinct forms. In this section, we propose three distinct accounts of the intuition of universality.

For that purpose, we adopt an axiological approach: our main object of study is a value function $V(.)$, which is defined on a life, which is itself defined as a vector $(a, b, \ldots, k)$ whose entries $a, b, \ldots, k$ correspond to the quality of each life-year. This section explores the formal constraints that the intuition of universality imposes on the structure of the value function $V(.)$. As such, this constitutes a first step toward a better understanding of the intuition of universality.

At the very outset, it should be stressed that a first, basic reading of the intuition of universality consists of stating that the value of a given life $(x, \ldots, x)$ in a country C should be exactly equal to the value of the same life in another country C'. That basic conception of the universality of the valuation of life is presented below, as the intuition U_0.

Intuition of Universality (U_0):

$$V((x, \ldots, x)_{country\ C}) = V((x, \ldots, x)_{country\ C'})$$

The formulation U_0 of the intuition of universality is quite intuitive: it makes a lot of sense to assume that the value of a life does not depend on the country where that life takes place. U_0 is thus a quite intuitive property that a value function $V(.)$ should satisfy.

Note, however, that the property U_0 is not really original, because it coincides merely with a standard anonymity condition. Anonymity being widely used in social evaluations—without an explicit reference to the intuition of universality—we believe that the property U_0 does not exhaust what the intuition of universality is about, and, in some sense, does not suffice to do justice to the

intuition of universality. Therefore we will, in the rest of this paper, take U_0 for granted—but as an anonymity condition—and explore the consequences of alternative formulations of the intuition of universality.

The intuition of universality can be formulated in terms of the variation of value induced by the addition of an extra life-year. One possible formulation of the intuition of universality consists of stating that the variation of value associated with the addition of an extra life-year should depend neither on the quality of the added life-year, nor on the quality of previous life-years, nor on the number of previous life periods (duration of initial lives). More formally, that formulation is:.

Intuition of Universality (U_1):

$$V((a, \ldots, g) + x) - V((a, \ldots, g)) = V((h, \ldots, m) + y) - V((h, \ldots, m))$$

The left-hand side of the above equality is the variation in value when a life-year with a quality x is added to a life of quality $(a, \ldots, g)$, whereas the right-hand side is the variation in value when a life-year with quality y is added to a life of quality $(h, \ldots, m)$.

The formulation U_1 of the intuition of universality states that the value of a life-year is universal, in the sense that it involves a *triple independence*: (i) independence with respect to the quality of the added life-year; (ii) independence with respect to the quality of previous life-years; (iii) independence with respect to the duration of initial lives to which a life-year is added (since the left-hand side and the right-hand side of the above equality may involve initial lives of unequal lengths). As such, the formulation U_1 captures a strong conception of universality.

To see how strong that conception of universality is, let us take the case of the addition of a life-year either in the United States, where average standard of living and life expectancy are high, or, alternatively, in Ghana, where standard of living and life-expectancy are lower. The condition U_1 states that the addition of a life-year in the United States (with US standard of living) has exactly the same value as the addition of a life-year in Ghana (with Ghana's standard of living). That condition states also that adding an extra life-year in Ghana with US standard of living has the same value as adding an extra life-year in the US with Ghana's standard of living.

Although it may seem appealing at first glance, the formulation U_1 of the intuition of universality has implications that are not attractive. In particular, the formulation U_1 conflicts with the monotonicity condition, stating that the value of a life increases with its quality. To see that conflict, let us take, here again, the example of the addition of a life-year in the United States or in Ghana, with

either US life quality (equal to 2) or Ghana's quality (equal to 1). Here are four alternative options:

A. Adding a life-year in Ghana with Ghana's standard of living
B. Adding a life-year in the US with the US's standard of living
C. Adding a life-year in Ghana with the US's standard of living
D. Adding a life-year in the US with Ghana's standard of living

The conception $U1$ of the intuition of universality implies that the value assigned to the addition of a life-year in case A must be equal to the value assigned to the addition of a life-year in case B, that is:

$$V((1,\ldots,1)+1)-V((1,\ldots,1))=V((2,\ldots,2)+2)-V((2,\ldots,2)).$$

However, by monotonicity, we have also that the value assigned to the addition of a life-year in case C exceeds the value assigned to the addition of a life-year in case A (where the added life-year has a lower quality, while everything else is left unchanged), that is:

$$V((1,\ldots,1)+2)-V((1,\ldots,1))>V((1,\ldots,1)+1)-V((1,\ldots,1)).$$

Still by monotonicity, we have also that the value assigned to the addition of a life-year in case B exceeds the value assigned to the addition of a life-year in case D (where, again, the added life-year has a lower quality, everything else remaining unchanged), that is:

$$V((2,\ldots,2)+2)-V((2,\ldots,2))>V((2,\ldots,2)+1)-V((2,\ldots,2)).$$

Given that $V((1,\ldots,1)+2)>V((1,\ldots,1)+1)$, and $V((2,\ldots,2)+2)>V((2,\ldots,2)+1)$, one obtains, from the first equality, the following inequality:

$$V((1,\ldots,1)+2)-V((1,\ldots,1)>V((2,\ldots,2)+1)-V((2,\ldots,2)).$$

This inequality means that adding a life-year with US standards in Ghana leads to a higher gain in value compared to adding a life-year with Ghana's standards in the US.

That inequality is in contradiction with the formulation $U1$ of the intuition of universality. According to that conception of universality, for sure a life-year with US standards in Ghana should be *equally good* to adding a life-year with Ghana's standards in the US. Hence, we reach here a contradiction, which implies that the formulation $U1$ of universality is not logically compatible with the monotonicity condition.

That proof by contradiction is formally close to the argument developed by Broome concerning the logical incompatibility of the intuition of neutrality for

the addition of a life with the principle of personal good.[13] Note that another way to prove that the property $U1$ is incompatible with monotonicity consists of examining the constraints that that formulation of the intuition of universality imposes on the structural form of the value function $V(.)$.

The property $U1$ implies that the variation in value associated with an extra life-year, i.e., $V((a, \ldots, g) + x) - V((a, \ldots, g))$, depends neither on the quality and quantity of other life-years, nor on the quality of the added life-year. The variation in value associated with the added life-years is thus a constant. Denoting that constant by c, one can deduce, by repeated substitutions, that:

$$V((a, \ldots, g) + x) = V((a, \ldots, g)) + c$$
$$[L \text{ years} + 1] \qquad [L \text{ years}]$$
$$\leftrightarrow V((a, \ldots, g) + x) = V((a, \ldots, f)) + c + c$$
$$[L \text{ years} + 1] \qquad [L - 1 \text{ years}]$$
$$\leftrightarrow V((a, \ldots, g) + x) = (L + 1)c$$
$$[L \text{ years} + 1].$$

We thus have that the formulation $U1$ of the intuition of universality implies that the value function $V(.)$ has a simple form: it is equal to the number of life-years multiplied by a constant. Hence, the formulation $U1$ of the intuition of universality implies that only the total lifetime matters, independent from the quality of life. Thus one can see, here again, that the formulation $U1$ of the intuition of universality is not compatible with the monotonicity property.

Given the natural appeal of the monotonicity condition, the logical incompatibility of the $U1$ formulation of the intuition of universality with monotonicity is quite problematic. This suggests that this formulation of the intuition of universality is too demanding, or too strong, which leads to incompatibilities with a property as simple as monotonicity. Therefore, in the rest of this section, we will propose to depart from the $U1$ formulation of the intuition of universality, and reformulate that intuition in different ways.

As we emphasized above, the formulation $U1$ of the intuition of universality was quite strong, since it involved a triple independence of the value of the extra life-year, with respect to the quality of the added life-year, with respect to the quality of previous life-years, and with respect to the number of previous life-years. In the remaining portion of this section, we will focus on weaker formula-

13 Broome, *Weighing Lives*, 238–39.

tions of the intuition of universality, which relaxes some of those independence requirements.

Intuition of Universality (U_2):

$$V((x, \ldots, x) + x) - V((x, \ldots, x)) = V((y, \ldots, y) + y) - V((y, \ldots, y))$$

The formulation U_2 of the intuition of universality states that the value of an extra life-year whose quality is equal to the quality of previous life-years should be universal, that is, independent from the quality of (previous and added) life-years, and also independent from the initial lengths of lives (since the left-hand side and right-hand side may involve initially lives of unequal lengths). Thus the formulation U_2 involves, as the formulation U_1, a triple independence of the value of an extra life-year.

However, the conception U_2 of universality is *weaker* than the conception U_1, since in U_2 the equality of the value of an extra life-year is conditional on its quality being equal to the quality of previous life-years, unlike what prevailed under U_1. Thus U_2 states the triple independence requirement only for lives of constant quality, not for lives of nonconstant quality. This limits the scope of the triple independence, and as such makes the conception U_2 of the intuition of universality weaker than the conception U_1.

In order to understand the implications of the formulation U_2 of the intuition of universality on the form of the value function $V(.)$, it is useful to notice that if the variation in value due to the addition of a life-year of constant quality is the same on both sides of the above equation, despite the fact that the (constant) quality of life-years is not the same on the left-hand side and right-hand side, and despite the fact that the lives being compared may initially differ in terms of sizes. As a consequence, the variation in value due to the addition of a life-year must be independent from the (constant) quality of life and from the length of the initial life. Therefore, this variation must be equal to a constant. Writing that constant with the letter c, we have,

$$V((x, \ldots, x) + x) - V((x, \ldots, x)) = c.$$

Hence, we obtain, by successive substitutions:

$$\begin{gathered} V((x, \ldots, x) + x) = V((x, \ldots, x)) + c \\ [L \text{ years} + 1] \qquad [L \text{ years}] \\ \leftrightarrow V((x, \ldots, x)) = V((x, \ldots, x)) + c + c \\ [L \text{ years} + 1] \quad [L - 1 \text{ years}] \end{gathered}$$

$$\leftrightarrow V((x, \ldots, x)) = V(x) + (L-1)c$$
$$[L \text{ years} + 1] \quad [1 \text{ year}].$$

The value function $V(x, \ldots, x)$ thus takes the form of a linear combination of the value of a single life-year $V(x)$, and of the length of that life.[14] Note that this formulation of the intuition of universality leads to a value function that is increasing in the quality of the added life-year (i.e., x), and, as such, satisfies the monotonicity condition, unlike the (stronger) formulation $U1$ of the intuition of universality.

Interestingly, the United Nations HDI, in its initial form, is additive in an index of life expectancy achievements, and in an index of GDP per capita achievements (and also of an index of education achievements).[15] Hence, the initial HDI has a functional form that is compatible with the conception $U2$ of the intuition of universality, in the sense that it values the addition of a life-year with constant quality independent from the (constant) quality, and independent from the length of initial lives. The standard HDI thus captures the formulation $U2$ of the intuition of universality.

Whereas the above discussion assumes a constant quality of life profile, it may be useful to generalize that discussion to the case where the lifetime quality profile is not constant, and is, for example, equal to $(a, b, \ldots, h)$. One could then consider a value function $V(.)$ that is, as above, additive, but takes the (more general) form: $V(X) + (L-1)c$, where X denotes the *generalized average* quality of lifetime.

If the generalized average of the quality of life X is equal to the quality of the added life-year (i.e., x), then that value function satisfies the property $U2$. Note also that, provided the generalized average of the quality of life X is increasing the quality of the added life-year (i.e., x), this value function also satisfies the monotonicity condition.[16] Moreover, the value function $V(.)$ allows for a certain degree of complementarity between lifetime quantity and lifetime quality, something that was not possible under the formulation $U1$ of the intuition of universality. Clearly, adding a life-year with a quality above the average quality (i.e., $x > X$) is here regarded as good (i.e., increases value), whereas adding a life-year with a quality inferior to the average quality (i.e., $x < X$) is bad (i.e., reduces value). Finally, adding a life-year with a quality exactly equal to the average quality (i.e., $x = X$) is neutral.

Those positive results hold thanks to the fact that $U2$ is a weaker formulation

14 Indeed, it is possible to add an extra term c on the right-hand side while still respecting the condition $U2$, which leads to $V(x, \ldots, x) = V(x) + cL$.

15 See United Nations Development Programme, *Human Development Report 1990*.

16 Note that the generalized average quality is not necessarily increasing in x. One could have $X = \min\{a, b, \ldots, h, x\} = a \neq x$, or $X = \max\{a, b, \ldots, h, x\} = d \neq x$.

of the intuition of universality than U_1. It is thus compatible with monotonicity and also with some degree of complementarity, but at the cost of weakening the requirement of universality (with respect to formulation U_1).

Let us now consider an alternative formulation of the intuition of universality, which imposes not an independence of the value of the extra life-year with respect to the quality of the added life-year, but an independence only with respect to the quality of previous life-years (which may not be constant, unlike under U_2), and with respect to the number of previous life-years. This is the formulation U_3 of the intuition of universality.

Intuition of Universality (U_3):

$$V((a, b, \ldots, g) + z) - V((a, b, \ldots, g)) = V((h, i, \ldots, m) + z) - V((h, i, \ldots, m))$$

The formulation U_3 of the intuition of universality involves a double independence of the valuation of an extra life-year: (i) independence with respect to the quality of previous life-years, and (ii) independence with respect to the quantity of previous life-years (since the number of previous life-years involved on the left-hand side and right-hand side may differ).

To understand the implications of the formulation U_3 of the intuition of universality on the structure of the value function $V(.)$, let us notice that, since the variation of value due to the addition of a life-year does not depend on the quality and quantity of previous life-years, it can only depend on the quality of the added life-year, that is, on z. Hence, we have:

$$V((a, b, \ldots, g) + z) - V((a, b, \ldots, g)) = v_L(z),$$

where $v_L(z)$ is a function of z.

Hence we obtain, by repeated substitution:

$$\begin{aligned} & V((a, b, \ldots, g) + z) = V((a, b, \ldots, g)) + v_L(z) \\ & \quad [L \text{ years} + 1] \qquad [L \text{ years}] \\ \leftrightarrow\ & V((a, b, \ldots, g) + z) = V((a, b, \ldots)) + v_{L-1}(g) + v_L(z) \\ & \quad [L \text{ years} + 1] \qquad [L - 1 \text{ years}] \\ \leftrightarrow\ & V((a, b, \ldots, g) + z) = V(a) + v_2(b) + \ldots + v_{L-1}(g) + v_L(z) \\ & \quad [L \text{ years} + 1] \qquad [1 \text{ year}]. \end{aligned}$$

where $v_2(.), \ldots, v_L(.)$ are functions of the quality of each life-year.

The formulation U_3 of the intuition of universality has thus a precise implication on the structure of the value function $V(.)$. Actually, it imposes that the

value function $V(.)$ is a sum of the transformed qualities of all life-years involved, with a number of terms equal to the number of life-years in the life under study.

The functional form for $V(.)$ implied by the property $U3$ satisfies the monotonicity condition, unlike the property $U1$. Moreover, it allows also for some degree of complementarity between the quality and quantity of life. As for the formulation $U2$ of the intuition of universality, those positive results arise thanks to the fact that the property $U3$ is a weaker formulation of the intuition of universality. That more limited universality requirement allows for some compatibility with monotonicity and with the intuition of complementarity, but at the cost of weakening the universality requirement.

In sum, this section showed that the intuition of universality for the valuation of a life-year can be formulated in quite distinct ways, which all have their particular implications for the structure of the value function that measures the value of a life as a whole. Note, however, that although this section allowed us to provide precise accounts of the intuition of universality, and to explore the consequences of those formulations on the structure of value functions, this section had little to say, in general, about how priorities should be given when allocating life-years within the population. Actually, the variations in value associated with the addition of a life-year do not have direct implications in terms of priority, except if one adopts the social objective of maximizing the sum, across all individuals, of values $V(.)$.

If one adopts that particular social objective, then an interesting thing to notice is that the formulations $U1$, $U2$, and $U3$ of the intuition of universality share an important direct implication in terms of social priority: these all imply *social indifference* regarding how life-years are allocated within the population. Thus those three conceptions of the intuition of universality lead the social evaluator to be indifferent with respect to how those life-years are distributed. That result is not particularly appealing: from an egalitarian perspective, one may prefer to give social priority to individuals whose lives are of low quality or of limited quantity. This view is clearly not compatible with the formulations $U1$, $U2$, and $U3$ of the intuition of universality, at least if the goal is to maximize the sum of values $V(.)$ across individuals.

It should be stressed, however, that there is no obvious reason why the social evaluator should take, as an objective, the maximization of the sum of individual values $V(.)$. Many other social goals exist, and in those cases the above formulations of the intuition of universality do not have direct implications in terms of priority. The goal of the next section is to develop an alternative approach, in terms of social preference orderings, in order to explore, under more general social objectives, the implications of the three conceptions of the intuition of universality developed above for the allocation of life-years within a population.

2. THREE ACCOUNTS OF THE INTUITION OF UNIVERSALITY: SOCIAL PREFERENCE APPROACH

Let us now examine the implications of the intuition of universality in terms of priority in the context of a problem of life-years allocation. For that purpose, let us define an allocation as a vector of quality of life for each individual in the population, whose size is supposed to be constant and equal to N. Formally, an allocation can be written as a vector $q = (q_i)_{i=1,\ldots,N}$, where $q_i = (q_{i1}, \ldots, q_{ili})$ is the life of individual i, who enjoys a life of length l_i. We denote by Q the set of all such allocations. We denote by Q^c the subset of Q that includes all allocations with constant quality along the life. Regarding individual longevities, we will define by l the vector of individual durations of life (number of life-years). We have that $l = (l_i)_{i=1,\ldots,N}$.

Let us now reformulate the three conceptions of the intuition of universality studied in section 1 in terms of their consequences concerning the social ranking of allocations. In this section, we denote that social preordering as $\leq_s$. That social preference relation is assumed to be reflexive, transitive, and complete. As usual, strict social preference is denoted by $<_s$, whereas social indifference is written as $\sim_s$.

Throughout this section, the intuition of universality will be formulated in terms of whether adding an extra life-year to a person i is equivalent to adding an extra life-year to a person j, which is, from a formal perspective, equivalent to stating that transferring a life-year from individual j to individual i leads to social indifference. Thus, even if the formulations of the intuition of universality developed below look like properties about transfers of life-years across individuals, these are only a formal way to formulate conditions of social indifference about who receives the extra life-year.

In terms of the preordering on allocations, the formulation $U1$ of the intuition of universality states that a change in who receives an additional life-year, everything else being left unchanged, leads to an allocation that is regarded, from a social perspective, as equally good as the initial allocation, independently from the quality of the added life-year, and independently from the quantity and quality of previous life-years.

> *Intuition of Universality* ($U1$): For all q, q' in Q, if $q_{it} = q_{it}'$ for all i, t in $\{1, 2, \ldots, \min(l_i, l_i')\}$, and if there exists i, j such that: $l_i' = l_i + 1$ and $l_j' = l_j - 1$ and for all $k \neq i, j$, $l_k' = l_k$, then $q \sim_s q'$.

From the perspective of $U1$, it does not matter whether an additional life-year is given to a person with a more or less long life, or with a life of more or less high quality: changing the recipient of the extra life-year leads neither to a social im-

provement, nor to a social worsening, but is just neutral. As such, the conception $U1$ captures some idea of universality in the valuation of life-years. It states that one is socially indifferent between allocating an extra life-year to a given life or to another life.

Note that this social indifference associated with who receives the extra life-year amounts to assuming that the social valuation of life-years satisfies a triple independence: (i) independence with respect to the quality of the added life-year; (ii) independence with respect to the quality of previous life-years; and (iii) independence with respect to the quantity of previous life-years.

What are the implications of the formulation $U1$ of the intuition of universality for the allocation of life-years within a population?

Although the formulation $U1$ may seem intuitive at first glance, it has implications that are not so attractive regarding the allocation of life-years. In particular, it is incompatible with a basic monotonicity property. The monotonicity property can be stated as follows.

> *Monotonicity*: For all q, q' in $\mathcal{Q}$, if $l = l'$, if $q_{it}' > q_{it}$ for some i, t in $\{1, \ldots, l_i\}$ and $q_{it}' = q_{it}$ for all other i, t in $\{1, \ldots, l_i\}$, then $q' >_s q$.

The monotonicity property is quite weak: it states that if some allocation q' involves a higher quality of life-years for some individuals in comparison to the allocation q, everything else remaining the same in q' and q, then the allocation q' is, from a social perspective, strictly better than the allocation q.

In order to see why the formulation $U1$ of the intuition of universality is logically incompatible with the monotonicity property, let us consider two allocations q and q' satisfying the conditions described in $U1$, and let us add a third allocation, denoted by q'', which is the same as the allocation q', except that it involves a strictly higher quality of life for the extra life-year enjoyed by individual i, that is, that $q_{ili}'' > q_{ili}'$.

It is easy to see that, when comparing allocations q and q'', the formulation $U1$ of the intuition of universality implies that there must be social indifference between q and q'', for the same reasons as there is social indifference between q and q'. We thus have, by property $U1$, that:

$$q \sim_s q'' \text{ and } q \sim_s q'.$$

This implies, by transitivity, that:

$$q' \sim_s q''.$$

However, the monotonicity property requires that:

$$q'' >_s q'.$$

Thus we reach here a contradiction. That contradiction implies that the social ranking of allocations cannot satisfy both the formulation $U1$ of the intuition of universality and the monotonicity property. A choice is to be made between those properties.

