Essays on Suffering-Focused Ethics

Magnus Vinding

Ratio Ethica

Contents

Research vs. non-research work to improve the world: In defense of more research and reflection

S-risk impact distribution is double-tailed

Beware underestimating the probability of very bad outcomes: Historical examples against future optimism

Radical uncertainty about outcomes need not imply (similarly) radical uncertainty about strategies

Some pitfalls of utilitarianism

Distrusting salience: Keeping unseen urgencies in mind

Popular views of population ethics imply a priority on preventing worst-case outcomes

Other Resources on Suffering-Focused Ethics

Preface

The essays found in this volume have all been published before as standalone pieces. My reasons for publishing them in a book are partly to extend their reach, by making them more widely available, and partly to better preserve them, in case some of them become inaccessible elsewhere.

Most of the essays in this collection were written and published after I wrote and published my book *Suffering-Focused Ethics: Defense and Implications* (2020), and many of them thus serve as extensions to that book, often going into greater detail on some of the key issues treated there.

All of the essays have to do with the reduction of suffering in one way or another. The essays in **Part I** mostly explore theoretical issues relating to the justification for suffering-focused moral views. The essays in **Part II** reply to various critiques of suffering-focused views. Finally, the essays in **Part III** are about practical issues concerning how we can best reduce suffering. (Of course, the distinction between theoretical and practical issues is by no means clear-cut, as exemplified by the essay "Lexical priority to extreme suffering — in practice", which in some sense bridges the two levels.)

Each essay can be read independently, and hence the essays can be read in any order whatsoever. The independent nature of the pieces means that there will be some repetition between them — e.g. the same idea might be introduced a couple of times. However, the essays still mostly complement each other, and I believe that they each provide some useful contributions.

It is my hope that these essays can help motivate further work on suffering-focused ethics, and not least that they can help motivate real-world efforts to reduce suffering in effective ways.

Magnus Vinding
Copenhagen
December 2022

Part I: Theoretical Issues

Why I used to consider the absence of sentience tragic

Whether one considers the absence of sentience bad or neutral — or indeed as good as can be — will tend to matter a lot for one's ethical and altruistic priorities. Specifically, it can have significant implications for whether one should push for smaller or larger future populations.

I used to be a classical utilitarian. That is to say, I used to agree with the statement "we ought to maximize the net amount of happiness minus suffering in the world". And given this view, I found it a direct, yet counterintuitive implication that the absence of sentience is tragic, and something that we ought to minimize by bringing about a maximally large, maximally happy population. My aim in this essay is to briefly present what I consider the main reason why I used to believe this, and also to explain why I no longer hold this view. I am not claiming that the reasons I had for endorsing my past view are shared by other classical utilitarians, yet I suspect they could be, at least by some.

The Reason: Striving for Consistency

My view that the absence of sentience is tragic and something that we ought to prevent mostly derived, I believe, from a wish to be consistent. Given the ostensibly reasonable view that death is bad, it would seem to follow, I reasoned, that since death merely amounts to a discontinuation of life — or, seen in a larger perspective, a reduction of the net amount of sentience — the reduction of sentience caused by not giving birth to a new happy life should be considered just as bad as the end of a happy life. This was counterintuitive, of course, yet I did not, and still do not, consider immediate intuitions to be the highest arbiters of moral wisdom, and so it did not seem that weird to accept this conclusion. The alternative, if I were to be consistent, would be to bring my view of death in line with my intuition that the absence of sentience is not bad. Yet this was too implausible, since death surely *is* bad.

This, I believe, was the reasoning behind my endorsing a moral obligation to produce a large, happy population. To not create such a large population would, in some ways, be the moral equivalent of committing genocide. My view is quite different now, however.

My Current View of My Past View

I now view this past reasoning of mine as akin to a deceptive trick, like a math riddle where one has to find where the error was made in a series of seemingly valid deductions. You accept that death is

tragic. Death means less sentient life than continued life, other things being equal. But a failure to bring a new individual into the world also means less sentient life, other things being equal. So why would you not consider a failure to bring an individual into the world tragic as well?

My current response to this line of reasoning is that death indeed is bad, but that it is not *intrinsically* bad. What is bad about death, I would argue, is the suffering and preference frustration that it involves, not the discontinuation of sentience per se (after all, a discontinuation of sentience occurs every night we go to sleep, which we rarely consider bad, much less tragic). This view is perfectly consistent with the view that it is not tragic to fail to create a new individual. Unlike the death of an existing person, the non-creation of a new person does not involve suffering, preference frustration, uncompleted life projects, and so on for the uncreated person.

Narrative self-deception: The ultimate elephant in the brain?

"*the elephant in the* **brain**, n. An important but unacknowledged feature of how our minds work; an introspective taboo."

The Elephant in the Brain is an informative and well-written book, co-authored by Kevin Simler and Robin Hanson. It explains why much of our behavior is driven by unflattering, hidden motives, as well as why our minds are built to be unaware of these motives. In short: because a mind that is ignorant about what drives it and how it works is often more capable of achieving the aims that it was built to achieve.

The book also seeks to apply this knowledge, to shed some light on the hidden motives of many of our social institution. Rather than being about high-minded ideals, our institutions often serve much less pretty, more status-driven purposes, such as showing off in various ways, as well as to help us better get by in a tough world.

All in all, I think *The Elephant in the Brain* provides a strong case for supplementing one's mental toolkit with a new, important tool, namely to continuously ask: How might my mind skillfully be avoiding confrontation with ugly truths about myself that I would prefer not to face? And how might such unflattering truths explain aspects of our public institutions and public life in general?

This is an important lesson, I think, and it makes the book more than worth reading. At the same time, I cannot help but feel that the book ultimately falls short when it comes to putting this tool to proper use. For the main critique that came to my mind while reading the book was that it seemed to ignore the biggest elephant in the brain by far — the elephant I suspect we would all prefer to ignore the most — and hence it failed, in my view, to take a truly deep and courageous look at the human condition. In fact, the book even seemed be a mouthpiece for this great elephant.

The great elephant I have in mind here is a tacitly embraced sentiment that goes something like: life is great, and we are accomplishing something worthwhile. As the authors write "life, for must of us, is pretty good", and they end the book on a similar note:

> In the end, our motives were less important than what we managed to achieve by them. We may be competitive social animals, self-interested and self-deceived, but we cooperated our way to the god-damned moon.

This seems to implicitly assume that what humans have managed to achieve, such as cooperating (i.e. two superpowers with nuclear weapons pointed at each other competing) their way to the moon, has been worthwhile all things considered. Might this, however, be a flippant elephant talking, rather than, say, a conclusion derived via a serious analysis of our condition?

The fact that people often get offended and become defensive when one even just questions the value of our condition — and sometimes also accuse the one raising the question of having a mental illness — suggests that we may indeed be disturbing a great elephant here; something we would strongly prefer not to think too deeply about.

It is important to note here that one should not confuse the cynicism required for honest exploration of the human condition with misanthropy, as Simler and Hanson themselves are careful to point out:

> The line between cynicism and misanthropy—between thinking ill of human *motives* and thinking ill of *humans*—is often blurry. So we want readers to understand that although we may often be skeptical of human motives, we love human beings. (Indeed, many of our best friends are human!) [...] All in all, we doubt an honest exploration will detract much from our affection for [humans]. (p. 13)

Similarly, an honest and hard-nosed effort to assess the value of human life and the human endeavor need not lead us to have less compassion for humans. Indeed, it might lead us to have much more compassion for each other.

Is Life "Pretty Good"?

With respect to Simler and Hanson's claim that "life, for must of us, is pretty good", it can be disputed whether this is indeed the case. According to the 2017 World Happiness Report, most people rated their life satisfaction at five or below on a scale from zero to ten, which arguably does not translate to being "pretty good". Indeed, one can argue that the scale employed in this report is biased, in that it does not allow for a negative evaluation of life.

But even if we were to concede that most people say that their lives are pretty good, one can still reasonably question whether most people's lives indeed *are* pretty good, and not least question whether such reports imply that the human condition is worthwhile in a broader sense.

Narrative Self-Deception: Is Life As Good As We Think?

Just as it is possible for us to be wrong about our own motives, as Simler and Hanson convincingly argue, could it be that we can also be wrong about how good our lives are? Furthermore, could it be

that we not only *can* be wrong but that most of us in fact *are* wrong about it most of the time? This is indeed what some philosophers argue, seemingly supported by psychological evidence.