Given the natural appeal of the monotonicity condition for the social ranking of allocations, this negative result supports giving up the intuition of universality, at least under its $U1$ formulation. Actually, if being universalist regarding the valuation of life-years implies violating monotonicity, and thus being socially indifferent between allocations that are clearly not equivalent at all, then the attractiveness of such a universalism can be questioned.

It should be stressed, however, that the above negative result only concerns the formulation $U1$ of the intuition of universality. As such, this cannot be generalized to all conceptions of the intuition of universality.

Let us now consider the implications of the second formulation of the intuition of universality on the allocation of life-years. Translated in terms of requirements regarding the social preference relation, the formulation $U2$ can be written as:

> *Intuition of Universality* ($U2$): For all q, q' in Q^C, if $q_{i1} = q_{i1}'$ for all i, and if there exists i, j such that: $l_i' = l_i + 1$ and $l_j' = l_j - 1$ and for all $k \neq i, j$, $l_k' = l_k$, then $q \sim_s q'$.

Thus $U2$ states that, when comparing allocations with constant quality among lives, a change in the recipient of an extra life-year from person j to person i (while keeping everything else unchanged) leads to social indifference. Whatever the durations of life for the individuals i and j, and whatever the qualities of their previous life-years, whether it is person i or person j that receives the extra life-year is neutral. Note that this conception of universality is weaker than conception $U1$, because it is here restricted to the subset of allocations in which the quality of life is constant along a given life.

In order to explore the implications of the formulation $U2$ of the intuition of universality in terms of priority, a first important step consists of examining the constraints that $U2$ imposes on the form of a social-welfare function. Actually, as shown in the appendix:

> *Characterization Theorem (Formulation U2 of the Intuition of Universality)*: A social-welfare function $W(.)$ satisfies the formulation $U2$ of the intuition of universality if and only if it takes the following form: $W(q) = F(q_i,$

$q_j, \ldots, q_N, \Sigma l_i)$, where q_i denotes the (constant) quality of life enjoyed at all life-periods by individual i under allocation q.

What is stated here is a representation result that takes the form of a logical equivalence: any social-welfare function that satisfies the property $U2$ must have that particular form, and, also, any social-welfare function that satisfies that form must also satisfy the formulation $U2$ of the intuition of universality. Interestingly, the form taken by the social-welfare function is simple: it is a function of the (constant) qualities associated with the life-years of all individuals, and also a function of the total lifetime of the population.[17]

An important corollary of this representation result is that, under the conception $U2$ of universalism, the particular distribution of life-years within the population does not matter; the only thing that matters concerning lifetime is the *total amount of life-years* that are lived. Whether the lifetime is shared more or less equally within the population does not matter.

That corollary is particularly counterintuitive. When considering the allocation of life-years within a population, a social planner may prefer, on the grounds of social justice, that life-years are distributed more equally across individual lives. Such an egalitarian perspective is incompatible with the formulation $U2$ of the intuition of universality. Being universalist under the $U2$ conception implies being socially indifferent between allocations that keep the total number of life-years constant, independently from how those life-years are distributed in the population.

To put it in different terms, the formulation $U2$ of the intuition of universality leads to being indifferent with respect to the distribution of life-years across individuals, and, as such, this is incompatible with the idea of giving priority to the poor, who can be here represented as individuals with shorter lives and lower qualities of life. The formulation $U2$ of the intuition of universality prevents giving priority to those disadvantaged individuals.

Note that this result only presupposes the formulation $U2$ of the intuition of universality, and is not based on a particular assumption concerning the way in which the social-welfare function aggregates value functions $V(.)$. Clearly, if the social ordering of allocations were based on the sum of individual value functions $V(.)$, as in section 1, one would also obtain social indifference with respect to the distribution of life-years within the population. This section provides a more general argument, according to which the formulation $U2$ of the intuition of universality leads inevitably to social indifference with respect to the distri-

17 Given that we consider populations of constant sizes, the social-welfare function $W(.)$ can also be regarded as a function of the average lifetime of the population.

bution of life-years, whatever the precise way (additive or not) in which value functions $V(.)$ enter the social objective.

That corollary of the formulation $U2$ of the intuition of universality tends to question the attractiveness of universality when formulated in that particular way. If being universalist implies being indifferent with respect to inequalities in length of life, then such a universalist perspective looks far from attractive. Ideally, we would like universalism to lead toward priority given to the disadvantaged, and, hence, toward more equality, and not to lead to indifference toward more inequality. We reach, here again, a negative result, but this negative result is relative to a particular formulation of the intuition of universality.[18]

Let us now turn to the third conception of universality developed in section 1. When reformulated in terms of its implications on the social preference ordering over allocations, the formulation U3 of the intuition of universality is defined as follows.

Intuition of Universality $(U3)$: For all q in Q, for all z in R_+, for all i, j, we have:

$$(\dots(q_{i1}, \dots, q_{ili}, z)\dots(\dots(q_{j1}, \dots, q_{jlj})\dots)$$
$$\sim_s (\dots(q_{i1}, \dots, q_{ili})\dots(\dots(q_{j1}, \dots, q_{jlj}, z)\dots).$$

The property $U3$ states that changing the recipient of an extra life-year with quality z leads to social indifference, whatever the quality and quantity of life-years lived by the possible recipients. As such, it captures some intuition of universality, in the sense that the social evaluator is indifferent between giving an extra life-year to one person or to another, whatever the lives of those persons are.

Although that conception of universality may seem appealing, it faces the same problem as the conception $U2$ studied above: by valuing transfers of life-years indifferently from the lives of the persons who are involved in the transfer (in terms of their quantity and quality), the conception $U3$ of universality goes against the idea of giving priority to the disadvantaged.

To see this, let us take a simple two-person example, involving persons i and j. The initial allocation is:

$$((q_{i1}, q_{i2}, \dots, q_{ili}), (q_{j1}, q_{j2}, \dots, q_{jlj})).$$

The formulation $U3$ of the intuition of universality states that changing the recip-

18 Note that this negative result is reached while assuming a representativity of the social preference ordering by means of a social-welfare function $W(.)$. However, as shown in the appendix, our result is actually more general, and does not necessarily require assuming the existence of such a representation.

ient of a life-year, from, let us say, person j to person i, leads to social indifference. If the last life-year of person j is reallocated to person i (with the associated quality q_{jlj}), we thus have:

$$((q_{i1}, q_{i2}, \ldots, q_{ili}), (q_{j1}, q_{j2}, \ldots, q_{jlj})) \sim_s ((q_{i1}, q_{i2}, \ldots, q_{ili}, q_{jlj}), (q_{j1}, q_{j2}, \ldots, q_{jlj-1})).$$

Then, by repeating reallocations of life-years successively, from person j to person i, one finally obtains:

$$((q_{i1}, q_{i2}, \ldots, q_{ili}), (q_{j1}, q_{j2}, \ldots, q_{jlj})) \sim_s ((q_{i1}, q_{i2}, \ldots, q_{ili}, q_{jlj}, \ldots, q_{j2}), (q_{j1})).$$

Thus property U_3 leads to social indifference between two allocations that are extremely different: whereas in the initial allocation, the lifetime is divided between persons i and j, in the final allocation, almost the entire lifetime is concentrated on person i, whereas only a single life-year remains for person j. That highly unequal distribution of lifetime does not seem to be as socially desirable as the initial allocation, but this is what formulation U_3 of the intuition of universality implies. Repeated use of the universality property U_3 leads to social indifference between allocations that are characterized by quite different degrees of inequality in the distribution of lifetime among persons.

From the perspective of social justice, one would prefer, on the contrary, to give priority in the allocation of life-years to disadvantaged individuals, who have either shorter lives or lives of worse quality. The intuition of universality is hardly compatible with giving priority to the disadvantaged. On the contrary, it leads to social indifference with respect to how life-years are allocated between persons. Here again, as for the formulation U_2, the intuition of universality goes against this ideal of giving priority to the disadvantaged.

Note also that, as for the conception U_2, the argument provided here does not rely on a particular functional form for the social objective. Obviously, if the social goal is, as in section 1, to maximize the sum of value functions $V(.)$, then we would also obtain social indifference with respect to who receives the extra life-year. But the argument developed here is more general, since this does not presuppose any particular social objective—that is, the social ordering does not need to be based on the mere sum of value functions $V(.)$. Thus, we reach a robust result on the conflict of conception U_3 of universality with giving priority to the disadvantaged.

In sum, this section leads to quite negative results concerning the implications of the intuition of universality. We showed that either the intuition of universality is incompatible with the monotonicity property (conception U_1), or leads to social indifference with respect to how life-years are allocated within the

population, which goes against the ideal of giving priority to the disadvantaged (conceptions U_2 and U_3).

Whereas this section reached some negative results concerning the intuition of universality, one may wonder whether abandoning that intuition in favor of the intuition of complementarity would allow obtaining more appealing implications. In particular, one may be curious to see whether adopting the intuition of complementarity would allow better meeting of the ideal of giving priority to the disadvantaged. That question is explored in the next section.

3. THE INTUITION OF COMPLEMENTARITY AND PRIORITY TO THE WORST-OFF

Under the intuition of complementarity, the value of a life-year depends on what that life-year allows—that is, on the quality of that life-year. At first glance, one may believe that the intuition of complementarity, by leading to assigning a higher value to life-years characterized by a higher quality (unlike the intuition of universality), could favor the allocation of life-years toward more life-years given to individuals who enjoy a high quality of life.

But that belief is actually wrong: when allocating life-years, the valorization of those years is only *one* aspect of the problem. Another crucial aspect concerns the priority that the social evaluator assigns to the well-being levels of the different individuals, and, in particular, their aversion to inequality.[19] When the aversion to inequality is large, it can offset the valorization dimension, and lead to assigning more life-years to individuals with low life quality. It is actually quite simple to combine the intuition of complementarity with giving priority to the disadvantaged.

To see that, let us assume that the value of an individual life takes a standard, time-additive form, that is:

$$V(q_i) = \sum_{t=i}^{l_i} u_i(q_{it}),$$

where $u_i(q_{it})$ represents the temporal utility associated with the life-year t for individual i.

For the sake of simplicity, we will assume here that lifetime is continuous rather than discrete, and thus consider the equivalent form in continuous time:

$$V(q_i) = \int_t^{l_i} u_i(q_{it})\, dt.$$

The social-welfare function takes the general form:

19 On the assignment of social priority, see Adler, *Well-Being and Fair Distribution* and *Measuring Social Welfare*.

$$W(q) = W(V(q_1), \ldots, V(q_N)).$$

Within that framework, the marginal social-welfare from increasing the duration of the life of individual i is given by the derivative:

$$\partial W/\partial l_i = (\partial W/\partial V(q_i))\, u_i(q_{il_i}).$$

The left-hand side of that equation is the variation in social welfare associated to a minor change in the duration of the life of individual *i*. This variation is equal to the product of two factors.

First, it depends on the degree of priority of the individual from a social perspective, which is captured by the factor $(\partial W/\partial V(q_i))$. This degree of priority clearly depends on the degree of inequality aversion exhibited by the social-welfare function. If person *i* is particularly disadvantaged, an inequality-averse social planner assigns a high weight to improving the well-being of that person. This first effect is the social-weighting effect.

Second, the marginal social welfare associated with a change in duration of the life of individual *i* depends also on the value of this extra life-year for the individual, based on the quantity and quality of their past life, and also based on the quality of the extra life-year itself. One can expect that a life profile with a higher quality of life will generally imply a higher value for an increase in the duration of life. That second effect is the individual valuation effect (which may depend on individual subjective preferences or some other objective approach to the valuation of individual lives).

In the case of increasing the lifetime of an individual whose life has low quality, the social-weighting effect and the individual-valuation effect go in opposite directions when the social planner is inequality averse. In that case, the social-weighting effect is strong, while the individual-valuation effect is low. On the contrary, when considering the marginal social welfare from increasing the duration of life of a person whose life has high quality, the opposite arises: the individual-valuation effect is large, while the social-weighting effect is low.

At the end of the day, whether a higher marginal social value is assigned to increasing the length of life of the person with a low or high life quality depends on the degree of inequality aversion exhibited by the social-welfare function, and also on individuals' valuations of life. It is quite possible that a higher marginal social value is assigned to increasing the duration of life of a person with a low life quality, despite the individual valuation effect. This is definitely the case when the social-welfare function exhibits a high degree of inequality aversion.

To show this, let us take a simple analytical example, where the function $u_i(q_{it})$ takes the following form:

$$u_i(q_{it}) = [{q_{it}}^{1-a} - {q_0}^{1-a}]/(1-a).$$

Moreover, let us suppose that the social-welfare function takes a standard Atkinson form:

$$W = \Sigma(V(q_i))^{1-e}/(1-e),$$

where the parameter e captures the sensitivity to inequalities in well-being across individuals.

In that analytical example, and supposing a constant quality of life $q_{it} = q^i$, the marginal social welfare from increasing the duration of life of person *i* is equal to:

$$\partial W/\partial l_i = (l_i u_i(q^i))^{-e} u_i(q^i).$$

Note that, when the ethical parameter *e* equals 0, the marginal social welfare from increasing the duration of life of individual *i* is equal to $u_i(q^i)$, and, hence, is increasing with the quality of life enjoyed by person *i*. In that case, the social-weighting effect is dominated by the individual-valuation effect, and so a larger priority is given to individuals with a higher quality of life.

But that is not the only possible case. Actually, under a large interval of values for the ethical parameter *e*, the opposite will take place, and the social-weighting effect will dominate the individual-valuation effect, leading to priority to the disadvantaged individuals.

It is straightforward to see that, when the ethical parameter *e* equals 1, the marginal social welfare from increasing the duration of life of individual *i* is equal to merely $1/l_i$—that is, to the inverse of person *i*'s duration of life. Hence, in that case, a higher priority will be given to individuals with a short life, and a lower priority to individuals with a longer life.

Alternatively, when *e* is superior to 1, an even larger priority is given to the disadvantaged individuals, since the marginal social welfare from increasing the length of life of a person *i* is then not only decreasing with the duration of life of that person, but also decreasing with the quality of life. Thus priority is here given to individuals with shorter lives and lives of lower quality.

Those examples suffice to illustrate that it is possible—and actually quite easy—to accommodate the intuition of complementarity with the ideal of giving priority to the disadvantaged. The intuition is that the marginal social value of increasing the duration of life of a person depends not only on the quality of that life (through the individual-valuation effect), but also on the weight that is given to improving the situation of that person in the social-welfare function (the social-weighting effect). When the latter dominates the former, priority is given to individuals with a shorter life and with a lower quality of life.

4. CONCLUDING REMARKS

Given that various policies—health policies, safety policies, development policies—influence mortality, and, hence, individual lifetimes, the valuation of life-years has become a necessary stage in the design of optimal policies. The definition of optimal policies in life-affecting domains requires governments to be able not only to weight life-years against resources, but, also, life-years enjoyed by some persons against life-years enjoyed by other persons. Moreover, at the descriptive level, the measurement of economic development requires the ability to weight achievements in terms of longevity in comparison to achievements on other dimensions of life, and, also, to make longevity achievements in some countries comparable with longevity achievements in other countries.

When considering the valuation of life-years, two basic intuitions arise: on the one hand, the intuition of universality, according to which the value of a life-year should be universal, and, hence, independent from the duration and the quality of lives considered, and, on the other hand, the intuition of complementarity, according to which the value of a life-year should depend on what that life-year allows for, and, hence, on its quality.

Those two intuitions are plausible, but hardly compatible: the intuition of universality requires that the value of a life-year is universal, and, hence, does not depend on its quality, which goes against the intuition of complementarity, which makes the valuation of a life-year dependent on its quality. Thus a choice is to be made between those two intuitions concerning the valuation of life-years.

In order to cast original light on that ethical dilemma, this paper proposes to provide several distinct accounts of the intuition of universality, and to explore their logical implications in terms of the valuation of life-years, and, also, in terms of the priority to be given to the disadvantaged when considering the allocation of lifetime within a population.

Our results suggest that the intuition of universality, whatever the precise formulation considered, leads to implications that are far from appealing. Our accounts of the intuition of universality lead either to a conflict with a basic principle of monotonicity (i.e., the conception U_1), or lead to a conflict with giving priority to the disadvantaged (i.e., conceptions U_2 and U_3). Those conflicts are quite problematic: imposing a universal valuation of life-years would lead to social indifference with respect to the distribution of lifetime within the population. Such social indifference would go against the ideal of equality, and, as such, is counterintuitive and hard to justify.

On the contrary, the intuition of complementarity can be compatible with the idea of giving priority to the disadvantaged, and, as such, does not imply a

social indifference with respect to how life-years are distributed within the population, unlike the intuition of universality. The underlying intuition is that the dependence of the valuation of a life-year on quality of life is only one aspect of the social valuation of life-years, which depends also on how individual interests are weighted in the social-welfare function. It is thus possible, when the social-welfare function exhibits a sufficiently high degree of inequality aversion, to conciliate the intuition of complementarity with the ideal of giving priority to individuals with low qualities of life.

All in all, this paper suggests that the intuition of universality, although it may seem appealing at first glance, leads, at the end of the day, to the opposite of what it aims at: by valuing all life-years in a uniform way, the intuition of universality is not compatible with giving priority to the disadvantaged, and, hence, tends to play against equality. On the contrary, the intuition of complementarity can be made compatible with the ideal of giving priority to the disadvantaged, and, hence, is more compatible with equality.

Paris School of Economics
marc.fleurbaey@psemail.eu

Université Catholique de Louvain
gregory.ponthiere@uclouvain.be

APPENDIX

The characterization result takes the form of an equivalence between a social-welfare function $W(q)$ satisfying property $U2$ and a social-welfare function taking the form $W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$.

To prove that equivalence result, we proceed in two steps.

Let us first prove that a social-welfare function taking the form $W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$ satisfies the property $U2$.

To see this, let us take a three-person case, with (constant) qualities of life a, b, and c, and durations of life m, n, and o. Let us denote by $a.m$ a life of m years with constant quality a. The allocation q is thus written $(a.m, b.n, c.o)$. Let us now compare that allocation with another allocation, q', where a life-year is transferred from the second person to the first person. The allocation q' is thus written as $(a.(m+1), b.(n-1), c.o)$.

The property $U2$ requires that there is social indifference between allocations q and q', that is: $W(a.m, b.n, c.o) = W(a(m+1), b(n-1), c.o)$.

It is easy to see that this equality is satisfied by any function taking the form

$W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$. Indeed, in our three-person case, the function takes the form: $W(a.m, b.n, c.o) = F(a, b, c, m + n + o)$. It does satisfy the equality mentioned above, since transfer of a life-year maintains the total number of life-years unchanged. Indeed, we have:

$$W(q) = F(a, b, c, m + n + o) = F(a, b, c, m + 1 + n - 1 + o) = W(q')$$

as required by property *U*2.

The same argument could be formulated for any case with $N > 3$, with any transfer of life-years. Thus we have that any social-welfare function taking the form $W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$ satisfies the property *U*2.

Let us now prove that any social-welfare function satisfying the property *U*2 takes the form $W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$. To prove this, let us turn back to our three-person case. We have, by repeated use of the property *U*2:

$$\begin{aligned} W(a.m, b.n, c.o) &= W(a.(m + 1), b.(n - 1), c.o) \\ &= W(a.(m + 2), b.(n - 1), c.(o - 1)) \\ &= W(a.(m + 3), b.(n - 2), c.(o - 1)) \\ &= \ldots \\ &= W(a.(m + n - 1 + o - 1), b.1, c.1) \\ &= W(a.(l_1 + l_2 + l_3 - (3 - 1)), b.1, c.1). \end{aligned}$$

Since the population size $N = 3$ is a constant, we thus have:

$$W(a.m, b.n, c.o) = F(a, b, c, l_1 + l_2 + l_3).$$

That is, the social-welfare function takes the form $W(q) = F(q_i, q_j, \ldots, q_N, \Sigma l_i)$. A similar proof could be provided for any $N > 3$.

Finally, it should be stressed that, whereas the above proof assumes the existence of a representation of the social ordering $\leq_s$, such an assumption is not necessary for the purpose at hand. Actually, it can be shown that there exists another preorder $\leq^*$ defined on $((q_i)_i, \Sigma l_i)$, which is such that:

$$((qi)i, \Sigma li) \geq^* ((qi')i, \Sigma li')$$

$$\leftrightarrow$$

$$(q1.(\Sigma li - N + 1), qi.1) \geq (q1'.(\Sigma li' - N + 1), qi'.1).$$

REFERENCES

Adler, Matthew. *Measuring Social Welfare: An Introduction*. New York: Oxford University Press, 2019.

———. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. New York: Oxford University Press, 2012.

Adler, Matthew, James Hammitt, and Nicolas Treich. "The Social Value of Mortality Risk Reduction: VSL versus the Social Welfare Function Approach." *Journal of Health Economics* 35 (May 2014): 82–93.