One philosopher who has argued along these lines is Thomas Metzinger. In his essay "Suffering", Metzinger reports on a pilot study he conducted in which students were asked at random times via their cell phones whether they would relive the experience they had just before their phone vibrated. The results were that, on average, students reported that their experience was not worth reliving 72 percent of the time. Metzinger uses this data, which he admits does not count as significant, as a starting point for a discussion on how our narrative about the quality of our lives might be out of touch with the reality of our felt, moment-to-moment experience:

> If, on the finest introspective level of phenomenological granularity that is functionally available to it, a self-conscious system would discover too many negatively valenced moments, then this discovery might paralyse it and prevent it from procreating. If the human organism would not repeat most individual conscious moments if it had any choice, then the logic of psychological evolution mandates concealment of the fact from the self-modelling system caught on the hedonic treadmill. It would be an advantage if insights into the deep structure of its own mind – insights of the type just sketched – were not reflected in its conscious self-model too strongly, and if it suffered from a robust version of optimism bias. Perhaps it is exactly the main function of the human self-model's higher levels to drive the organism continuously forward, to generate a functionally adequate form of self-deception glossing over everyday life's ugly details by developing a grandiose and unrealistically optimistic inner story – a "narrative self-model" with which we can identify?

Metzinger continues to conjecture that we might be subject to what he calls "narrative self-deception" — a self-distracting strategy that keeps us from getting a realistic view of the quality and prospects of our lives:

> a strategy of flexible, dynamic self-representation across a hierarchy of timescales could have a causal effect in continuously remotivating the self-conscious organism, systematically distracting it from the potential insight that the life of an anti-entropic system is one big uphill battle, a strenuous affair with minimal prospect of enduring success. Let us call this speculative hypothesis "narrative self-deception".

If this holds true, such self-deception would seem to more than satisfy the definition of an elephant in the brain in Simler and Hanson's sense: "an important but unacknowledged feature of how our minds work; an introspective taboo."

To paraphrase Metzinger: the mere fact that we find life to be "pretty good" when we evaluate it from the vantage point of a single moment does not mean that we in fact find most of our experiences "pretty good", or indeed even worth (re)living most of the time, moment-to-moment. Our single-moment evaluations of the quality of the whole thing may well tend to be gross, self-deceived overestimates. And recent studies suggest that this is indeed the case.

Another philosopher who makes a similar case is David Benatar, who in his book *Better Never to Have Been* argues that we tend to overestimate the quality of our lives due to well-documented psychological biases:

> The first, most general and most influential of these psychological phenomena is what some have called the Pollyanna Principle, a tendency towards optimism. This manifests in many ways. First, there is an inclination to recall positive rather than negative experiences. For example, when asked to recall events from throughout their lives, subjects in a number of studies listed a much greater number of positive than negative experiences. This selective recall distorts our judgement of how well our lives have gone so far. It is not only assessments of our past that are biased, but also our projections or expectations about the future. We tend to have an exaggerated view of how good things will be. The Pollyannaism typical of recall and projection is also characteristic of subjective judgements about current and overall well-being. Many studies have consistently shown that self-assessments of well-being are markedly skewed toward the positive end of the spectrum.

Is "Pretty Good" Good Enough?

Beyond doubting whether most people would say that their lives are "pretty good", and beyond doubting that a single moment's assessment of one's quality of life actually reflects this quality all that well, one can also question whether a life that is rated as "pretty good", even in the vast majority of moments, is indeed good enough to render it worth starting for its own sake.

This is, for example, not necessarily the case on tranquilist or antifrustrationist views of value, according to which experiential wellbeing consists of the absence of suffering or preference frustrations. Similar to Metzinger's point about narrative self-deception, one can argue that, if tranquilist or antifrustrationist views happen to be plausible views of the value of our experiences (upon closer inspection), we should probably expect to be quite blind or resistant to this fact. And interesting to note in this context is that many of the traditions that have placed a strong emphasis on paying attention to our direct experience, including some strands of Buddhism, seem to have converged on views very similar to tranquilism and antifrustrationism.

Can the Good Lives Outweigh the Bad?

One can also question the value of our condition on a more collective level, by focusing not only on a single (self-reportedly) "pretty good" life, but on *all* individual lives. In particular, we can question whether the good lives of some can justify the miserable lives of others.