Anand, Sudhir. "Recasting Human Development Measures." UNDP Human Development Report Discussion Paper (March 2018): 1–55.

Becker, Gary S., Thomas J. Philipson, and Rodrigo R. Soares. "The Quantity and Quality of Life and the Evolution of World Inequality." *American Economic Review* 95, no. 1 (March 2005): 277–91.

Broome, John. *Ethics out of Economics*. Cambridge: Cambridge University Press, 1999.

———. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Costa, Dora, and Richard Steckel. "Long-Term Trends in Health, Welfare, and Economic Growth in the United States." In *Health and Welfare during the Industrialization*, edited by Richard Steckel and Rodrik Floud, 47–90, Chicago: University of Chicago Press, 1997.

Fleurbaey, Marc, and Didier Blanchet. *Beyond GDP: Measuring Welfare and Assessing Sustainability*. Oxford: Oxford University Press, 2013.

Fleurbaey, Marc, and Gregory Ponthiere. "Prevention against Equality?" *Journal of Public Economics* 103 (July 2013): 68–84.

Jones-Lee, Michael. *The Economics of Safety and Physical Risk*. London: Blackwell, 1989.

Miller, Tim. "Variations between Countries in Values of Statistical Life." *Journal of Transportation Economics and Policy* 34, no. 2 (May 2000): 169–88.

Nordhaus, William. "The Health of Nations: The Contributions of Improved Health to Living Standards." In *Measuring Gains from Medical Research: An Economic Approach*, edited by Kevin Murphy and Robert Topel, 9–40, Chicago: University of Chicago Press, 2003.

Ravallion, Martin. "Troubling Trade-offs in the Human Development Index." *Journal of Development Economics* 99, no. 2 (November 2012): 201–9.

Sen, Amartya. "Mortality as an Indicator of Economic Success and Failure." *Economic Journal* 108, no. 446 (January 1998): 1–25.

———. "On the Development of Basic Income Indicators to Supplement GNP Measures." *United Nations Bulletin for Asia and the Far East* 24 (1973): 1–11.

United Nations Development Programme. *Human Development Report 1990*. New York: Oxford University Press, 1990.

———. *Human Development Report 2010. The Real Wealth of Nations: Pathways to Human Development*. Houndmills: Palgrave Macmillan, 2010.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1593

CRITICAL LEVELS, CRITICAL RANGES, AND IMPRECISE EXCHANGE RATES IN POPULATION AXIOLOGY

Elliott Thornley

How do we determine whether one population is at least as good as another? Here is one easy answer. We use a number to represent each person's welfare—how good their life is for them—with the size of the number proportional to how good their life is. Positive numbers represent good lives, negative numbers represent bad lives, and zero represents lives that are neither good nor bad. We then sum these numbers to get the value of each population. A population X is at least as good as a population Y iff the value of X is at least as great as the value of Y. A theory of how populations relate with respect to goodness is called a *population axiology*, and we can call this population axiology the *Total View*.

The Total View implies that we can improve populations by adding lives that are barely worth living, and some find this implication distasteful. We can avoid this implication by first subtracting some positive constant from the number representing a person's welfare and then summing the results. Call these population axiologies *critical-level views*.

Critical-level views cannot account for two intuitions that many people find appealing. The first is that there is a *range* of welfare levels such that adding lives at these levels makes a population neither better nor worse. The second is that populations of different sizes may be *incommensurable*, so that neither population is better than the other and yet nor are they equally good. In that case, we might prefer to subtract a range of positive constants from the number representing a person's welfare and then calculate the value of a population relative to each constant within the range. We can then claim that X is at least as good as Y iff the value of X is at least as great as the value of Y relative to each constant within the range. If neither X nor Y is at least as good as the other, they are incommensurable. Call these population axiologies *critical-range views*.

Critical-level and critical-range views fall within the more general class of *critical-set views*. I offer a characterization and taxonomy of these views below,

along with six objections that tell against various views in this taxonomy. Some views imply repugnant or sadistic conclusions. Other views make neutrality implausibly greedy. Each view implies at least one implausible discontinuity, and no view can account for the incommensurability between lives and between same-size populations without extra theoretical resources.

I then offer a view that retains much of the appeal of critical-set views while avoiding many of the aforementioned pitfalls. The *Imprecise Exchange Rates View* has its start in the observation that there are often no precise truths about whether it is worth undergoing some bad for the sake of some good. It makes sense of this observation by claiming that various *exchange rates* between goods and bads are imprecise. This imprecision renders certain combinations of goods and bads incommensurable with other combinations. The view thus provides a natural explanation of incommensurability between lives and between same-size populations, avoids all forms of sadism along with the most concerning instances of repugnance and greediness, and has many other advantages besides.

I characterize and taxonomize critical-set views in section 1 and object to them in section 2. I introduce the Imprecise Exchange Rates View in section 3, canvas its advantages in section 4, and address some objections in section 5. I sum up in section 6.

1. CRITICAL-SET VIEWS

Foundational to critical-set views is the notion of a *life*. I follow Broome in loosely defining a life as "how things are for a person," where this phrase is understood to include all those things that can affect that life's *welfare*, how good the life is for the person living it.[1] This definition jars somewhat with our ordinary understanding of a life. Depending on our theory of welfare, it might count events occurring after a person's death as part of their life. But for our purposes, this terminological strangeness is of little consequence. The definition also allows that more than one person can live the same life. This possibility simplifies the ensuing discussion.

Advocates of critical-set views assume that welfare is both measurable on an interval scale and interpersonally level comparable. Measurability on an interval scale allows us to talk meaningfully about ratios of differences in welfare, so that claims like the following are meaningful: "The difference in welfare between the life Ada would have as an artist and the life Ada would have as a baker is twice the size of the difference in welfare between the life Ada would have as a baker

1 Broome, *Weighing Lives*, 94–95.

and the life Ada would have as a consultant." Interpersonal level comparability allows us to compare the welfare of different people, so that claims like the following are meaningful: "The life Ada would have as an artist contains more welfare than the life Bob would have as a baker." This claim is equivalent to the claim that "The life Ada would have as an artist is personally better than the life Bob would have as a baker." In other words, "The life Ada would have as an artist is better *for her* than the life Bob would have as a baker is *for him*." I mostly use the terminology of personal betterness below.

Advocates of critical-set views claim that each life's welfare can be represented by a real-valued function w, so that a life x is at least as personally good as a life y iff $w(x) \geq w(y)$, and the difference in welfare between x and y is k times the difference in welfare between y and z iff $|w(x) - w(y)| = k|w(y) - w(z)|$. This assumption implies that each pair of lives is *commensurable* with respect to welfare. That is, for all possible lives x and y, x is at least as personally good as y or y is at least as personally good as x. I will call $w(x)$ the *welfare level* of life x.

Critical-set views typically go on to sort lives into absolute categories. Which category a life falls in depends on how it compares to some standard: a life is *personally good* iff it is better than the standard, *personally bad* iff it is worse than the standard, and *personally neutral* iff it is neither better nor worse than the standard. The category of personally neutral lives can be refined further. Following Rabinowicz, I will say that a life is personally *strictly* neutral iff it is equally good as the standard and personally *weakly* neutral iff it is incommensurable with the standard.[2] The standard in question is defined differently by different authors. Some define it as nonexistence.[3] Others define it as a life constantly at a neutral level of temporal welfare.[4] Still others define it as a life without any good or bad components—features of a life that are good or bad for the person living it.[5] With one caveat, critical-set views are compatible with each definition.[6]

So much for comparing lives. Comparing populations—sets of lives—requires more machinery. Critical-set views start by designating some (gapless)

2 Rabinowicz, "Getting Personal," 80–81. Gustafsson calls these lives "neutral" and "undistinguished," respectively ("Population Axiology and the Possibility of a Fourth Category of Absolute Value").

3 Arrhenius and Rabinowicz, "The Value of Existence."

4 Broome, *Weighing Lives*, 68; Bykvist, "The Good, the Bad, and the Ethically Neutral," 101.

5 Arrhenius, "Future Generations," 26.

6 The caveat is that *neutral-range views*—explained below—cannot be paired with the latter two definitions. Neutral-range views claim that all lives are personally commensurable with each other and that some lives are personally incommensurable with the standard. That means that the standard cannot be a life. I thank an anonymous reviewer for pointing this out.

set of welfare levels to be the *critical set*. This critical set is defined to be the set of all welfare levels such that adding lives at these welfare levels to a population makes that population neither better nor worse. Each welfare level within this critical set is called a *critical level*. These critical levels play a key role in determining a life's *contributive value*, which we can understand as the contribution that a life makes to the value of a population. On critical-set views, the contributive value $c(x)_q$ of a life x relative to a critical level q is calculated by subtracting q from the welfare level $w(x)$:[7]

$$c(x)_q = w(x) - q.$$

The value of a population X relative to a critical level q is the sum of the contributive values of each life x_i in X relative to q:

$$v(X)_q = \sum_i c(x_i)_q.$$

And a population X is at least as good as a population Y iff $v(X)_q \geq v(Y)_q$ relative to each q in the critical set Q. If neither X nor Y is at least as good as the other, they are incommensurable.

Here is an example to illustrate. Suppose that we have two populations, X and Y. X contains one person at welfare level 5. Y contains three people at welfare level 2. On a critical-set view with a single critical level at 0, X is worse than Y.[8] On a view with a single critical level at 4, X is better than Y.[9] On a view with multiple critical levels including 0 and 4, X is incommensurable with Y because the value of X is not at least as great as the value of Y relative to $q = 0$ and the value of Y is not at least as great as the value of X relative to $q = 4$.

The characterization prior to this example constitutes the common core of critical-set views. The following four choice points divide the class. First, a critical-set view's critical set can comprise either a single critical level or multiple critical levels, forming a critical range. The former are *critical-level views*, and the latter are *critical-range views*. On critical-level views, lives at the critical level are *contributively strictly neutral*, by which I mean that adding these lives to a population leaves the new population equally good as the original. On critical-range

7 Critical-set views can also incorporate some real-valued function f applied to the welfare level and critical level. This function could be prioritarian: strictly increasing and strictly concave. I leave out the f purely for simplicity's sake. My discussion applies to any critical-set view on which f is strictly increasing. Any critical-set view on which f is not strictly increasing will violate *Dominance over Persons*, which says that for any populations X and Y featuring all the same people, if each person is at least as well off in X as they are in Y and some person is better off in X than they are in Y, then X is better than Y.

8 $v(X)_0 = (5 - 0) = 5$ and $v(Y)_0 = (2 - 0) + (2 - 0) + (2 - 0) = 6$.

9 $v(X)_4 = (5 - 4) = 1$ and $v(Y)_4 = (2 - 4) + (2 - 4) + (2 - 4) = -6$.

views, lives within the critical range are *contributively weakly neutral*, by which I mean that adding these lives to a population renders the new population incommensurable with the original. On all critical-set views, adding lives at welfare levels above the critical set makes a population better and adding lives at welfare levels below the critical set makes a population worse. I will call such lives *contributively good* and *contributively bad*, respectively.

The second choice point concerns the personally neutral set. This too can comprise either a single personally neutral level or a personally neutral range. *Neutral-level views* claim that lives at the personally neutral level are personally *strictly* neutral, so that they are personally equally good as the standard. *Neutral-range views* claim that lives within the personally neutral range are personally *weakly* neutral, so that they are personally incommensurable with the standard. From now on, I drop the "personally" from expressions like "personally neutral set." "Neutral set" refers to the set of welfare levels such that lives at those levels are personally neutral. "Critical set" refers to the set of welfare levels such that lives at those levels are contributively neutral.

The third choice point is one on which I have already taken a stand. Critical-range and neutral-range views can interpret their critical and neutral ranges as ranges of incommensurability, parity, indeterminacy, some other value relation, or any combination of the aforementioned phenomena.[10] I adopt the language of incommensurability in this paper, but my discussion can be translated into other terms without significant change to its import.

The fourth choice point concerns the relative positions of the critical and neutral sets. The options available at this stage depend on the directions taken at the first and second choice points, so I outline them in figure 1. The numbers at each terminus indicate which of the objections listed below apply to that view.

Many of the views in this taxonomy have never been advocated in print, but I lay them all out here for the sake of completeness. Four views that have been defended in print are the Total View, a positive critical-level view, a critical-range view, and a neutral-range view. I diagram them below. Horizontal lines denote that lives at the corresponding welfare level are personally/contributively strictly neutral. Boxes denote that lives at the corresponding welfare levels are personally/contributively weakly neutral. Lives at welfare levels above (below) the horizontal line or shaded box are personally/contributively good (bad). The numbers are purely illustrative.

10 For incommensurability, see Blackorby, Bossert, and Donaldson, "Quasi-Orderings and Population Ethics." For parity, see Qizilbash, "The Mere Addition Paradox, Parity and Vagueness" and "On Parity and the Intuition of Neutrality"; and Rabinowicz, "Broome and the Intuition of Neutrality." For indeterminacy, see Broome, *Weighing Lives*.

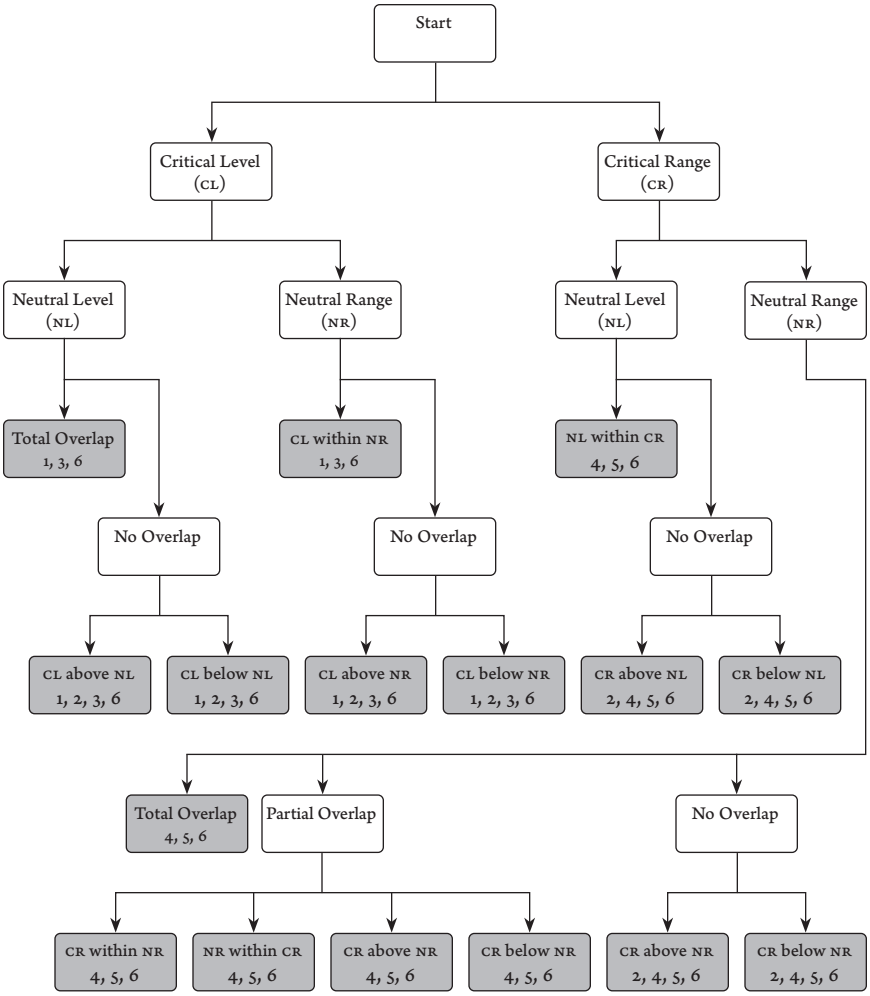

FIGURE 1 Taxonomy

First, the Total View (fig. 2), which is defended by Hudson, Tännsjö, and Huemer, among others.[11] There is a single coinciding neutral level and critical level, so that a life is personally good (bad/strictly neutral) iff it is contributively good (bad/strictly neutral). Any two populations are commensurable.

11 Hudson, "The Diminishing Marginal Value of Happy People"; Tännsjö, "Why We Ought to Accept the Repugnant Conclusion"; Huemer, "In Defence of Repugnance."

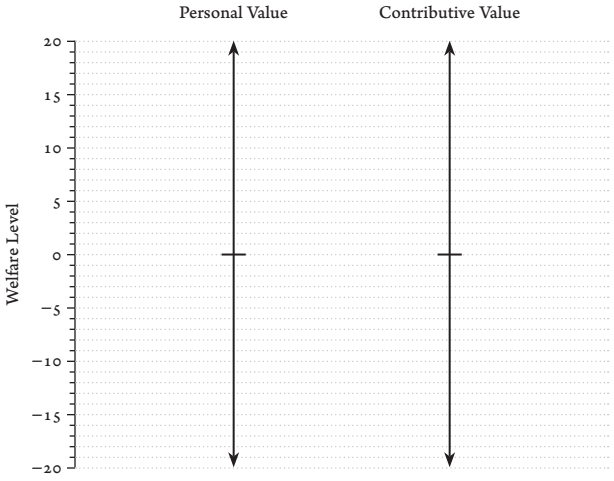

FIGURE 2 The total view

Second, a positive critical-level view (fig. 3), defended by Blackorby, Bossert, and Donaldson.[12] There is a single critical level above a single neutral level, so a life can be personally good without being contributively good. Any two populations are commensurable.

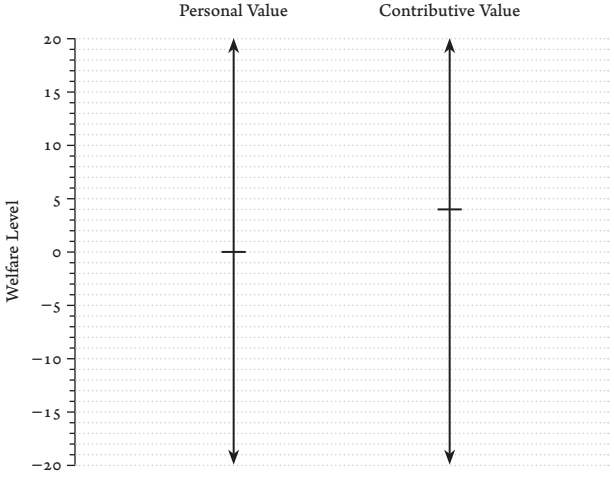

FIGURE 3 A positive critical-level view

12 Blackorby, Bossert, and Donaldson, *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*; Bossert, "Anonymous Welfarism, Critical-Level Principles, and the Repugnant and Sadistic Conclusions."

Third, a critical-range view. A view of this kind is defended by Broome, who interprets the critical range as a range of indeterminacy, along with Qizilbash and Rabinowicz, who each interpret the critical range as a range of parity.[13] There is a single neutral level but a critical range, so any overlap between the neutral and critical sets can be partial at most. In figure 4, I present a version of the view in which the neutral level coincides with the lowest welfare level in the critical range. On critical-range views, some pairs of populations are incommensurable.

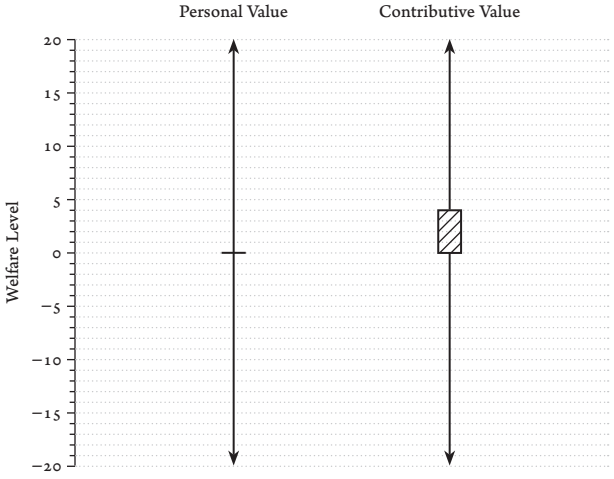

FIGURE 4 A critical-range view

Finally, a neutral-range view (fig. 5). Rabinowicz discusses a view of this kind in more recent work, and Gustafsson defends a view of this form in which there is a neutral and critical range for temporal welfare levels as well as lifetime welfare levels.[14] On neutral-range views, there is a neutral range and critical range that totally overlap, so a life is personally good (bad/weakly neutral) iff it is contributively good (bad/weakly neutral). Some pairs of populations are incommensurable.

13 Broome, *Weighing Lives*; Qizilbash, "The Mere Addition Paradox, Parity and Vagueness" and "On Parity and the Intuition of Neutrality"; Rabinowicz, "Broome and the Intuition of Neutrality."

14 Rabinowicz, "Getting Personal"; Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value."

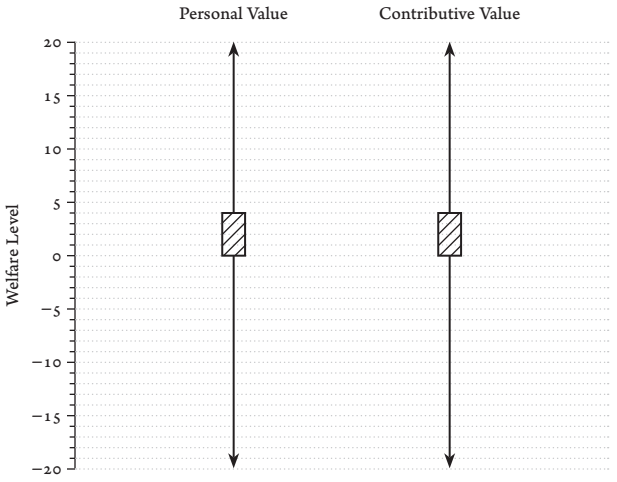

FIGURE 5 A neutral-range view

2. OBJECTIONS TO CRITICAL-SET VIEWS

Many varieties of critical-set view are subject to the same objections. Each view must reckon with at least three of the following six.

2.1. Maximal Repugnance

Any critical-set view on which lives barely worth living are contributively good will imply the

> *Repugnant Conclusion*: Each population of wonderful lives is worse than some population of lives barely worth living.[15]

And any critical-set view on which lives barely worth *not* living are contributively bad will imply the

> *Mirrored Repugnant Conclusion*: Each population of awful lives is better than some population of lives barely worth not living.[16]

Both of these consequences arise because, on critical-set views, a population of

15 Parfit, *Reasons and Persons*, 388.

16 Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value," 85. Carlson calls this claim the "Reverse Repugnant Conclusion" ("Mere Addition and Two Trilemmas of Population Ethics," 297). Broome calls it the "Negative Repugnant Conclusion" (*Weighing Lives*, 213).

enough contributively good (bad) lives can be better (worse) than any other population.

However, as Rabinowicz notes, the repugnance of these conclusions is attenuated if lives at a wide range of welfare levels are personally neutral.[17] In that case, lives barely worth living are much better than lives barely worth not living. What makes the Repugnant Conclusion and its mirror troubling is the presumed similarity of lives barely worth living and lives barely worth not living. With that in mind, I define *Maximal Repugnance* as follows:

> *Maximal Repugnance*: There is a life x and a life y that is identical but for one fewer gumdrop's worth of pleasure and one more hangnail's worth of pain such that (1) each population of wonderful lives is worse than some population of x lives and (2) each population of awful lives is better than some population of y lives.

Note that I drop the specification that x is barely worth living and y is barely worth not living. This feature is not necessary for repugnance. Suppose, for example, that we accept a view that implies Maximal Repugnance for a life x that is significantly personally good. This move mitigates the force of implication 1: we might be quite happy to accept that each population of wonderful lives is worse than some population of significantly personally good lives. But it exacerbates the implausibility of implication 2: if x is significantly personally good, then y is personally good, and it is hard to believe that each population of awful lives is better than some population of personally good lives. More generally, at least one of implications 1 and 2 will be implausible no matter how good x and y are.

Given that one fewer gumdrop's worth of pleasure and one extra hangnail's worth of pain can push a life's welfare level from above the critical level to below it, all critical-level views imply Maximal Repugnance.

2.2. *Sadism*

Any view on which there is no overlap between the critical set and the neutral set implies some sadistic conclusion. If the critical set is above the neutral set and there is some welfare level between the two, the view implies the original

> *Sadistic Conclusion*: Each population of awful lives is better than some population of personally good lives.[18]

That is because lives at a welfare level above the neutral set and below the critical

17 Rabinowicz, "Broome and the Intuition of Neutrality," 406, and "Getting Personal," 79.

18 Arrhenius, "An Impossibility Theorem for Welfarist Axiologies," 256.

set are personally good but contributively bad. And on critical-set views, adding enough contributively bad lives to a population can make that population worse than any other.

If the critical set is below the neutral set and there is some welfare level between them, the view implies the

> *Mirrored Sadistic Conclusion*: Each population of wonderful lives is worse than some population of personally bad lives.[19]

That is because lives at a welfare level below the neutral set and above the critical set are personally bad but contributively good. And on critical-set views, adding enough contributively good lives to a population can make that population better than any other.

We could endorse a critical-set view on which there is no overlap between the neutral set and the critical set and yet no welfare level between the two sets.[20] These kinds of views imply only weaker forms of sadism. If the critical set is above the neutral set, the view implies a

> *Weaker Sadistic Conclusion*: Each population of awful lives is better than some population of personally neutral lives.

If the critical set is below the neutral set, the view implies a

> *Weaker Mirrored Sadistic Conclusion*: Each population of wonderful lives is worse than some population of personally neutral lives.

These conclusions are more plausible than the pair above, but that is faint praise. In fact, comparison with the previous subsection will show that they could equally be called Stronger Mirrored and Stronger Repugnant Conclusions, respectively.[21]

All views with no overlap between the critical set and the neutral set imply some form of sadism.

19 Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value," 85.

20 That is possible if welfare levels are *not dense* (by which I mean that there is some pair of distinct welfare levels with no welfare level between them) or if the neutral set and critical set are such that exactly one of them is open at the end where they meet (e.g., if the neutral set is $[0, 1)$ and the critical set is $[1, 2]$).

21 I use the words "weaker" and "stronger" rather than "weak" and "strong" to distinguish these conclusions from the Weak Sadistic Conclusion and Strong Repugnant Conclusion that appear in Gustafsson ("Population Axiology and the Possibility of a Fourth Category of Absolute Value," 86) and Meacham ("Person-Affecting Views and Saturating Counterpart Relations," 270), respectively.

2.3. Strong Superiority across Slight Differences

Consider a sequence of lives beginning with a contributively good life x_1. We reach x_2 by making x_1 slightly worse. Perhaps x_2 is identical to x_1 but for one extra hangnail's worth of pain. We reach x_3 by making x_2 slightly worse, and so on. After a finite number of slight detriments we reach x_n, a contributively bad life.

On critical-level views, each life is either contributively good, contributively strictly neutral, or contributively bad. That means that, in our sequence, there is some contributively good life x_k such that x_{k+1} is either contributively strictly neutral or contributively bad. That in turn implies that x_k has positive contributive value, while x_{k+1}'s contributive value is nonpositive. Adding positive numbers can never yield a nonpositive number, and vice versa, so critical-level views imply that any population of lives x_k is better than any population of lives x_{k+1}. Call this implication *Strong Superiority across Slight Differences* (SSASD).[22]

We might claim that this implication is of little concern: x_k is contributively good and x_{k+1} is not, so the strong superiority of x_k over x_{k+1} should come as no surprise. But this level of description masks the difficulty. Consider a case in which each life in our x-sequence is long and turbulent, featuring soaring highs and crushing lows. Amid these peaks and troughs, we might expect a hangnail to pale almost into axiological insignificance. But critical-level views imply that this drop in the ocean can make all the difference: there will be a long, turbulent life x_k such that any population of lives x_k is better than any population of lives x_{k+1} identical but for the extra hangnail. Two corollaries of this implication bring out its implausibility: a population of just a single life without the hangnail is better than any population of lives with it, and a population of just a single life with the hangnail is worse than any population of lives without it.

2.4. Strong Noninferiority across Slight Differences

This instance of SSASD might spur us to adopt a critical-range view. On critical-range views, lives at a range of welfare levels are contributively weakly neutral. If this range is wide enough, our x-sequence will contain no lives x_k and x_{k+1} such that x_k is contributively good and x_{k+1} is contributively strictly neutral or bad. If x_k is the last contributively good life in the sequence, then x_{k+1} will be contributively *weakly* neutral. That means that critical-range views can avoid SSASD, because it is not the case that any population of contributively good lives is better than any population of contributively weakly neutral lives. Instead, each population of con-

22 For discussions of superiority and noninferiority in axiology, see Arrhenius and Rabinowicz, "Value Superiority"; Nebel, "Totalism without Repugnance"; and Thornley, "A Dilemma for Lexical and Archimedean Views in Population Axiology."

tributively good lives is incommensurable with some population of contributively weakly neutral lives. Here is an example to warm us up for the proof.

Suppose that all the welfare levels between 0 and 4 inclusive are critical. And suppose that $w(x_k) = 4.01$ and $w(x_{k+1}) = 3.99$. Population X consisting of a single life x_k is better than population Y consisting of a single life x_{k+1}, because $v(X) > v(Y)$ for each critical level q in the critical set Q. But X is incommensurable with population Z consisting of two lives x_{k+1}. X has greater value than Z relative to $q = 4$, but Z has greater value than X relative to $q = 0$.[23]

More generally, each contributively weakly neutral life has positive contributive value relative to some critical level q.[24] That implies that each population has less value than some sufficiently large population of contributively weakly neutral lives relative to that q. Therefore, each population is not better than some sufficiently large population of contributively weakly neutral lives.

However, critical-range views still imply *Strong Noninferiority across Slight Differences*: for some x_k and x_{k+1} in our x-sequence, any population of lives x_k is *not worse than* any population of lives x_{k+1}. To see how, return to our example above. No matter how many lives x_k are contained in X, and no matter how many lives x_{k+1} are contained in Z, X will have greater value than Z relative to $q = 4$. Therefore X is not worse than Z, no matter what their respective sizes. More generally, for any contributively good life x_k and any contributively weakly neutral life x_{k+1}, there exists some q such that x_k has positive contributive value relative to q and x_{k+1} has nonpositive contributive value relative to q. So relative to this q, any population of lives x_k has greater value than any population of lives x_{k+1}. That in turn implies that any population of lives x_k is not worse than any population of lives x_{k+1}. This kind of discontinuity is innocuous considered in itself. But as I demonstrate below, critical-range views imply that Strong Noninferiority across Slight Differences occurs in some counterintuitive places.

23 $v(X)_4 = (4.01 - 4) = 0.01$ and $v(Z)_4 = (3.99 - 4) + (3.99 - 4) = -0.02$; $v(X)_0 = (4.01 - 0) = 4.01$ and $v(Z)_0 = (3.99 - 0) + (3.99 - 0) = 7.98$.

24 We might think that lives at the lowest welfare level in the critical range are a counterexample to this claim. They do not have positive value relative to any critical level q in the critical range Q. But these lives are not contributively weakly neutral. On our definitions, they are contributively bad. Here is why. Suppose $w(x)$ is the lowest welfare level in the critical range Q. Then, for any population X, the value of X is at least as great as the value of X plus a life at $w(x)$ relative to each q in Q, so X is at least as good as X plus a life at $w(x)$. But the value of X plus a life at $w(x)$ is *not* at least as great as the value of X relative to each q in Q (in particular, it is not at least as great relative to critical levels q that are not the lowest in the critical range), so X plus a life at $w(x)$ is not at least as good as X. Therefore, X plus a life at $w(x)$ is worse than X, and x is contributively bad. This is strange because $w(x)$ is in the critical range, but this strangeness turns out to be of little consequence. We just need to bear in mind that only lives within the boundaries of the critical range are contributively weakly neutral.

Consider a new sequence. Each life in this sequence features a blank period, free of any good or bad components. We can imagine it as a minute of dreamless sleep. The first life in the sequence y_0 also features a period of constant happiness of length n hours, and nothing else. The second life y_1 is identical, except that the happiness lasts $n - 1$ hours. y_2's happiness lasts $n - 2$ hours, and so on. Call all such lives featuring only good and neutral components *straightforwardly better than blank*. Life y_n features only the blank period and so qualifies as a *blank life*, featuring no good or bad components whatsoever.[25] Life y_{n+1} features the blank period plus one hour of suffering, y_{n+2} features the blank period plus two hours of suffering, and so on. The last life in the sequence is y_{2n}, featuring the blank period plus n hours of suffering. Call all such lives featuring only bad and neutral components *straightforwardly worse than blank*.

Intuitively, the first discontinuity in this sequence occurs between y_{n-1} and y_n. That is, y_{n-1} is strongly noninferior to y_n: any population of lives y_{n-1} featuring one hour of happiness is not worse than any population of blank lives y_n. And, again intuitively, the second discontinuity in this sequence occurs between y_n and y_{n+1}. That is, y_{n+1} is strongly *nonsuperior* to y_n: any population of lives y_{n+1} featuring one hour of suffering is *not better* than any population of blank lives y_n. These two claims remain intuitive when we replace "hours" with "minutes," "seconds," "milliseconds," and so on.

But critical-range views must deny at least one of these claims. Recall that on critical-range views, more than one welfare level is critical. Therefore, in any sequence with sufficiently small differences in welfare between adjacent lives, more than one life is contributively weakly neutral. We can make the differences in welfare between adjacent lives in our y-sequence arbitrarily small by replacing hours with smaller units of time, so for some such unit, more than one life in our y-sequence is contributively weakly neutral.

Suppose for illustration that when the unit of time is seconds, y_{n-1} and y_n are the contributively weakly neutral lives. In that case, y_{n-2} (the last contributively good life) is strongly noninferior to y_{n-1} (the first contributively weakly neutral life). In other words, any population of lives featuring two seconds of happiness is not worse than any population of lives featuring one second of happiness. That implies that a population of just a *single* life featuring two seconds of happiness is not worse than any population of lives featuring one second of happiness. But this consequence seems implausible. The only difference between the lives is the duration of happiness; the latter population can feature an arbitrarily longer total duration of happiness, and yet the latter population can never be better than the former.

25 Broome, *Weighing Lives*, 208.

We get a mirror of this implication if we suppose instead that y_n and y_{n+1} are the contributively weakly neutral lives. In that case, any population of lives featuring two seconds of suffering is not better than any population of lives featuring one second of suffering. Though this latter population can feature an arbitrarily longer total duration of suffering, it can never be worse than a population of just a single life featuring two seconds of suffering. This too seems implausible.

Nothing hinges on the particular lives chosen to illustrate this dynamic. Any critical-range view will imply that (1) a population of just a single straightforwardly better-than-blank life is not worse than any population of straightforwardly better-than-blank lives identical but for a slightly smaller quantity of good, or (2) a population of just a single straightforwardly worse-than-blank life is not better than any population of straightforwardly worse-than-blank lives identical but for a slightly smaller quantity of bad.

2.5. Maximal Greediness

Critical-range views face another difficulty. As Broome points out, they imply that contributively weakly neutral lives can "swallow up" and neutralize goodness and badness.[26] Here is an illustration of what that means. Suppose again that all welfare levels between 0 and 4 inclusive are critical. And suppose that population *A* consists of a single life x at welfare level 20. We reach population *B* by making two changes. We reduce *x*'s welfare level by 1 and add a life *y* at welfare level 2. The combined effect of these changes might seem bad. We made one person worse off and added a life that is contributively weakly neutral. But our critical-range view implies that these changes are not bad. Neither *A*'s nor *B*'s value is at least as great as the other relative to each q in Q, so the two populations are incommensurable.[27] Our critical-range view also implies that *A* is incommensurable with *C* (in which *x*'s welfare level is 18 and there are two lives at welfare level 2) and *D* (in which *x*'s welfare level is 17 and there are three lives at welfare level 2) and so on. This process can continue indefinitely. *A* will also be incommensurable with a population *Z*, in which *x*'s welfare level is extremely low and there is some large number of contributively weakly neutral lives. Broome and I find this "greedy neutrality" concerning, but others are happy to bite the bullet.[28] In any case, the worry can be sharpened.

Note first that the size of population *A* need not be restricted to a single life:

26 Broome, *Weighing Lives*, 169–70 and 202–5.

27 Relative to $q = 4$, $v(A)_4 = (20 - 4) = 16$ and $v(B)_4 = (19 - 4) + (2 - 4) = 13$. Relative to $q = 0$, $v(A)_0 = (20 - 0) = 20$ and $v(B)_0 = (19 - 0) + (2 - 0) = 21$.

28 Rabinowicz, "Broome and the Intuition of Neutrality"; Frick, "On the Survival of Humanity"; Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value."

adding enough contributively weakly neutral lives can neutralize any finite loss of welfare for existing people. And suppose that blank lives are contributively weakly neutral. In that case, for any arbitrarily good population and any arbitrarily bad population, there is some population of blank lives—featuring no good or bad components whatsoever—such that the good population plus the blank lives is not better than the bad population. This implication seems difficult to accept.

It gets worse. Consider again our *y*-sequence above. Given that the unit of time is sufficiently small, critical-range views imply that more than one life in this sequence is contributively weakly neutral. For illustration, suppose that the blank life y_n and the straightforwardly better-than-blank life y_{n-1} are contributively weakly neutral. In that case, we can replace "blank lives" with "straightforwardly better-than-blank lives" in the above paragraph. For any arbitrarily good population and any arbitrarily bad population, there is some population of straightforwardly better-than-blank lives—featuring no bad components whatsoever and some happiness—such that the good population plus the straightforwardly better-than-blank lives is not better than the bad population. The former population might feature only neutral and good components, the latter population might feature only bad components, and yet this critical-range view implies that the former is not better than the latter.

If the straightforwardly worse-than-blank life y_{n+1} is contributively weakly neutral, we get a mirror of this implication. For any arbitrarily good population and any arbitrarily bad population, there is some population of straightforwardly worse-than-blank lives—featuring no good components whatsoever and some suffering—such that the bad population plus the straightforwardly worse-than-blank lives is not worse than the good population. Call implications of this kind *Maximal Greediness*.

Shifting the critical range away from blank lives fails to mitigate the difficulty. If the critical range is above or below the welfare level of a blank life, then some other life in our *y*-sequence will be contributively weakly neutral. No matter where the critical range is placed, we get Maximal Greediness.

2.6. No Incommensurability between Lives or between Same-Size Populations

On critical-level views, a population's value can be represented by a real number. Since any two real numbers are commensurable (a is at least as great as b or b is at least as great as a), critical-level views imply that any two populations are commensurable: X is at least as good as Y or Y is at least as good as X.

However, universal commensurability seems implausible. Consider the fol-

lowing small improvement argument.[29] Suppose that X consists of ten wonderful lives and Y consists of one hundred very good lives. Neither X nor Y is better than the other.[30] If any two populations are commensurable, X and Y are equally good. But if X and Y are equally good, then any population better than Y is better than X. Y^+, consisting of one hundred slightly-better-than-very-good lives, is better than Y but not better than X. Therefore, X and Y are not equally good. They are incommensurable.

Critical-range views can account for this incommensurability. They can claim that X has greater value than Y relative to one level in the critical range and that Y has greater value than X relative to another level. But this explanation cannot account for all plausible instances of incommensurability. In particular, it cannot account for the incommensurability of same-size populations.

This is easiest to see in the single-life case. Critical-set views assume that a life's welfare can be represented by a real number. Since any two real numbers are commensurable, this assumption implies that any two lives are commensurable: x is at least as good as y or y is at least as good as x.

Now note critical-set views' equation for the value of a population X relative to a critical level q:

$$v(X)_q = \sum_i (w(x_i) - q).$$

Since this equation is a sum of welfare levels minus the critical level, assuming that a life's welfare can be represented by a real number implies that a population's value relative to a critical level can be represented by a real number. That in turn implies that the value of any two populations relative to a critical level is commensurable. Formally,

1. For any populations X and Y and any critical level q, $v(X)_q \geq v(Y)_q$ or $v(Y)_q \geq v(X)_q$.

Now let X and Y stand for arbitrary same-size populations and q stand for an arbitrary critical level such that $v(X)_q \geq v(Y)_q$. Substituting in the equations for $v(X)_q$ and $v(Y)_q$ gives us the following inequality:

$$\sum_i (w(x_i) - q) \geq \sum_i (w(y_i) - q).$$

This inequality can also be expressed as follows, with n representing the size of populations X and Y:

$$(\sum_i w(x_i)) - nq \geq (\sum_i w(y_i)) - nq.$$

29 De Sousa, "The Good and the True"; Chang, "The Possibility of Parity."

30 Those who disagree should tweak the numbers or adjectives.

The terms involving q can then be canceled from each side:

$$\sum_i w(x_i) \geq \sum_i w(y_i).$$

Therefore, the inequality is true for all values of q, and X is at least as good as Y. Since X, Y, and q were arbitrary, we can conclude:

2. For any same-size populations X and Y and any critical level q, if $v(X)_q \geq v(Y)_q$, then X is at least as good as Y.

Together, 1 and 2 imply:

3. For any same-size populations X and Y, X is at least as good as Y or Y is at least as good as X.

In other words, critical-set views imply that any two same-size populations are commensurable.

However, universal commensurability of same-size populations seems implausible. Consider another small improvement argument. Suppose that x is a turbulent life, featuring soaring highs and crushing lows, and that y is a drab life, featuring only Muzak and potatoes.[31] If we fix the relative quantities of x's highs and lows in the right way, neither x nor y is better than the other. Yet x and y cannot be equally good because a slightly less drab life y^+—featuring Muzak, potatoes, and ketchup—is better than y but not better than x. Therefore, x and y are incommensurable. Similar arguments suggest the incommensurability of other pairs of same-size populations.

Partly on the basis of such arguments, advocates of critical-set views have started to incorporate incommensurability and indeterminacy into their theories of personal betterness. Broome, for example, states that some pairs of lives are obviously indeterminately related but offers no explanation for why this is so.[32] Rabinowicz, meanwhile, offers a fitting-attitudes analysis of parity—one species of incommensurability—according to which two lives are on a par iff it is permissible to prefer either life to the other.[33] And Gustafsson accounts for incommensurability between lives by claiming that there is a neutral range of temporal welfare levels.[34] Adding a moment within this range to a life renders the new life incommensurable with the original.