A story that gives many people pause on this question is Ursula K. Le Guin's *The Ones Who Walk Away from Omelas*. The story is about a near-paradisiacal city in which everyone lives deeply meaningful and fulfilling lives — that is, everyone except a single child who is locked in a basement room, forced to live a life of squalor:

> The child used to scream for help at night, and cry a good deal, but now it only makes a kind of whining, "eh-haa, eh-haa," and it speaks less and less often. It is so thin there are no calves to its legs; its belly protrudes; it lives on a half-bowl of corn meal and grease a day. It is naked. Its buttocks and thighs are a mass of festered sores, as it sits in its own excrement continually.

The story's premise is that this child must exist in this condition for the happy people of Omelas to enjoy their lives, which then raises the question of whether the enjoyment found in these lives can morally outweigh and justify the misery of this single child. Some citizens of Omelas seem to decide that this is not the case: the ones who walk away from Omelas.

Sadly, our world is much worse than the city of Omelas on every measure. For example, in the World Happiness Report cited above, around 200 million people reported their quality of life to be in the absolute worst category. If the story of Omelas gives us pause, we should also think twice before claiming that the "pretty good" lives of some people can outweigh the self-reportedly very bad lives of these hundreds of millions of people, many of whom decide to end their own lives by suicide.

Beyond that, one can question whether the "pretty good" lives of some humans can in any sense outweigh or justify the enormous amount of suffering humanity that imposes on non-human animals, including the torturous suffering we impose on more than a trillion fish each year, as well as the suffering that we impose upon the tens of billions of chickens and turkeys who live out their lives under the horrific conditions of factory farming, many of whom end their lives by being boiled alive.

My aim in this essay has not been to draw any conclusions about the value of our condition. Rather, my aim has been to argue that we likely have an elephant in our brain that leads us to evaluate our lives, individually as well as collectively, in overoptimistic terms, and to ignore the many

considerations that might suggest a negative conclusion. This is an elephant that pushes us toward the conclusion that "it's all pretty good and worthwhile", and which disposes us to flinch away from serious, sober-minded engagement with questions concerning the value of our condition, including whether it would be better if there had been no sentient beings at all.

On purported positive goods “outweighing” suffering

Summary

Many moral views hold that purported positive goods, such as pleasure, can morally “outweigh” or “cancel out” suffering. Yet this notion of outweighing is more problematic than is commonly recognized, since it is not obvious in what sense such outweighing is supposed to obtain, nor what justifies it. Clarifying and justifying this notion of “outweighing” is thus a problem facing the moral views that rely on it. In contrast, strongly suffering-focused views, and harm-focused views more generally, do not face this problem.

Introduction

The premise that suffering can always, at least in principle, be outweighed by pleasure is entailed by moral theories such as classical utilitarianism and some other positive consequentialist views. Yet defenders of these views rarely provide an elaborate defense of this premise. For example, as far as I can tell, little is said to justify this premise in seminal defenses of classical utilitarianism, such as Bentham, 1789; Mill, 1863; and Sidgwick, 1874, nor in more modern defenses, such as Hewitt, 2008 and Lazari-Radek & Singer, 2014.

This is quite a glaring omission, since many philosophers have argued against the premise that suffering can (always) be outweighed by pleasure, and have done so in different ways.

(Note that the adoption of a suffering-focused ethic is not predicated on the view that pleasure can never outweigh suffering; there are many other views and arguments that can support the view that suffering deserves special priority, Gloor, 2016; Vinding, 2020, part I.)

Views that reject outweighing

Happiness as an incommensurate good

Philosopher Clark Wolf defends the view that happiness and suffering have positive and negative value, respectively, but rejects the view that “pleasures and pains can cancel one another out in the way that classical utilitarians usually assume” (Wolf, 1997, sec. I). In Wolf’s view, pleasure is not a truly opposite counterpart to suffering, and hence it cannot “cancel out” or “make up for” suffering, even as he maintains that pleasure does have intrinsic positive value (Wolf, 1997).

Negating interpersonal compensation

A position with similar implications is the view that suffering can be outweighed by pleasure *intra*personally, but not *inter*personally. This view is defended in Ryder, 2011, ch. 3; Harnad, 2016. Arguments that support the view that pleasure cannot outweigh suffering interpersonally can also be found in Vinding, 2020, ch. 3.