Gustafsson's move strikes me as a step in the right direction. However, his view cannot account for the incommensurability between same-length lives for

31 Parfit, "Overpopulation and the Quality of Life," 148.

32 Broome, "Loosening the Betterness Ordering of Lives."

33 Rabinowicz, "Getting Personal."

34 Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value."

the same reason that critical-range views cannot account for the incommensurability between same-size populations. Gustafsson might claim that any two lives of the same length are commensurable, but this claim seems implausible. The small improvement argument involving drab and turbulent lives remains convincing if we specify that the lives are the same length.

Rabinowicz's account is incomplete but, I believe, more promising. He claims that "life wellbeing is a many-dimensional concept," that "specifying its level requires characterizing a life with respect to several relevant dimensions," and that "different weight assignments" to these relevant dimensions give rise to incommensurability between lives.[35] This notion of "different weight assignments" forms the core of the Imprecise Exchange Rates View.

3. IMPRECISE EXCHANGE RATES

Some trade-offs are worth making. For example, going to the dentist to prevent tooth decay is a trade-off worth making. The good of having healthy teeth outweighs the bad of the trip. Other trade-offs are worth *not* making. Getting up at 4 AM and walking to work to save the £2 bus fare is a trade-off worth not making. The bad outweighs the good. Still other trade-offs are neither worth making nor worth not making, and a small improvement fails to break the deadlock. Here is an example.

A parent says to their child, "No dessert unless you finish your dinner." The child knows exactly what finishing dinner involves. They are all too familiar with the taste of peas and can see one hundred of them left on the plate. They also know what dessert will be like. The jelly is sitting on the counter and promises to taste as good as it always has. In this case, the trade-off may be neither worth making nor worth not making. And a small improvement to the child's predicament need not resolve the issue. Suppose that the parent takes pity on the child and removes one pea from the plate. That need not ensure that finishing dinner is now a trade-off worth making.

I claim that cases of this kind are evidence that various *exchange rates*—between pairs of goods, between pairs of bads, and between goods and bads—are imprecise. This imprecision renders certain goods incommensurable with other goods, certain bads incommensurable with other bads, and certain combinations of goods and bads incommensurable with other combinations. In the child's case, eating both the peas and the jelly is incommensurable with eating neither. This incommensurability between goods, bads, and their combinations

35 Rabinowicz, "Getting Personal," 81.

is the source of incommensurability between lives. The child's life in which they eat the peas and jelly is incommensurable with the otherwise identical life in which they eat neither.

That is one motivation for the Imprecise Exchange Rates (IER) view. Now for the formalization. Recall that critical-set views begin with an ordering of lives by welfare. The IER view begins instead with a set of orderings: one for each dimension of good and bad within a life. The exact form of the view thus depends on our theory of welfare. If we accept the simplest hedonist theory, there are just two orderings: one of happiness and one of suffering. If we accept an objective list theory, there are more orderings: perhaps one of love, one of virtue, one of false belief, etc. Welfare levels are thus given by vectors. Suppose, for example, that we accept an objective list theory on which happiness (h), love (l), suffering (s), and false belief (f) are the dimensions of good and bad. Then the welfare level of a life x is as follows:

$$w(x) = \langle h(x), l(x), s(x), f(x) \rangle.$$

I assume that h, l, s, and f are real-valued functions. I also assume that the values of each function are interpersonally level comparable (so that we can make claims like "The life Ada would have as an artist features more happiness than the life Bob would have as a baker") and measurable on a ratio scale (so that we can make claims like "The life Ada would have as an artist features twice the suffering of the life Ada would have as a baker"). Blank lives—featuring no good or bad components whatsoever—score 0 on each dimension.

Each ratio scale is independent, so we cannot yet compare values across dimensions. We cannot make claims like "In the life Ada would have as an artist, her happiness outweighs her suffering." Comparisons of this kind are only possible given a specified *proto-exchange-rate* r: a vector of two or more real numbers strictly greater than 0 and summing to 1 denoting the relative weight granted to each dimension of good and bad. On the objective list theory above, for example, each proto-exchange-rate r will take the form $\langle r_h, r_l, r_s, r_f \rangle$, where r_h denotes the weight granted to happiness, r_l denotes the weight granted to love, and so on. Letting x represent the life Ada would have as an artist, the claim that her happiness outweighs her suffering relative to a given r will be true iff $r_h h(x) > r_s s(x)$.

On the IER view, only welfare levels *relative to a given* r can be expressed as a real number. Continuing with our example objective list theory, the equation is as follows:

$$w(x)_r = r_h h(x) + r_l l(x) - r_s s(x) - r_f f(x).$$

The value of a population relative to r is the sum of the welfare levels of each of its lives relative to r:

$$v(X)_r = \sum_i w(x_i)_r.$$

We then account for incommensurability by claiming that there are multiple proto-exchange rates r in the set of all admissible proto-exchange rates R. A life x is at least as good as a life y iff $w(x)_r \geq w(y)_r$ relative to each r in R. And a population X is at least as good as a population Y iff $w(X)_r \geq w(Y)_r$ relative to each r in R.[36]

In what follows, I mostly discuss a simple hedonist version of the IER view, in which the welfare level of a life x is given by a vector of happiness and suffering, $\langle h(x), s(x)\rangle$, with the functions h and s normalized so that the proto-exchange-rate r composed of $r_h = 0.5$ and $r_s = 0.5$ falls within the set R. I adopt hedonism purely for the sake of simplicity. Its two dimensions are sufficient to illustrate the most important advantages and drawbacks of the IER view. My discussion below applies equally to variants of the view with more dimensions.

4. ADVANTAGES OF THE IMPRECISE EXCHANGE RATES VIEW

The IER view has several advantages over critical-set views. Here are four.

4.1. Some Incommensurability between Lives and between Same-Size Populations

The first advantage is that the IER view offers a simple and plausible account of incommensurability between lives and between same-size populations. Recall that a life is at least as good as another iff its welfare level is at least as great relative to each r in R. If R contains more than one r, then some pairs of lives are incommensurable: neither is at least as good as the other.

Consider an example. Suppose that R contains each r in which $0.4 \leq r_h \leq 0.6$. Since $r_h + r_s = 1$, $r_s = 1 - r_h$. In that case, life x—at welfare level $\langle 4, 1\rangle$—is incommensurable with life y—at welfare level $\langle 10, 6\rangle$. The welfare level of x is greater relative to $r_h = 0.4$, but the welfare level of y is greater relative to $r_h = 0.6$.[37] This is as it should be. Taking on the extra suffering in y for the sake of the extra happiness is a trade-off neither worth making nor worth not making.

The IER view also gives us the right result in small improvement cases. A

36 Rabinowicz offers a similar formalization ("Getting Personal," 83–84). His formalization, however, takes a set of permissible preferential ratio scales over the set of lives as primitive. It does not specify how the dimensions of welfare weigh against each other.

37 $w(x)_{r_h=0.4} = 0.4 \times 4 - 0.6 \times 1 = 1$ and $w(y)_{r_h=0.4} = 0.4 \times 10 - 0.6 \times 6 = 0.4$; $w(x)_{r_h=0.6} = 0.6 \times 4 - 0.4 \times 1 = 2$ and $w(y)_{r_h=0.6} = 0.6 \times 10 - 0.4 \times 6 = 3.6$.

slightly improved life y^+ at welfare level $\langle 10 + e, 6\rangle$ comes out better than y and incommensurable with x. That is because the IER view accounts for the incommensurability between lives while respecting a certain kind of dominance:

> *Dominance over Dimensions*: For any lives x and y and any set of proto-exchange-rates R, if for each good dimension g, x features at least as much g as y, and for each bad dimension b, x features at most as much b as y, x is at least as good as y. If, in addition, x features more g than y for some g or less b than y for some b, x is better than y.[38]

Another implication is related. Let us say that two proto-exchange rates *differ in optimism* iff they differ in the total weight granted to all dimensions of good taken together.[39] The implication is that if R contains proto-exchange rates that differ in optimism, then only lives featuring identical quantities of good and bad can be equally good.[40] That means that lives at welfare levels such as $\langle 4, 4\rangle$ and

38 Here is a sketch of the proof. Life x is at least as good as life y relative to any R iff $r_h h(x) - r_s s(x) \geq r_h h(y) - r_s s(y)$ for any $0 < r_h < 1$ and $r_s = 1 - r_h$. Rearranging this equation gives $r_h(h(x) - h(y)) + r_s(s(y) - s(x)) \geq 0$. If x dominates y, then $h(x) \geq h(y)$ and $s(y) \geq s(x)$, so each term on the left-hand side of the inequality in the previous sentence is nonnegative. Therefore, the weak inequality holds. If, in addition, x features more happiness or less suffering than y, then at least one term on the left-hand side of the inequality is positive, so the strict inequality holds. This proof can be extended to any number of dimensions of good and bad.

39 Here is an example. Return briefly to our objective list theory on which happiness, love, suffering, and false belief are the dimensions of good and bad, and consider the following three proto-exchange-rates: $r_1 = \langle 0.3, 0.2, 0.1, 0.4\rangle$, $r_2 = \langle 0.2, 0.3, 0.1, 0.4\rangle$, and $r_3 = \langle 0.3, 0.3, 0.1, 0.3\rangle$. Proto-exchange-rates r_1 and r_2 are distinct because r_1 assigns more weight to happiness while r_2 assigns more weight to love. But they are equally optimistic because they both assign a weight of 0.5 to both dimensions of good taken together. Proto-exchange-rate r_3, meanwhile, differs in optimism from both r_1 and r_2 because r_3 assigns a weight of 0.6 to both dimensions of good taken together.

40 To see this result, note first that equally good lives must have the same welfare level relative to each proto-exchange-rate. If x has a greater welfare level than y relative to some proto-exchange-rate, y is not at least as good as x, and so the pair cannot be equally good. Now let $g(x)$ denote the total quantity of good in x, $b(x)$ denote the total quantity of bad in x, and so on, and let r_1 and r_2 denote the total weight assigned to dimensions of good relative to proto-exchange-rates that differ in optimism. If x and y are equally good, then

$$r_1 g(x) - (1 - r_1)b(x) = r_1 g(y) - (1 - r_1)b(y)$$

and *mutatis mutandis* for r_2. Rearranging these equations gives

$$r_1(g(x) - g(y) + b(x) - b(y)) + b(x) - b(y) = 0$$

and *mutatis mutandis* for r_2. Since both expressions equal 0, they equal each other. Canceling $b(x) - b(y)$ from each side gives

⟨5, 5⟩ come out incommensurable on the IER view. This result is exactly what we want. Undergoing the extra suffering for the sake of the extra happiness is a trade-off neither worth making nor worth not making. If lives at ⟨4, 4⟩ and ⟨5, 5⟩ were judged equally good, the view would generate counterintuitive verdicts in small improvement cases. For example, a life at ⟨4, 4⟩ would be worse than a life at ⟨5, 5 − *e*⟩ for any $e > 0$. From now on, I assume that *R* contains proto-exchange-rates that differ in optimism.

The above three points are true of populations as well as lives. If *R* contains more than one *r*, then some pairs of populations (including same-size populations) are incommensurable. If one population weakly (strictly) dominates another over dimensions, then it is at least as good (better). And if *R* contains proto-exchange-rates that differ in optimism, then only populations featuring identical quantities of good and bad can be equally good.

4.2. No Sadism

Recall that critical-set views positing no overlap between the critical set and the neutral set imply some sadistic conclusion: either each population of awful lives is better than some population of lives that are not personally bad, or each population of wonderful lives is worse than some population of lives that are not personally good.

The IER view can avoid this drawback. More precisely, the IER view avoids sadism if we make the plausible claim that blank lives are personally strictly neutral. This claim implies that *only* blank lives are personally strictly neutral since, as we saw in the last subsection, no lives differing in their quantities of good or bad can be equally good. The extension of personal strict neutrality then matches the extension of contributive strict neutrality since, on the IER view, only blank lives are contributively strictly neutral. Adding any other kind of life changes the quantity of good or bad in the population, and no populations differing in their quantities of good or bad can be equally good.

This coincidence of personal and contributive strict neutrality suffices to establish that each category of personal value coincides with the corresponding category of contributive value. That is because the IER view then determines

---

$$r_1(g(x) - g(y) + b(x) - b(y)) = r_2(g(x) - g(y) + b(x) - b(y)).$$

Since $r_1 \neq r_2$, the expression $g(x) - g(y) + b(x) - b(y)$ must equal 0. That is true iff there exists some *k* such that $g(x) - g(y) = k$ and $b(x) - b(y) = -k$. If $k > 0$, then $g(x) > g(y)$ and $b(x) > b(y)$. In that case, *x* is better than *y* by strict dominance, so they cannot be equally good. If $k < 0$, then *y* is better than *x* by strict dominance. The only remaining possibility is that $k = 0$, in which case $g(x) = g(y)$ and $b(x) = b(y)$. Therefore, *x* and *y* are equally good only if they feature identical quantities of good and bad.

each life's personal and contributive category in the same way: its value is compared to the value of a blank life relative to each proto-exchange rate in *R*. That implies that a life is personally good (bad/strictly neutral/weakly neutral) iff it is contributively good (bad/strictly neutral/weakly neutral). Therefore, the IER view avoids all instances of sadism.

With the coincidence of each personal and contributive category of value on the IER view established, I often drop the words "personal" and "contributive" in what follows. In figure 6, I graph these coincident categories for lives at different welfare levels on the IER view with $0.4 \leq r_h \leq 0.6$. A life is good (bad/weakly neutral) iff the point picked out by its quantity of suffering on the horizontal axis and its quantity of happiness on the vertical axis falls within the dark (light/white) region. Lives at the origin are blank and hence strictly neutral.

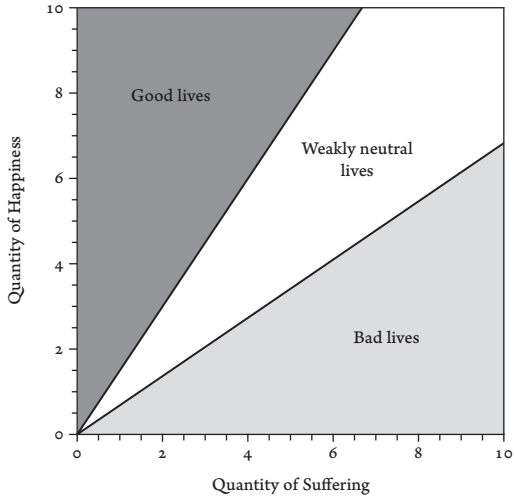

FIGURE 6 Coincident categories at different welfare levels.

4.3. Less Concerning Superiority and Noninferiority

As we saw above, critical-level views imply a concerning instance of Strong *Superiority* across Slight Differences (SSASD) in our x-sequence: there exists some long, turbulent life x_k such that any population of lives x_k is *better than* any population of lives x_{k+1} identical but for an extra hangnail. Critical-range views, meanwhile, imply only Strong *Noninferiority* across Slight Differences in our x-sequence: there exists some long, turbulent life x_k such that any population of lives x_k is *not worse than* any population of lives x_{k+1} identical but for an extra hangnail. But on critical-range views, at least one discontinuity of this kind must occur in a counterintuitive place in our y-sequence, so that there exists some

life y_k featuring only neutral components and happiness such that a population of just a single life y_k is not worse than any population of lives each featuring a slightly shorter duration of happiness, or there exists some life y_j featuring only neutral components and suffering such that a population of just a single life y_j is not better than any population of lives each featuring a slightly shorter duration of suffering.

The IER view avoids both of these problems. Consider first SSASD. Suppose, for illustration, that an extra hangnail adds 0.02 to a life's quantity of suffering. Suppose also that some turbulent life x_k has welfare level $\langle 9, 9\rangle$. Life x_{k+1} then has welfare level $\langle 9, 9.02\rangle$. Since x_k dominates x_{k+1}, population X consisting of a single life x_k is better than population Y consisting of a single life x_{k+1}. But X is incommensurable with population Z, consisting of two lives x_{k+1}. X has greater value than Z relative to $r_h = 0.4$, but Z has greater value than X relative to $r_h = 0.6$.[41]

We get the same result with lives at many other welfare levels. In fact, the IER view avoids SSASD in all but a small minority of cases. To see those cases in which SSASD is implied, let $\langle h(x_k), s(x_k)\rangle$ and $\langle h(x_k), s(x_k) + 0.02\rangle$ be the welfare levels of x_k and x_{k+1} respectively. Life x_k is strongly superior to life x_{k+1} iff x_k is good and x_{k+1} is strictly neutral or bad, or x_k is strictly neutral and x_{k+1} is bad. This condition is satisfied iff x_k's welfare level is nonnegative relative to the most pessimistic proto-exchange rate $r_h = 0.4$, x_{k+1}'s welfare level is nonpositive relative to the most optimistic proto-exchange rate $r_h = 0.6$, and at least one of x_k's or x_{k+1}'s welfare levels is non-zero relative to some r in R.[42] That yields two inequalities: $0.4h(x_k) - 0.6s(x_k) \geq 0$ and $0.6h(x_k) - 0.4(s(x_k) + 0.02) \leq 0$. Plotting these two inequalities gives us the region in figure 7.

A life x_k is strongly superior to an otherwise identical life x_{k+1} with an extra hangnail iff the point picked out by $s(x_k)$ on the horizontal axis and $h(x_k)$ on the vertical axis lies within the unshaded region. This is a welcome result. As we can see, an extra hangnail triggers strong superiority only when added to lives featuring very small quantities of happiness and suffering. The IER view thus gives hangnails their proper axiological due. In blank and nearly blank lives, they can be consequential. In turbulent lives, they pale almost into axiological insignificance.[43]

41 $v(X)_{r_h=0.4} = 0.4 \times 9 - 0.6 \times 9 = -1.8$ and $v(Z)_{r_h=0.4} = (0.4 \times 9 - 0.6 \times 9.02) + (0.4 \times 9 - 0.6 \times 9.02) = -3.624$; $v(X)_{r_h=0.6} = 0.6 \times 9 - 0.4 \times 9 = 1.8$ and $v(Z)_{r_h=0.6} = (0.6 \times 9 - 0.4 \times 9.02) + (0.6 \times 9 - 0.4 \times 9.02) = 3.584$.

42 The hangnail's worth of pain ensures that this last condition is met.

43 Reflecting this graph in the line $h = s$ gives the region of lives that can be pushed from bad or strictly neutral to good by an increase of 0.02 in that life's quantity of happiness. Perhaps this small jump corresponds to a gumdrop's worth of pleasure. As in figure 7, the region includes only lives featuring very small quantities of happiness and suffering.

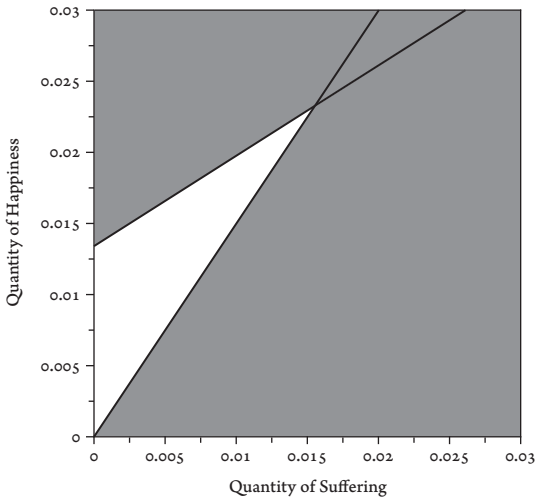

FIGURE 7 Welfare levels at which an extra hangnail triggers strong superiority

I write "almost" because an added hangnail can trigger strong *noninferiority*, even in turbulent lives. Consider again the case in which x_k's welfare level is $\langle 9, 9 \rangle$ and x_{k+1}'s welfare level is $\langle 9, 9.02 \rangle$. Given $r_h = 0.5$, $w(x_k)_{r_h=0.5} = 0.5 \times 9 - 0.5 \times 9 = 0$, and $w(x_{k+1})_{r_h=0.5} = 0.5 \times 9 - 0.5 \times 9.02 = -0.01$. Adding zeroes can never yield a negative number, and vice versa, so any population of lives x_k has greater value than any population of lives x_{k+1} relative to $r_h = 0.5$. That ensures that x_k is strongly noninferior to x_{k+1}: any population of lives x_k is not worse than any population of lives x_{k+1}.

More generally, an extra hangnail will trigger strong noninferiority whenever at least one of the lives being compared is weakly neutral. In that case, the extra hangnail will push the life's value from positive to negative relative to some r_h. Relative to that r_h, any population of lives without the hangnail has greater value than any population of lives with the hangnail. Therefore, any population of lives without the hangnail is not worse than any population of lives with the hangnail.

This too is a welcome result. Suppose we must choose between two populations. Each population consists of lives at only one welfare level, one population's lives are better than the other's, and at least one population consists of lives that are neither good nor bad. Then it is not worse to choose the population consisting of the better lives, regardless of the populations' respective sizes.

And importantly, the IER view does not imply strong noninferiority across straightforwardly better-than-blank lives or strong nonsuperiority across straightforwardly worse-than-blank lives, as critical-range views do. To see why, consider

a life y_k with welfare level $\langle a, 0\rangle$ and a life y_{k+1} with welfare level $\langle b, 0\rangle$. Suppose that $a > b > 0$, so that y_k is better than y_{k+1} and both are straightforwardly better than blank. Since both lives feature no suffering whatsoever, $w(y_k)_r$ and $w(y_{k+1})_r$ are positive relative to each r in R. That implies that for any r in R and any number m, there is some number n such that a population of n lives y_{k+1} has greater value than a population of m lives y_k relative to r. So for any number m, there is some number n such that a population of n lives y_{k+1} is better than a population of m lives y_k. The result is that y_k is not strongly noninferior to y_{k+1}.[44] A parallel line of argument proves that no straightforwardly worse-than-blank life is strongly nonsuperior to any other straightforwardly worse-than-blank life.

4.4. Less Concerning Greediness

Recall that critical-range views imply Maximal Greediness: for any population of awful lives and any population of wonderful lives, (1) there is some population of straightforwardly better-than-blank lives such that the population of awful lives is not worse than the population of wonderful lives plus the straightforwardly better-than-blank lives, or (2) there is some population of straightforwardly worse-than-blank lives such that the population of wonderful lives is not better than the population of awful lives plus the straightforwardly worse-than-blank lives. This disjunction follows from critical-range views' claim that lives at more than one welfare level are contributively weakly neutral and their assumption that any two lives are commensurable. Together, these imply that some straightforwardly better-than-blank life or some straightforwardly worse-than-blank life is contributively weakly neutral. And on critical-range views, adding enough contributively weakly neutral lives to a population can make that population incommensurable with any other.

The IER view agrees that lives at more than one welfare level are contributively weakly neutral. On the IER view with $R = \{r\colon 0.4 \leq r_h \leq 0.6\}$, for example, lives at $\langle 4, 3\rangle$ and $\langle 5, 4\rangle$ are both weakly neutral. But, as we have seen, it denies the assumption that any two lives are commensurable. Lives at $\langle 4, 3\rangle$ and $\langle 5, 4\rangle$ are one such incommensurable pair. As a result, the IER view avoids Maximal Greediness. Blank lives—with welfare level $\langle 0, 0\rangle$—have a value of 0 relative to each r in R, and so are contributively *strictly* neutral. Adding them to a population leaves the new population equally good as the original, so blank lives cannot swallow up goodness or badness.

Straightforwardly better-than-blank lives, meanwhile—with welfare level

44 Indeed, y_k is not even *weakly* noninferior to y_{k+1}. For the distinction between strong and weak noninferiority, see Thornley, "A Dilemma for Lexical and Archimedean Views in Population Axiology," 6.

$\langle a, 0\rangle$, $a > 0$—have positive value relative to each r in R, and so are contributively good. Adding them improves a population, so straightforwardly better-than-blank lives cannot swallow up and neutralize goodness. And *mutatis mutandis* for straightforwardly worse-than-blank lives. They cannot swallow up and neutralize badness. Therefore, the IER view implies neither disjunct of Maximal Greediness.

On the IER view, only lives featuring some positive quantity of good can neutralize badness, and only lives featuring some positive quantity of bad can neutralize goodness. This is as it should be.

5. OBJECTIONS TO THE IMPRECISE EXCHANGE RATES VIEW

The above four points constitute the main advantages of the IER view. Below are two objections.

5.1. Some Incommensurability between Good Lives and Weakly Neutral Lives

On the IER view, some good lives are incommensurable with some weakly neutral lives. Take a life x with welfare level $\langle 1, 0\rangle$ and a life y with welfare level $\langle 8, 7\rangle$. Life x is good, because $w(x)_r$ is positive relative to each $0.4 \leq r_h \leq 0.6$. Life y is weakly neutral, because $w(y)_r$ is positive relative to each $r_h > 0.4\dot{6}$ and negative relative to each $r_h < 0.4\dot{6}$. Yet x is incommensurable with y, because $w(x)_r < w(y)_r$ relative to each $r_h > 0.5$ and $w(x)_r > w(y)_r$ relative to each $r_h < 0.5$.

Although this consequence might seem odd, we ought to accept it. The reasons are twofold. First, the implication is not unique to the IER view. It is an inevitable consequence of admitting the possibility of lives both weakly neutral and close-to-strictly neutral, as Gustafsson and Rabinowicz note.[45] To see why, recall that strictly neutral lives are equally good as the standard and that weakly neutral lives are incommensurable with the standard. These definitions imply that strictly neutral lives are incommensurable with weakly neutral lives. As Raz notes, a small improvement or detriment to either of two incommensurable objects typically does not remove their incommensurability.[46] Such small tweaks can make a difference only when one of the two objects is almost better than the other. Therefore, if a strictly neutral life is neither almost better nor almost worse than some weakly neutral life, then some good life (slightly better than the strictly neutral life) and some bad life (slightly worse than the strictly neutral life) will also be incommensurable with the weakly neutral life.

45 Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value," 96; Rabinowicz, "Getting Personal," 86.

46 Raz, *The Morality of Freedom*, 326.

Second, incommensurability between some good lives and some weakly neutral lives follows from three claims that we should be reluctant to deny. The first is that a life featuring a positive quantity of good and no bad whatsoever (like a life at welfare level ⟨1, 0⟩) is good. The second is that a turbulent, neutral life (like a life at welfare level ⟨8, 7⟩) can be better than another neutral life (like a life at welfare level ⟨7, 7⟩). The third is that a good life at welfare level ⟨1, 0⟩ and a turbulent life at welfare level ⟨8, 7⟩ are such that neither is better than the other and a small improvement either way fails to break the deadlock.

5.2. Some Instances of Maximal Repugnance

On the IER view, life x with welfare level $\langle a, 0\rangle$ is good and life y with welfare level $\langle 0, a\rangle$ is bad for any $a > 0$. That implies that each population of wonderful lives is worse than some population of x-lives, and each population of awful lives is better than some population of y-lives. As a need only be larger than 0, lives x and y could be very similar. They could be identical but for x's featuring an extra gumdrop and y's featuring an extra hangnail. Therefore, the IER view implies Maximal Repugnance. Gustafsson, Broome, and Rabinowicz note that any view admitting the possibility of strictly neutral lives has implications of this kind, and they take it to be a reason to reject such views.[47]

However, I claim that ruling out the IER view on this basis is premature. Note first that implying this instance of Maximal Repugnance seems preferable to the alternative, which is to claim that lives with welfare level $\langle a, 0\rangle$ or $\langle 0, a\rangle$ for some $a > 0$ are contributively weakly neutral. As we have seen, that claim commits critical-set views to Maximal Greediness.

Note also that the IER view implies Maximal Repugnance only when lives x and y are nearly blank. If a life is turbulent, featuring a lot of happiness and suffering, then much more than a few extra gumdrops are required to move that life from bad to good. If we hold a life's quantity of suffering fixed at 6, for example, then the last contributively bad life has welfare level $\langle 4, 6\rangle$ and the first contributively good life has welfare level $\langle 9, 6\rangle$. Once again, the IER view is giving gumdrops and hangnails their proper axiological due. In nearly blank lives, they are significant. In turbulent lives, they fade into the background.

My final point is related. It is common in population axiology to think of lives barely worth living as drab. Parfit asked us to imagine lives in which the only pleasures are "muzak and potatoes."[48] But a Muzak and potatoes life can have

47 Gustafsson, "Population Axiology and the Possibility of a Fourth Category of Absolute Value," 96; Broome, "Loosening the Betterness Ordering of Lives," 8; Rabinowicz, "Getting Personal," 86–87.

48 Parfit, "Overpopulation and the Quality of Life," 148.

a welfare level of ⟨*a*, o⟩ only if its protagonist is very different from you and me. We—and everyone else endowed with an ordinary human psychology—would inevitably suffer boredom were we to live such a life, and lives at welfare level ⟨*a*, o⟩ feature no bad whatsoever. So, when we picture lives at ⟨*a*, o⟩, we should not imagine how we would feel sitting down to another bowl of mashed potatoes. Imagine instead a life of dreamless sleep, topped off with a gumdrop's worth of pleasure. When I conceive of ⟨*a*, o⟩ lives in this way, the IER view's implications no longer strike me as so repugnant.

6. CONCLUSION

The variety of possible critical-set views is dizzying, but each variety has serious drawbacks. On critical-level views, two extra hangnails can mark the difference between a good life and a bad life, even when the lives in question are long and turbulent. That means that a population of just a single life without the hangnails is better than any population of lives with them. It also means that each population of wonderful lives is worse than some population of lives without the hangnails, while each population of awful lives is better than some population of lives with them. On critical-range views, meanwhile, each population of wonderful lives and each population of awful lives is such that adding enough lives featuring only good and neutral components to the former makes it no better than the latter, or adding enough lives featuring only bad and neutral components to the latter makes it no worse than the former. What is more, some discontinuity in contributive value must occur in a counterintuitive place, so that a population of just a single life featuring only dreamless sleep and some duration of happiness is not worse than any population of lives identical but for a slightly shorter duration of happiness, or a population of just a single life featuring only dreamless sleep and some duration of suffering is not better than any population of lives identical but for a slightly shorter duration of suffering. Some varieties of critical-set views are sadistic, and no variety can account for the incommensurability between lives and between same-size populations without extra theoretical resources.

The IER view comes equipped with the required theoretical resources. It diagnoses as the source of incommensurability the fact that some trade-offs are neither worth making nor worth not making and a small improvement fails to break the deadlock. The resulting incommensurability between lives allows us to claim both that blank lives are strictly neutral and that a wide range of turbulent lives are weakly neutral, so that the IER view captures the advantages of both critical-level and critical-range views and charts the narrow course between Maximal Greediness and the most concerning instances of Maximal Repugnance.

Making the size of the contributively neutral range depend on a life's quantity of goods and bads has another nice consequence: it gives gumdrops and hangnails their proper axiological due. When a life is nearly blank, one fewer gumdrop and one extra hangnail can take it from good to bad. When a life is turbulent, gumdrops and hangnails pale almost into axiological insignificance. And because the IER view determines a life's categories of personal and contributive value in the same way, it escapes all forms of sadism.

In sum, the IER view is a worthy successor to critical-set views. It retains much of their appeal, while avoiding many of their pitfalls.[49]

University of Oxford
elliott.thornley@philosophy.ox.ac.uk

REFERENCES

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

———. "An Impossibility Theorem for Welfarist Axiologies." *Economics and Philosophy* 16, no. 2 (October 2000): 247–66.

Arrhenius, Gustaf, and Wlodek Rabinowicz. "The Value of Existence." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson, 424–44. New York: Oxford University Press, 2015.

———. "Value Superiority." In *The Oxford Handbook of Value Theory*, edited by Iwao Hirose and Jonas Olson, 225–48. New York: Oxford University Press, 2015.

Blackorby, Charles, Walter Bossert, and David Donaldson. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press, 2005.

———. "Quasi-Orderings and Population Ethics." *Social Choice and Welfare* 13, no. 2 (1996): 129–50.

Bossert, Walter. "Anonymous Welfarism, Critical-Level Principles, and the Repugnant and Sadistic Conclusions." In *The Oxford Handbook of Population Ethics*, edited by Gustaf Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns. Oxford: Oxford University Press, forthcoming.

Broome, John. "Loosening the Betterness Ordering of Lives: A Response to

49 I thank Hilary Greaves, Teruji Thomas, Tomi Francis, Kacper Kowalczyk, Alice van't Hoff, Todd Karhu, Nikhil Venkatesh, Jessica Fischer, Aidan Penn, Michal Masny, Farbod Akhlaghi, and two anonymous reviewers for helpful comments and discussion.

Rabinowicz." In *The Oxford Handbook of Population Ethics*, edited by Gustaf Arrhenius, Krister Bykvist, Tim Campbell, and Elizabeth Finneron-Burns. Oxford: Oxford University Press, forthcoming.

———. *Weighing Goods*. Oxford: Blackwell, 1991.

———. *Weighing Lives*. Oxford: Oxford University Press, 2004.

Bykvist, Krister. "The Good, the Bad, and the Ethically Neutral." *Economics and Philosophy* 23, no. 1 (March 2007): 97–105.

Carlson, Erik. "Mere Addition and Two Trilemmas of Population Ethics." *Economics and Philosophy* 14, no. 2 (October 1998): 283–306.

Chang, Ruth. "The Possibility of Parity." *Ethics* 112, no. 4 (July 2002): 659–88.

De Sousa, Ronald B. "The Good and the True." *Mind* 83, no. 332 (October 1974): 534–51.

Frick, Johann. "On the Survival of Humanity." *Canadian Journal of Philosophy* 47, nos. 2–3 (2017): 344–67.

Gustafsson, Johan E. "Population Axiology and the Possibility of a Fourth Category of Absolute Value." *Economics and Philosophy* 36, no. 1 (March 2020): 81–110.

Hudson, James L. "The Diminishing Marginal Value of Happy People." *Philosophical Studies* 51, no. 1 (January 1987): 123–37.

Huemer, Michael. "In Defence of Repugnance." *Mind* 117, no. 468 (October 2008): 899–933.

Meacham, Christopher J. G. "Person-Affecting Views and Saturating Counterpart Relations." *Philosophical Studies* 158, no. 2 (March 2012): 257–87.

Nebel, Jacob M. "Totalism without Repugnance." In *Ethics and Existence: The Legacy of Derek Parfit*, edited by Jeff McMahan, Tim Campbell, James Goodrich, and Ketan Ramakrishnan, 200–31. Oxford: Oxford University Press, 2021.

Parfit, Derek. "Overpopulation and the Quality of Life." In *Applied Ethics*, edited by Peter Singer, 145–64. Oxford: Oxford University Press, 1986.

———. *Reasons and Persons*. Oxford: Clarendon Press, 1984.

Qizilbash, Mozaffar. "The Mere Addition Paradox, Parity and Vagueness." *Philosophy and Phenomenological Research* 75, no. 1 (July 2007): 129–51.

———. "On Parity and the Intuition of Neutrality." *Economics & Philosophy* 34, no. 1 (2018): 87–108.

Rabinowicz, Wlodek. "Broome and the Intuition of Neutrality." *Philosophical Issues* 19, no. 1 (October 2009): 389–411.

———. "Getting Personal: The Intuition of Neutrality Reinterpreted." In *Studies on Climate Ethics and Future Generations*, vol. 2, edited by Paul Bowman and Katharina Berndt Rasmussen, 59–90. Stockholm: Institute for Future

Studies, 2020. https://www.iffs.se/en/publications/working-papers/studies-on-climate-ethics-and-future-generations-vol-2/.

Raz, Joseph. *The Morality of Freedom*. Oxford: Oxford University Press, 1986.

Tännsjö, Torbjörn. "Why We Ought to Accept the Repugnant Conclusion." *Utilitas* 14, no. 3 (November 2002): 339–59.

Thornley, Elliott. "A Dilemma for Lexical and Archimedean Views in Population Axiology." *Economics and Philosophy*. Published ahead of print, September 6, 2021. https://doi.org/10.1017/s0266267121000213.

Journal of Ethics and Social Philosophy
VOL. 22, NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1421

QUALITY OF WILL ACCOUNTS AND NON-CULPABLY DEVELOPED MENTAL DISORDERS

Matthew Lamb

A FAMILIAR FACT about our practice of blame is that an agent's ignorance sometimes, but not always, excuses what would otherwise be a blameworthy wrongdoing. This aspect of blameworthiness is the epistemic condition of blameworthiness. Dylon McChesney and Mathieu Doucet rightly note that any viable account of the epistemic condition must properly account for the significance of ignorance that is due to an agent's mental disorder. As they note,

> your reaction to someone who does not notice your distress because he is an inconsiderate jerk is (we hope!) quite different from your typical reaction to someone who does not notice your distress because she is depressed or on the autism spectrum. Reactive attitudes like blame and resentment are standard in the first case, but inappropriate in the second.[1]

This seems exactly right. An important commitment of our ordinary practice of blame is that mental disorders sometimes excuse an agent for what would otherwise be blameworthy ignorance. If an account of the epistemic condition cannot capture this commitment, then the account is not viable. Call this the disorder-based viability constraint.

McChesney and Doucet use the disorder-based viability constraint to argue (i) against George Sher's account of the epistemic condition and (ii) in favor of a quality of will view.[2] Against Sher's account, they argue as follows:

1. Mental disorders that "(*a*) involve the agent's constitutive dispositions and traits and (*b*) explain the agent's ignorance" sometimes (but not always) excuse.

1 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 235.

2 McChesney and Doucet, "Culpable Ignorance and Mental Disorders."

2. All mental disorders that meet conditions *a* and *b* fail to excuse on Sher's view.
3. Thus, Sher's view falls short of the disorder-based viability constraint.[3]

McChesney and Doucet then argue that since a quality of will view can tie blameworthiness to the agent's moral concerns, a quality of will view can accurately capture the range of cases where mental disorders excuse.

However, I argue that their quality of will approach also fails the disorder-based viability constraint. When it comes to cases where the agent developed a mental disorder in adolescence, our ordinary practice of blame sometimes takes this fact to be excusing. Any account of the epistemic condition that meets the disorder-based viability constraint needs to accurately account for the full range of cases where developing a disorder in adolescence is an excuse. Yet McChesney and Doucet's view cannot capture the full range of those cases. Thus, their view falls short of the disorder-based viability constraint.

1. QUALITY OF WILL ACCOUNTS

Let us begin with an overview of the quality of will account defended by McChesney and Doucet.[4] Their view holds that when an agent is blameworthy for x, it is because x reflects a morally objectionable aspect of the agent's moral concerns.[5] Accordingly, an agent's epistemic relation to his wrongdoing matters for blameworthiness on their view insofar as it bears on the moral concern expressed by the wrongdoing. For instance, if the agent is ignorant about the wrongness of the action because he simply is not concerned with what matters morally (e.g., fairness), then the ignorance reflects deficient moral concerns. And so the ignorance is blameworthy. But if the agent does not know better about the wrongness of the action because his attention is limited by fatigue rather than a deficiency in his moral concerns, then his ignorance does not reflect poor moral concern. In turn, the ignorance would not be blameworthy. The same applies to ignorance caused by mental disorders. When the presence of the agent's mental disorder–based ignorance is not explained by the agent's lack of moral concerns

3 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 231.

4 McChesney and Doucet cite Arpaly and Schroeder (*In Praise of Desire*) and Smith ("Responsibility for Attitudes") as the sort of account they are building on. Other quality of will views include Harman, "Does Moral Ignorance Exculpate?"; Scanlon, *Moral Dimensions*; and Talbert, "Moral Competence, Moral Blame, and Protest."

5 For readers who hold that there are distinct types of blame with corresponding distinct types of blameworthiness, one can understand McChesney and Doucet as concerned with blameworthiness as the appropriateness of moral resentment.

but is instead explained by the disorder, then the ignorance is excused; the ignorance is not an appropriate target of resentment.[6]

2. BUILDING A COUNTEREXAMPLE

In what follows, I argue that McChesney and Doucet's quality of will view lacks the resources for adequately addressing some cases of disorder-based ignorance where (i) the ignorance reflects a deficiency in moral concern, (ii) the disorder is developed (and maintained) through no fault of the agent during adolescence, and (iii) the disorder poses an unreasonably demanding difficulty for avoiding the ignorance.

Consider the following scenario.

> *Narcissistic Joe*: As a young child, Joe's life contains multiple risk factors for developing narcissistic personality disorder, such as having a cruel, authoritarian, and neglectful family at home. In his youth, while his peers are developing into empathetic, healthy individuals, Joe's desire for self-esteem develops in the unhealthy direction of having an overly inflated sense of self-importance that is maintained at the expense of others. Moreover, young Joe is neither diagnosed nor treated for his disorder. As a result of developing this disorder in his childhood and not receiving treatment, Joe grows into a young adult who finds it incredibly difficult to be empathetic. Frequently in his young adult life, Joe's narcissism results in him being ignorant of the moral significance of others' well-being.

Joe's ignorance of the importance of others' well-being is tied to a mental disorder that he developed during childhood. Moreover, let us consider a period of Joe's young adult life where there have been some opportunities to recognize that he has a serious personality disorder and that he should seek help, but *not* to an extent where he could reasonably be expected to do so. When it comes to this period of Joe's life, does his disorder-based ignorance warrant blame as resentment?[7]

To see why Joe's ignorance does not merit resentment, let us imagine the following. Joe has inconvenienced you by lying and he showed no regard for how this impacted you. Your initial reaction may understandably be one of resentment. But when you share what happened with a colleague, you learn more about Joe.

6 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 244.

7 McChesney and Doucet accept that their view may preclude personality disorders from the category of excusing disorders. The Narcissistic Joe case aims to show that this leads to violations of the disorder-based viability constraint. See McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 245–46.

You learn that he is not just an ordinary jerk. Due to his childhood and deficient opportunities for seeking therapy, Joe suffers from narcissistic personality disorder. And while it is not impossible for him to see the wrongness of lying and manipulating others to get ahead, it is especially difficult for him. As your colleague tells you, it would be unreasonable to expect Joe's disorder-based ignorance to be resolved by Joe simply deciding to be more considerate; his disorder calls for professional help. And while there is nothing that makes it impossible for him to seek help, the way that a person with narcissistic personality disorder views the world makes it especially difficult (but not impossible) for Joe to even see that there is a problem with himself. His personality disorder that has been acquired in childhood sets him up to think of himself as exceptional and to tend to give this assumption more credence than the counterevidence he might get exposed to. Thus, even an expectation that he recognizes that there is a problem in the first place would itself be unreasonably demanding.[8] After learning of Joe's history and the difficulty he now faces for knowing better, the initial blame and resentment you held should no longer seem appropriate. Now the appropriate response is to withdraw (or at least severely mitigate) your blaming reaction toward Joe for the ignorant wrongdoing. Sure, Joe is ignorant because he is a narcissistic jerk, but what other kind of young adult could he reasonably be expected to grow into? He developed a mental disorder during adolescence that calls for professional help. If you maintain your resentment, that would be unjustly harsh toward Joe.[9]

I hope we can now see that Joe's case is one of disorder-based ignorance that reflects poorly on the agent's moral concerns, yet resentment is plausibly not appropriate. However, this alone does not raise a problem for McChesney and Doucet's quality of will view. They rightly note that their view has resources to deem some cases of disorder-based ignorance that reflect poorly on the agent to be cases where the individual should not be blamed.[10] But, as I argue, these resources are inadequate.

3. INADEQUATE RESOURCES

In the final section of their article, McChesney and Doucet highlight the fact that just because ignorance reflects an individual's poor moral concerns, it does not follow that their view deems the person blameworthy. This is because there's nothing about a person's ignorance reflecting poor moral concern that necessarily precludes the existence of "independent reasons for supposing that [the individ-

8 Ronningstam, "Narcissistic Personality Disorder."

9 This is not to say that it is inappropriate to feel upset, insulted, or even frustrated.

10 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 245–46.

ual] ought to be exempt from blame."[11] McChesney and Doucet do not say what exactly these independent reasons are, just that they would be "very different from the reasons we have offered here."[12] I take this to mean that the reasons, whatever they may be, would be reasons that are independent of the *epistemic condition* of blameworthiness. If this is right, then there are two general categories of reasons that can serve as independent reasons for exempting the agent from blame.

One category of independent reasons pertains to the agent failing a condition of moral responsibility that is not the epistemic condition. When a reason in this category occurs, the fact that the person is ignorant (i.e., their epistemic relation to the wrongness) would not itself explain the lack of blameworthiness. Instead, the lack of blameworthiness would be tied to the person's deficiency in control or moral agency. For instance, consider someone who meets the diagnostic criteria for narcissistic personality disorder because that person *lacks* the general ability to understand the fact that other people's well-being matters. In such a case, their view could say that the person has a deficiency in moral agency, such that when he is ignorant due to his lack of capacity, he is not blameworthy. This would not be because he fails the *epistemic* condition, but because he fails a prerequisite for even being a candidate for blameworthiness in the first place: having sufficient capacities for moral agency. However, this would not apply to *all* cases of ignorance rooted in narcissistic personality disorder. There is nothing about the diagnostic criteria that requires a person to lack that capacity.[13] My point here is just to highlight one way that there could be independent reasons in a case of mental disorder–based ignorance where blame is not appropriate.

The other category of independent reasons consists of reasons that are independent of moral responsibility itself rather than only being independent of the epistemic condition. Reasons in this category could make an individual exempt from blame by *overriding* the responsibility-based reasons for blame. A paradigmatic example of this type of reasons is a forward-looking consideration, such as the ineffectiveness of engaging in blame to correct behavior compared to the effectiveness of showing compassion, patience, and understanding. For instance, consider a case where the mental disorder explains why the person's moral concerns are frequently deficient, but where the person still meets the conditions for moral agency and responsibility. On McChesney and Doucet's view, this person is not off the hook *via the epistemic condition* since the ignorance does reflect an objectionable deficiency in moral concern. However, if our goal is to encourage this person to foster a tendency to take steps that are conducive to consider-

11 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 245–46.

12 McChesney and Doucet, "Culpable Ignorance and Mental Disorders," 246.

13 American Psychiatric Association, "Personality Disorders."

ing the significance of others' well-being, being resentful toward him might be counterproductive to our goal. The value of this goal of improving the person's behavior might give us overriding reasons *not* to blame the agent, even if the conditions for being morally responsible for the ignorance are met.

However, even with these resources for holding that an individual sometimes should not be blamed despite the disorder-based ignorance reflecting poor moral concerns, the case of Narcissistic Joe can still highlight a problem for McChesney and Doucet's view. There is nothing about Joe's case that requires us to build in an independent reason for exempting Joe from blame. While it is true that some cases of narcissistic personality disorder involve a lack of certain capacities necessary for moral agency, it need not occur in all cases where the diagnostic criteria are met. In fact, as the case of Joe is written, it is set up to where Joe has the various capacities needed for meeting the non-epistemic conditions of responsibility. He did not fail to develop a *capacity* for empathy, even though it is especially difficult for him to be empathetic. Similarly, there is nothing about a case of narcissistic personality disorder that requires us to build in reasons for exempting the agent from blame that are independent of concerns about moral responsibility–based blame (e.g., pragmatic reasons for withholding blame). For instance, suppose the person Joe wrongs is a passing stranger whose reaction, whether resentful or sympathetic, has no bearing on the likelihood of Joe seeking professional therapy. In short, there is no reason we cannot set up the Joe case to be one where there is no independent reason for exempting Joe from blame. Yet if what I have said above is correct about the significance of Joe's adolescence and deficiency of reasonable opportunities to pursue treatment, the attitude of resentment is inappropriate. And this is so even in the absence of *independent* reasons for withholding blame. Thus, McChesney and Doucet's view mistakenly deems Joe's ignorance as warranting resentment.

4. ANOTHER COUNTEREXAMPLE

Their view's inadequate resources for capturing the full range of cases where disorder-based ignorance is not worthy of resentment is not limited to ignorance due to personality disorders. The view also faces problems when it comes to more familiar disorders, such as depression. Consider a case of Joe's sister, Michele, who develops major depressive disorder in adolescence. Michele is currently a young adult whose life, strictly speaking, contains opportunities to seek professional help, but not to the extent that getting professional help is something that could reasonably be expected of her. During this period of her life, she frequently suffers from episodes of depression where she fails to care about the

right sort of things, such as her friendships and other important relationships. Moreover, this is not a case of her being too fatigued to act on her actual concern for her friendships. Instead, her depression is simply manifested as a lack of interest and concern for a great number of things, including being a good friend. For instance, when she thinks about keeping a promise to a friend, it is not impossible for her to see that it is worth doing, but it is very difficult for her to judge it as worth doing. Due to this disorder-based difficulty, she fails to judge the promise to be worth keeping.

Michele's ignorance reflects her deficient moral concern for the value of promise keeping and friendship. Yet she is not being an ordinary jerk. She is suffering from major depressive disorder. And in this particular case, her disorder-based ignorance does not warrant resentment. Any account of the epistemic condition that meets the disorder-based viability constraint must be able to capture this verdict about her ignorance. However, since we are not supposing that there are independent reasons to exempt Michele from being an appropriate target of blame, McChesney and Doucet's view holds that Michele's ignorance is blameworthy. Their view thereby falls short of the disorder-based viability constraint when it comes to cases like Michele's.

5. CONCLUSION

The significance of Narcissistic Joe and Michele is *not* that disorder-based ignorance always excuses. Their significance is that they highlight a category of mental disorder–based ignorance that plausibly excuses. Cases of mental disorder–based ignorance that fall into this category are instances of ignorance rooted in the agent's mental disorder, where (i) the ignorance reflects deficient moral concern(s), (ii) the disorder is developed (and maintained) through no fault of the agent during adolescence, and (iii) the disorder imposes a difficulty in avoiding or correcting the ignorance, such that an expectation to overcome said difficulty is unreasonably demanding. When these conditions are met and there are no independent reasons for exempting the agent from blame, then McChesney and Doucet's view takes the ignorance as not an excuse. Yet some of those, such as Joe's and Michele's, are cases where ordinary practice takes the disorder-based ignorance as not warranting resentment. Thus, their quality of will view falls short of the disorder-based viability constraint.

University of Rochester
mlamb6@ur.rochester.edu

REFERENCES

American Psychiatric Association. "Personality Disorders." In *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC: American Psychiatric Association, 2013.

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

Harman, Elizabeth. "Does Moral Ignorance Exculpate?" *Ratio* 24, no. 4 (November 2011): 443–68.

McChesney, Dylon, and Mathieu Doucet. "Culpable Ignorance and Mental Disorders." *Journal of Ethics and Social Philosophy* 14, no. 3 (February 2019): 227–48.

Ronningstam, Elsa. "Narcissistic Personality Disorder." In *Gabbard's Treatments of Psychiatric Disorders*, 5th ed., edited by Glen O. Gabbard, 1073–86. Washington, DC: American Psychiatric Association, 2014.

Scanlon, T. M. *Moral Dimensions: Permissibility, Meaning and Blame*. Cambridge, MA: Harvard University Press, 2008.

Smith, Angela. "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115, no. 2 (January 2005): 236–71.

Talbert, Matthew. "Moral Competence, Moral Blame, and Protest." *Journal of Ethics* 16, no. 1 (March 2012): 89–109.

Journal of Ethics and Social Philosophy
VOL. 22 NO. 3 · SEPTEMBER 2022
https://doi.org/10.26556/jesp.v22i3.1308

THE SHERIFF IN OUR MINDS

ON THE MORALITY OF THE MENTAL

Samuel Director

MANY PEOPLE believe that our thoughts can be morally wrong. Many regard rape and murder fantasies as wrong. In a recent essay, George Sher disagrees with this and argues that "the realm of the purely mental is best regarded as a morality-free zone," wherein "no thoughts or attitudes are either forbidden or required."[1] Sher argues that "each person's subjectivity is a limitless, lawless wild west in which absolutely everything is permitted."[2] Sher calls this view the "Wild West of the Mind."

I argue against Sher's position. In section 1, I summarize Sher's view. In section 2, I outline and criticize Sher's argument for the Wild West of the Mind. Sher identifies two features of the mental realm that he thinks put our thoughts beyond the scope of morality. The first feature of the mental realm that Sher appeals to is that rules against actions have discrete boundaries, while rules against thoughts do not. I argue that this problem is equally true of actions and thoughts, meaning that this argument fails to show how thoughts are morally different from actions. The second feature of the mental realm that Sher points to is that our mental lives are impermeable to and disconnected from other people, meaning that they cannot wrong others. I argue that our thoughts, despite being impermeable and disconnected, can wrong others by inflicting unfelt harms upon them.

In section 3, I outline additional objections to Sher's view. First, I argue that Sher's view should actually be understood as being about the permissibly of any unheard utterance, not just about the permissibility of private thoughts. This clarification, I argue, renders his view implausible. Second, I argue that our thoughts can inflict unfelt harms on others, making them sometimes impermissible. Third, I argue that Sher's position on thought-action composites is implausible.

1 Sher, "A Wild West of the Mind," 484.

2 Sher, "A Wild West of the Mind," 484.

1. SHER'S ARGUMENT

Sher's thesis is the following:

> *Wild West of the Mind* (WWM): For any purely private thought, *T*, that is expressed in an agent's, *S*'s, mind, *T* is not morally wrong.

Sher offers several clarifications of WWM. First, he is not denying that one's thoughts "can reflect badly on his character."[3] He agrees that some thoughts suggest that an individual is vicious. Second, he agrees that one's thoughts can be epistemically wrong (i.e., epistemically unjustified). With these clarifications, Sher's claim is that "where a person's private mental contents are concerned," condemnation based on viciousness or epistemic wrongness "are the only forms of condemnation that are in place."[4] As Sher later argues, these forms of condemnation are not sufficient to support the view that our private thoughts can be wrong.

Lastly, Sher is not addressing cases where an individual's actions are made better or worse in virtue of her thoughts. For example, imagine a case in which Smith pushes Jones to hurt him and another case where Smith pushes Jones to save him from being hit by a train. Smith's thoughts are clearly morally relevant. It would be implausible to deny that Smith's intentions make a moral difference in how we should evaluate his actions. To avoid this implication, Sher distinguishes between purely private thoughts and thoughts that have both a public and a private component. The former exist only in the mind and have no physical expression in the world, while the latter include both thought and action. In the cases of Smith and Jones, we are dealing with "composite occurrences that have both public and private components."[5] Sher's claim is only that our purely private thoughts cannot be wrong.

2. PROBLEMS WITH SHER'S ARGUMENT:

Sher defends WWM by arguing that all available arguments against WWM are bad and also by outlining positive reasons for WWM. I will not address Sher's negative argument. Instead, I object to his positive argument for WWM. In his positive argument, Sher identifies two features of the mental that seem to put our thoughts beyond the scope of morality.

The first feature of the mental that Sher appeals to is that rules against actions have discrete boundaries, while rules against thoughts do not. As Sher says,

3 Sher, "A Wild West of the Mind," 484.

4 Sher, "A Wild West of the Mind," 484.

5 Sher, "A Wild West of the Mind," 485.

"when morality or the law forbids me to shoot you, the act that it forbids begins with my decision to pull the trigger and ends with the bullet penetrating your body."[6] However, when it comes to thoughts, "if morality were to forbid me to think of shooting you, the prohibition would inevitably diffuse itself, like dye poured into water, among innumerable other thoughts and feelings."[7] In short, to avoid thinking about shooting someone, one needs to refrain from a host of other thoughts about that person. Sher continues: normal people "can easily resist the transition from the impulse to shoot you to the deed itself," but most normal people "have far less control over … [their] inferences and associations; so, to guard against thinking about shooting you, I would also have to avoid many other thoughts."[8] If it is wrong for me to fantasize about shooting someone, I would "have to avoid dwelling on the wrong that [they] have done to me," because that thought may lead me to think about shooting them.[9] More generally, "to know which thoughts to avoid, I would have to know which ones might lead to the forbidden thought" and avoid them too.[10] So, actions can be morally evaluated because they have discrete boundaries; but if we were to morally evaluate our thoughts, we must also evaluate all the thoughts that lead to our purportedly bad thoughts. This seems implausible.

Sher's criticisms apply equally to actions. To say that it is wrong for Smith to kill Jones also means that it is wrong for Smith to do things outside of killing Jones but that will lead to killing Jones. For example, buying a gun, researching how to dispose of a body, etc., are all wrong. So, it seems that actions are just as permeable as thoughts. Since this is not problematic for actions, it should not be problematic for thoughts.

Sher might respond: if Smith buying a gun will not lead to him killing Jones, then buying the gun is not wrong. Only those actions that most likely would lead to the killing are wrong, not the ones leading up to it, because they do not cause the killing.[11]

I agree that if buying the gun has no causal connection to Smith killing Jones, then it is not wrong for Smith to buy the gun. But given Sher's reasoning for the permeability of thoughts, I believe that the problem applies equally to actions. Recall that Sher's reasoning for why the wrongness of thoughts can spread so easily while the wrongness of actions cannot: normal people "can easily resist

6 Sher, "A Wild West of the Mind," 492.

7 Sher, "A Wild West of the Mind," 492.

8 Sher, "A Wild West of the Mind," 492.

9 Sher, "A Wild West of the Mind," 492–93.

10 Sher, "A Wild West of the Mind," 493.

11 I am thankful to an anonymous referee for raising this objection.

the transition from the impulse to shoot you to the deed itself," but most normal people "have far less control over … [their] inferences and associations; so, to guard against thinking about shooting you, I would also have to avoid many other thoughts."[12] Sher's argument seems to rely on the claim that the jump from thought *X* to murderous thought *Y* is not within our voluntary control, while the jump from action *X* to murderous action *Y* is within our control. This explains how blame for thoughts can permeate very far. Would this same principle not be able to apply to some actions? Suppose that Smith knows that if he were to buy a gun, this would vastly increase the likelihood that he would kill Jones. Would it not be wrong for him to buy the gun? Also, it is otherwise permissible for Smith to drive down Jones's street, but if he does so knowing that this will fill him with uncontrollable rage, leading to him killing Jones, then this is wrong. Using Sher's reasoning, we can show that some actions are equally permeable to thoughts in terms of blame.

One might say that, in general, we have more control over the shift from thought to action than we do over the shift from thought to thought. While that may be true, the important takeaway is that this is not a unique problem for thoughts. For those links between thoughts that are voluntary, we can avoid the problem posed by Sher.[13]

Lastly, the proponent of my position can bite the bullet say that if we regard thought *X* as wrong, we should regard all thoughts that knowingly lead to *X* as being somewhat wrong. Like I said above, this seems like the right answer for otherwise permissible actions that knowingly lead to a bad action.

12 Sher, "A Wild West of the Mind," 492.

13 One might object that the jump from one action to the other has more barriers in place than the jump from one thought to the other. For example, as a referee pointed out to me, Smith (in the previous example) may have a lock on his door, he may see a police officer as he drives, etc. All of these are physical objects that will likely deter Smith from moving from action *X* to action *Y* (murder). It seems less clear that there are similar barriers for thoughts. This suggests that there is more control (and thus more responsibility) when Smith moves from action to action than when he goes from thought to thought. I am not convinced that this is always the case. There are many thoughts that can, so to speak, unlock doors in our minds. Suppose that Smith is a loving spouse who cares very much about fidelity. Smith never fantasizes about his attractive coworker, because he knows that if he were to start doing so, this would unlock a door in his mind to fantasizing about many other women in his life. In the same way that the lock on the door deters Smith from leaving the house, the concern for fidelity deters Smith from fantasizing. I would guess that the concern for fidelity is a more effective deterrent than a locked door. However, even if the reviewer is correct that this difference makes thoughts more permeable than actions, I think it would only show a difference in degree, not in kind. In principle, we can be responsible for thoughts that likely lead to other thoughts and actions that likely lead to other actions. But, it may be the case that this principle applies fewer times to thoughts than to actions.

The second feature of the mental that Sher points to is the fact that our mental lives are impermeable to and disconnected from other people. Sher argues that the mental and the public are different, in that "what is going on in each person's subjectivity is always independent of, and is often wildly at variance with, what is concurrently going on in his public neighborhood."[14] Given the "impermeability and independence of each person's subjectivity," Sher argues that "each subjectivity is almost literally a world of its own."[15] Sher claims that, in light of this gulf between our minds, our thoughts about other people are representations of those people that do not have moral standing. As he says, "the 'people' who populate our mental landscapes are only shadow people, and you can't have a moral obligation . . . to a shadow."[16]

This is questionable. In the next section, I argue that one's private thoughts can be intrinsically wrong and that they can affect the well-being of others, regardless of whether those thoughts are about shadow people.

3. Arguments against WWM

Here, I develop several objections to WWM that Sher does not consider.

3.1. What Is Special about Our Minds?

Consider the following cases:

> *Hate in the Head*: Smith, who hates Jews, privately thinks that the Holocaust was morally justified.
>
> *Hate in an Empty Room*: Smith, who hates Jews, is in an empty room, on top of an empty mountain, etc., and publicly states his belief that the Holocaust was morally justified.

The only difference between the cases is whether the hateful sentence is uttered out loud or merely in Smith's head. We can stipulate that, in both cases, nobody will find out about it, it will not make Smith more likely to do something bad, etc. Sher's view initially seems to suggest that there is a moral difference between these cases. That seems implausible.

Sher might respond that, in Hate in an Empty Room, Smith's words have the capacity to cause harm, while in Hate in the Head, Smith's thoughts lack such a capacity. But, this is not the case. I have stipulated that nobody will hear Smith.

14 Sher, "A Wild West of the Mind," 494.

15 Sher, "A Wild West of the Mind," 494.

16 Sher, "A Wild West of the Mind," 494.

We can imagine a version of Hate in the Head in which someone reads Smith's mind and is harmed by his thoughts. Thus, the capacity to harm is not a genuine difference between these cases.[17]

These cases suggest that the mental realm does no work in Sher's account. If the mental *qua* mental were doing any work, then we should regard the above cases as morally different. But there seems to be no reason for doing so. Thus, on Sher's view, the mental does no intrinsic work.

Contrast these theses:

> *Wild West of the Mind* (WWM): For any purely private thought, *T*, that is expressed in an agent's, *S*'s, mind, *T* is not morally wrong.
>
> *Permissibility of Unheard Utterances* (PUU): for any utterance, *U*, it is not wrong for *S* to utter *U*, in her mind or in speech, so long as nobody is ever aware that S uttered *U*.

Sher intends to endorse WWM. But if he cannot offer a principled reason for regarding Smith's action as permissible in Hate in the Head but wrong in Hate in an Empty Room, then he is really committed to PUU. In other words, when we press Sher's account, it seems that he cannot hold that our thoughts are not wrong *because they are in our mind* but rather must hold that our thoughts are not wrong *because others do not know what we are thinking about*. If so, Sher must revise his position to say that our utterances and feelings, mental or otherwise, are only wrong if they are made known to others.

This has two important implications. First, Sher's position, contrary to his opinion, is not about the mental at all. Second, as I will argue in the coming paragraphs, PUU is false. If Sher is committed to PUU, then his view is false.

Again, PUU says that our utterances, mental or verbal, are not wrong if others do not find out about them. We can see the implausibility of this thesis by considering the following cases:

> *Joke in an Empty Room*: Smith, who is anti-Semitic, says a horribly offensive Holocaust joke in an empty room.
>
> *Causal Impotence Hitler Vote*: Voting is a kind of utterance. Smith lives in Germany in 1933. Smith knows that Hitler will win the election and thus knows that he is causally impotent over the outcome. And Smith knows

17 One might wonder whether Sher's view entails that certain thoughts become impermissible only when someone is in an MRI machine and the contents of their thoughts can be discerned. This would be a strange implication.

> that nobody will ever find out about his vote. So, he votes for Hitler, because he hates Jews.

In both cases, we can stipulate that nobody will be directly harmed by Smith's action, and nobody will find out about Smith's actions. However, it seems intuitively clear that there is something morally wrong about both of these actions. The precise explanation of why these actions seem wrong, I contend, is that they involve an agent either endorsing an evil action or expressing and endorsing an evil belief. In Joke in an Empty Room, Smith expresses and endorses a morally repugnant belief, and in Causal Impotence Hitler Vote, Smith endorses an evil agent and his policies. This judgment can be summed up in the following thesis:

> *Endorsement*: It is *prima facie* wrong to endorse morally wrong ideas or agents.[18]

The Endorsement principle seems like the best explanation of my intuitions in the aforementioned cases. Additionally, a further intuitive defense can be offered in favor of Endorsement. Consider these cases:

> *Smith*: Smith is a typical person with typical beliefs, none of which are all that objectionable. He is generally nice to people in his life.
>
> *Nazi Jones*: Jones is identical in all respects to Smith, but he also harbors horribly anti-Semitic beliefs. Although he never acts on these beliefs, Jones essentially subscribes to the Nazi political ideology.

By stipulation, both Smith and Jones will be generally nice people for most of their lives. So, the only difference between them is that Jones is a closeted Nazi. Intuitively, it seems clear to me that Jones is a morally worse person than Smith.[19] The only explanation for this is that Jones endorses morally wrong ideas.

18 Corvino endorses a similar view ("Naughty Fantasies"). The argument for Endorsement is that it explains what I believe are clear intuitions in Joke in an Empty Room and Causal Impotence Hitler Vote; the defense of Endorsement does not rely on any reference to Unfelt Harms. The Endorsement principle helps to clarify something seemingly paradoxical in my view. If it is wrong to have a certain thought, then is it not wrong to hold the thought in one's head long enough to judge that it is wrong? According to the Endorsement principle, what is wrong is not the mere having of a thought in one's head but rather endorsement of its content. For example, if one were to write a paper about the wrong of rape, one would need to have the concept of rape in one's thoughts. This is permissible, because this person is not endorsing rape.

19 Sher's view is not about one's character, so I do not mean this example to be mainly about character. Instead, I mean that the fact that Nazi Jones has a worse character than Smith is best explained by the wrongness of Jones's mental actions. So, the point here concerns a conduct-based evaluation, not a primarily character-based one.

If this account is true, and if our intuitions in the above cases are correct, then not all unheard utterances are permissible. Thus, PUU is false. And, given that WWM reduces to PUU, WWM is false.[20]

One might object that the Endorsement principle seems to rely on the claim that our beliefs are within our control, which sounds false. Thus, the Endorsement principle never gets off the ground. I have several responses to this objection.[21]

First, although it is less commonly endorsed, one could be a doxastic voluntarist of a certain variety. As I see it, doxastic voluntarism is the view that agents exercise at least some control (either direct or indirect) over some of their beliefs. I think that it is beyond the scope of this paper to enter into the debate about the voluntariness of belief. But what I can say here is that if it turns out that my view requires doxastic voluntarism to be true, this would be a less common but still defensible position.[22]

Second, it seems clear to me that there is something *prima facie* wrong about endorsing morally wrong ideas or agents. Suppose that Smith had complete control over his political beliefs and he chose to be a Nazi. This would be wrong. Of course, nobody has direct control over their beliefs. But, what this example shows is that there can be something *prima facie* wrong about some endorsements. Now, modify the case to real life and add that Smith became a Nazi believer as a result of causes that he could not control. It seems like the wrongness of Smith's belief might be defeated by the fact that he could not control his beliefs. But the important insight is that there is wrongness there to be defeated, meaning that endorsements can be *prima facie* wrong.

Why does this matter? This would show that many of our thoughts that involve bad endorsements are *prima facie* wrong but that the wrongness is defeated. This would still run contrary to Sher's position. As I read Sher, he wants to argue that there is no moral valence *at all* with our thoughts. As he says, the mental is "morality-free."[23] Beyond this, my position would also be able to maintain that any endorsements we make that are, in some way, within our volun-

20 Sher might respond that there is still a difference between Hate in the Head and Hate in an Empty Room, namely that verbal utterances have clear boundaries, while thoughts do not. Perhaps the precursors to a private thought are much more difficult to map out, while the precursors to a public utterance are clear. To this, I again reply that I think thoughts and actions are just as porous. Just as I cannot perfectly map out all of the chemical reactions that lead to my thoughts, nor can I do so for actions. Also, actions stem from our mental life, meaning that the permeability of thoughts would have to apply to actions as well.

21 Thank you to an anonymous referee for bringing this point to my attention.

22 For classic arguments, see Steup, "Doxastic Voluntarism and Epistemic Deontology"; and Alston, *Epistemic Justification*.

23 Sher, "A Wild West of the Mind," 483.

tary control can be morally evaluated. I agree that many beliefs seem to not be under our direct control. But, many beliefs also seem to be under our indirect control. We can choose to look at (or ignore) evidence against our position, we can choose to seek out people who disagree (or choose not to), etc. These are all ways that we exercise some indirect control over what we believe. To the extent that Smith's Nazi beliefs and endorsements are the result of his indirect control, then he can be blamed for his endorsement of those beliefs.

3.2. Unfelt Harms

Here I argue, based on the possibility of unfelt harms, that Sher's view is false. Consider two cases:

> *Chris*: Chris lives a normal life, and all of his friends often think positively about him. Chris never finds about his friends' thoughts.
>
> *Alastair*: Alastair lives a similar life to Chris. But all his friends are constantly fantasizing about grotesque ways of killing him, stealing his money, etc. They will not ever do these things but they still fantasize about them. They never speak to each other about these fantasies, and Alastair never finds out about any of this. Alastair desires that his friends think well of him.

Intuitively, it seems that Alastair's life is going worse for him than Chris's. It does not seem to matter that neither of them will find out about their friends' thoughts. Even with that stipulation in place, it seems obvious that Chris's life is going better than Alastair's.

The intuition that Alastair's life is going worse than Chris's can be explained by the concept of unfelt harm. Many philosophers have defended the view that individuals can be harmed by actions that never affect their subjective experience.[24] Boonin advances a sustained argument for the possibility of unfelt harms. Although I lack space to outline Boonin's full argument, it is motivated by cases like this:

> *Adultery*: Bob wants his marriage to Carol to be monogamous and he believes that it is, but in fact Carol cheats on him regularly.[25]

Most people have the intuition that Bob is being harmed by Carol's adultery, even if he never finds out about it. Or, as Boonin puts it, "if Carol's acts really are

24 Several authors have defended the possibility of posthumous harm: see Feinberg, *The Moral Limits of the Criminal Law*; Parfit, *Reasons and Persons*; Pitcher, "The Misfortunes of the Dead"; and Boonin, *Dead Wrong*.

25 Boonin, *Dead Wrong*, 17.

harming Bob despite the fact that her acts are having no effect on Bob's mental states … then Carol is inflicting unfelt harm on Bob."[26] I lack sufficient space to launch a full-scale defense of the unfelt harm position, but I take it to be *prima facie* intuitive that, in cases like Adultery, unfelt harm is occurring. At the very least, the objector to this position must say that Bob is not being harmed at all and that our intuitions are being misled by something. This would be a *prima facie* counterintuitive position.[27]

If unfelt harms are possible, then Alastair is being harmed by his friends' thoughts. If Alastair can be harmed without knowing it, and as long as we agree that the frustration of our desires can be harmful to us, then it follows that Alastair is being harmed by his friends' thoughts.[28] Alastair desires that his friends not engage in fantasies about killing him. Thus, when his friends engage in these fantasies, they frustrate his desires and harm him. Since well-being and harm are moral concepts, it follows that our private thoughts can be wrong in virtue of causing someone to have less well-being and to be in a harmed state.

One might immediately worry that even if we agree that unfelt harms are genuine harms, it needs to be argued that they can constitute wrongs. This is especially important, given that many of our thoughts are not within our control even if they are harmful.[29] My argument that the mental is morally laden is not meant to include involuntary thoughts. For example, intrusive thoughts, sudden thoughts, images popping into one's head, etc., would count as involuntary. On my view, these thoughts may be harmful in some sense, but the harm does not rise to the level of a wrong because the agent has an excuse—namely, that the thought was not within her control. This is especially important for people with certain mental disorders, such as obsessive-compulsive disorder, where the agent feels excessive responsibility and guilt over her thoughts. For such individuals,

26 Boonin, *Dead Wrong*, 20.

27 One might wonder how the adultery case, which involves an action inflicting an unfelt harm, can be used to support the claim that a thought can inflict unfelt harm. The point of the adultery case is to illustrate that something that an agent does not know can still harm him. If I am correct that mental and verbal utterances are morally on a par, and if verbal utterances are actions, then it would follow that both external utterances and thoughts should be considered in the same category as physical acts. And, if physical acts can inflict unfelt harm, then external verbal speech acts and thoughts should be treated in the same way. Put more simply, as long as we agree that something an agent does not know about can harm him, I do not see a reason to think that unfelt actions inflict harm while unfelt thoughts or utterances cannot.

28 I am not claiming that desire frustration is *always* bad for us. My argument does not rely on such a strong claim. But everyone agrees that the frustration of some clearly reasonable desires harms us, and Alastair's desire for his friends to think well of him is a reasonable desire.

29 Sher's argument does not rely on the claim that our thoughts are not within our control.

the realization that the majority of their thoughts are not within their control is freeing. I am only concerned with voluntary thoughts. Although many thoughts are not voluntary, the relevant ones are. I want to discuss thought categories like indulged fantasies, prolonged voluntary daydreaming, etc. Any time an agent gets a thought and chooses to indulge it and follow it is a voluntary thought.

With that clarification, we can now assess the move from unfelt harm to unfelt wrong in the world of thoughts. The basic argument can be made in two ways. First, if we agree that unfelt harm is a genuine harm, and if we agree that harms of any kind are *prima facie* wrong, then we should agree than unfelt harms are *prima facie* wrong. The burden of proof shifts to the objector to say why unfelt harms are genuine harms but cannot move into the realm of being wrongs. Second, there are clear examples of unfelt wrongs. For example, suppose that my neighbor watches me shower every day through the window from his house. Given that he is on his property, he has not trespassed onto my land. He has done nothing that affects my subjective experience. Thus, the best explanation of why this peeping Tom is acting wrongly is some kind of unfelt harm. The burden would then be on the objector to explain this case as wrongful without reference to unfelt harm. One might object that, in this case, the wrong can be explained by a privacy rights violation instead of unfelt harm. Still, it seems clear to me that the act is not just violating my rights, it also harms me. And we can devise a case of an unfelt wrong that does not have this feature. For example, in the previous case of Alastair, it seems wrong for his friends to be constantly gossiping about him, but he has no right against this. Thus, unfelt harms leading to an unfelt wrong best explain the wrongness of their gossip.[30]

Sher might respond by saying that the subject of the harm is not really

30 One might object that my examples jump between being examples of unfelt harms and unfelt rights violations. But I have tried to lump those into one category and infer from unfelt rights violations to unfelt wrongs. However, it may be the case that unfelt rights violations are harmful and thus wrong, while unfelt harms (which do not violate anyone's rights) do not count as genuine harms. I have several responses to this objection. First, whether it is an unfelt rights violation or an unfelt harm, it still seems clear that, from the third-person perspective, Alastair's life is going worse than Chris's. Perhaps unfelt rights impose a more substantial wrong on an agent. But from the external point of view, the unfelt harm directed at Alastair (even if he has no right against it) makes his life go worse. To put it differently, his life would be going better for him if his friends were not gossiping about him. Both Boonin and Pitcher make a similar move in their defenses of unfelt harms. Second, it seems arbitrary to me to say that unfelt harms do not genuinely wrong an agent while unfelt rights violations do. I agree that unfelt rights violations might be worse, but they are not different in any relevant ways. Both unfelt harms and unfelt rights violations are unknown to the agent, do not affect her subjective experience, etc. It seems strange to say that one is morally relevant while the other is not.

Alastair; instead, shadow-Alastair is affected by the gossip. And, harms against shadow-Alastair do not matter. Suppose that an author needlessly kills her main character. This may be unnecessary, but it would be odd to say that it is wrong, because the character is not real. Shadow-Alastair is no different from this fictional character. Thus, we should not be concerned with harms done to shadow-Alastair. In a sense, this objection would allow Sher to agree with the possibility of unfelt harms while still responding to my objection, because he would be saying that the unfelt harm is taking place but is directed at a shadow person, whose interests are not morally relevant. While I see the motivation for this view, I am skeptical of the ontological commitment involved in positing shadow people. Where do they exist? Do they come into existence whenever I conceive of them? Or do we have a certain number of preexisting shadow selves that exist? But suppose that Sher had a good response to these worries—I still believe that the unfelt harms objection succeeds. If we were to ask Alastair's friends to whom they are directing their comments, they would say that they mean them in reference to the real Alastair, not to shadow-Alastair. If Sher is to claim that the gossipers' comments are actually addressed at shadow-Alastair, then he is committed to a highly revisionary view about how these speakers use language. In other words, for Sher's view to succeed, the gossipers would have to be mistaken about whom they are referring to as the object of the gossip. My view maintains that they are referring to exactly the person they claim to be referring to.

3.3. *Action-Thought Composites*

Sher is keen to point out that his view is only about purely private thoughts, not about thought-act composites. A purely private thought is one that never leads to a corresponding action, while a thought-act composite is a thought that does lead to an action. Recall the cases of Smith and Jones. In one case, Smith pushes Jones with the intention of saving his life, and in the other case, Smith pushes Jones in front of a train with the intention of killing him. Smith's thoughts clearly make a moral difference. One might be inclined to look at the cases of Smith and Jones and say that this is evidence that our thoughts have a moral valence. After all, the difference between an attempted murder and an attempted lifesaving is Smith's intentions. This might lead one to say that Smith's murderous thoughts are wrong, even if they are never put into practice. Sher denies this; as he says,

> to warrant condemnation for maliciously injuring someone, a person must not only harbor the malice but actually inflict the injury. Thus, taken by itself, the claim that it is wrong to give public expression to private

> malice does not imply that there is anything wrong with simply entertaining the malice.[31]

This is counterintuitive. Sher is saying that a malicious thought is not wrong on its own. But the conjunction of a malicious thought and a malicious action is wrong. However, Sher presumably wants to say that the conjunction of a malicious thought and a malicious action is worse than just a malicious action. If Smith and Jones (who is Jewish) are boxing, and Smith hits Jones because he wants to win and because he hates Jewish people, this is clearly worse than Wilson (who is not anti-Semitic) hitting Jones during a boxing match only because he wants to win. So, Sher seems to be committed to the following position: the conjunction of a malicious thought and a malicious action is worse than just a malicious action, but the malicious thought is not wrong on its own. This is puzzling; if the malicious thought is not wrong on its own, then how does it add any wrongness to the malicious action? Sher seems to claim that a malicious thought, which is not independently wrong, somehow contributes to the wrongness of malicious thought-action composites.[32] This view violates the following principle about wrongness:

> *No New Wrongness*: If an action, *X*, is not wrong, it cannot make some further action, *Y* (to which it is conjoined), more wrong than *Y* already was.

This principle seems hard to deny. Yet, Sher is committed to denying it, which puts him in a counterintuitive position. Either he must admit that his view extends to both purely private thoughts and thought-act composites (which he does not want to do), or he must offer an account of the seemingly magical emergent wrongness of malicious thoughts that comes into being when they are added to malicious actions.[33]

Sher could object that No New Wrongness is false on the grounds that an

31 Sher, "A Wild West of the Mind," 483.

32 One might object that there is no such thing as a malicious action without a malicious thought. While I agree in a sense that some actions depend constitutively for their wrongness on the motive behind them, there are certainly still bad actions that have no thoughts attached to them. Killing someone without malice is still wrong. So, it seems perfectly reasonable to claim that there can be wrong actions independent of the thoughts that connect with them. But the criticism above would not apply to actions like lying, because whether something is a lie or not depends on the intentions of the speaker. But not all actions, or even most, are this way.

33 Consider the following potential counterexample: Smith swings a bat. This is not wrong. Now, Smith swings the bat at the same speed and intentionally hits a child. Here, an action that was not wrong is added to another action and increases the wrongness of that action. This is not a genuine counterexample. The two actions are *X* (swinging a bat) and *Y* (hitting

organic unity is formed whenever a malicious thought and action are combined. In other words, even if there is nothing wrong about malicious thoughts, there might be something wrong in the state of affairs that combines malicious thoughts and actions, and this wrongness could be more than the wrongness of the malicious action on its own. While this would be a possible response, it does not seem more plausible than my view. What is gained by saying that an organic unity is formed when a malicious thought and intention are added together instead of saying that the thought is independently wrong? While I am not skeptical of organic unities, it seems too convenient to posit them here. At the very least, they should be avoided when possible, and I have provided a way to avoid an organic unity here.

One might object that this argument misunderstands the connection between wrongfulness and the agent's mental states at the time of action. It could be that the person who unknowingly pushes Jones to his death is acting wrongly but is not blameworthy, due to ignorance, etc. Perhaps the agent's mental states are relevant to judgments of culpability but not relevant to judgments of wrongness. Thus, my argument in this section may rest on a false presupposition about how mental states and actions interact in terms of moral evaluation.

Although it is controversial, I am sympathetic to subjectivism about moral obligations, which I understand to be the view that actions are right or wrong depending on what the agent believes at the time of action. If this view is true, then attributions of culpability and wrongness go hand in hand. On other days of the week, I am sympathetic to the ambiguity view, which I understand to be the view that "right" and "ought" can have both objective and subjective senses. So, on this view, there is some sense of rightness that goes with culpability and some sense that tracks the objective facts of the situation. All of this is to say that I think there is good reason to connect wrongfulness and culpability.[34]

But I realize that the preceding is controversial. Given this, I think my view can succeed even if wrongness and culpability are completely separate. If this were the case, I could reword the No New Wrongness principle to be about culpability instead. It would then read: if an action, *X*, is not blameworthy, it cannot make some further action, *Y* (to which it is conjuncted), more blameworthy than *Y* already was. I think the same problem for Sher's view could be generated at the level of blameworthiness for thoughts, even if wrongness ends up not being the

a child). The wrongness of *Y* is not increased by *X*. Hitting a child is what makes *Y* wrong, not the method by which the hitting occurs.

34 For a helpful discussion of this debate, see Mason, "Objectivism and Prospectivism about Rightness"; Graham, "In Defense of Objectivism about Moral Obligation"; and Zimmerman, "Is Moral Obligation Objective or Subjective?"

correct label. Sher would then need to explain how a thought that was not blameworthy suddenly becomes blameworthy when it is added to a wrong action.

4. CONCLUSION

Sher's argument for WWM fails. Contrary to Sher's view, the mind is not a wild west. There is a moral sheriff who governs our thoughts.[35]

Florida Atlantic University
sdirector@fau.edu

REFERENCES

Alston, William P. *Epistemic Justification: Essays in the Theory of Knowledge*. Ithaca, NY: Cornell University Press, 1989.

Boonin, David. *Dead Wrong: The Ethics of Posthumous Harm*. Oxford: Oxford University Press, 2019.

Corvino, John. "Naughty Fantasies." *Southwest Philosophy Review* 18, no. 1 (January 2002): 213–20.

Feinberg, Joel. *The Moral Limits of the Criminal Law*. Vol. 1, *Harm to Others*. Oxford: Oxford University Press, 1984.

Graham, Peter A. "In Defense of Objectivism about Moral Obligation." *Ethics* 121, no. 1 (October 2010): 88–115.

Mason, Elinor. "Objectivism and Prospectivism about Rightness." *Journal of Ethics and Social Philosophy* 7, no. 2 (March 2013): 2–19.

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

Pitcher, George. "The Misfortunes of the Dead." *American Philosophical Quarterly* 21, no 2 (April 1984): 183–88.

Sher, George. "A Wild West of the Mind." *Australasian Journal of Philosophy* 97, no. 3 (2019): 483–96.

Steup, Matthias. "Doxastic Voluntarism and Epistemic Deontology." *Acta Analytica* 15 (2000): 25–56.

Zimmerman, Michael. "Is Moral Obligation Objective or Subjective?" *Utilitas* 18, no. 4 (December 2006): 329–61.

35 I would like to thank George Sher, David Boonin, Chris Freiman, Chris Heathwood, Alex Zambrano, and an anonymous reviewer from JESP for their comments on this paper. Thanks are also due to Emily Director for her continuing support, mentally and otherwise.