Epicurean and Buddhist axiologies

Views that hold that the value of happiness lies chiefly in its absence of negative features are another class of views that reject outweighing (Vinding, 2020, ch. 2; Ajantaival, 2021/2022). Such views have been endorsed by Epicurus, Arthur Schopenhauer, and William James in the West (Vinding, 2020, ch. 2), as well as by strands of Buddhism in the East (Breyer, 2015). Similar views been defended in recent times in Gloor, 2016; 2017; Sherman, 2017; Knutsson, 2019.

Unlike the interpersonal-only asymmetry defended by Ryder and Harnad, variations of this view also tend to imply an intrapersonal asymmetry. For example, Schopenhauer explicitly maintained that his "present well-being" could not "undo [his] previous sufferings" (Schopenhauer, 1819, vol II, p. 576). In Schopenhauer's view (ibid.),

> it is quite superfluous to dispute whether there is more good or evil in the world; for the mere existence of evil decides the matter, since evil can never be wiped off, and consequently can never be balanced, by the good that exists along with or after it.

Ethics as being about problems

Schopenhauer's view seems closely related to the view that ethics is about solving problems (see e.g. Gloor, 2016; Fehige, 1998). A way to justify this view may be to argue that only the existence of such problematic states imply genuine victims, while failures to create supposed positive goods (whose absence leaves nobody troubled) do not imply any real victims — such "failures" are mere victimless "crimes".

According to this view, we cannot meaningfully "cancel out" or "undo" a problematic state found somewhere by creating some unproblematic state elsewhere. The problematic states in question need not be limited to suffering. Negative consequentialist views that see ethics as being about solving problems may be concerned about problematic states more generally, such as injustice, preference frustration, and premature death (cf. Animal Ethics, 2012).

Lexical views

Another class of views are those that maintain that the most extreme forms of suffering have greater moral importance than anything else (see e.g. Mayerfeld, 1999, pp. 178-179; Leighton, 2011, ch. 9; Tomasik, 2015; Gloor, 2016; Vinding, 2020, ch. 4-5). Such views can be compatible with views that say that suffering can sometimes be outweighed by pleasure, as well as with views that say that it never can. But in either case, these lexical views still contradict the premise that suffering can *always* be outweighed by pleasure, and they thus constitute another important set of views to contend with for those who defend that premise.

Problematic cases

The following problematic cases are worth considering in relation to the view that suffering can always be outweighed by pleasure.

Happy sadists

There are various versions of this thought experiment. One version is to consider torture in the Colosseum: one individual is tortured horrifically for the enjoyment of a large crowd. According to the premise of outweighing mentioned above, the torture can be justified if the resulting pleasure of the crowd is sufficiently large (cf. Scarre, 1996, p. 156).

Suffering deemed unbearable and irredeemable

A related case to consider is that of suffering that the sufferer, while experiencing it, considers unbearable and impossible to outweigh (Tomasik, 2015; Vinding, 2020, ch. 4).

Two issues to address

The cases above serve as a good starting point for understanding and discussing the two principal issues that face the notion that pleasure can always, at least in principle, outweigh suffering.

I. Clarification

In *what sense* is pleasure supposed to be able to outweigh suffering? (Some relevant considerations on this can be found in Knutsson, 2016.) In particular, in what sense is the pleasure of many people supposed to outweigh the torturous suffering of a single individual when that individual considers their own suffering to be unbearable and unoutweighable?

II. Justification

Relatedly, what justifies the notion that extreme suffering can be outweighed by pleasure?

These are anything but trivial questions. In fact, they require elaborate explanation. And, to echo a remark I have made elsewhere, the assumption that suffering can always (in principle) be outweighed by pleasure cannot simply be considered plausible by default, especially given its controversial status and the many arguments that have been made against it (for an overview of these, see Vinding, 2020, part I).

Strongly suffering-focused views do not share this problem

It is worth noting that strongly suffering-focused views — and strong negative consequentialist views more generally — do not entail this notion that some states of the world can "outweigh" or "cancel out" bad states elsewhere. To be sure, suffering-focused views do tend to hold that some states of suffering can be deemed *worse*, and hence more deserving of priority, than some other states of suffering. Yet this is a fundamentally different claim, as it does not involve any outweighing in the sense of thinking that suffering, including extreme suffering in particular, can be "cancelled out" or "made up for" by different states elsewhere.

There are, of course, also problems when it comes to judgments about which states are worst and most deserving of priority. But it is worth noting that, first, it is quite uncontroversial that we often can and should make such judgments. For example, everyone agrees that it makes sense to triage in a way that favors patients undergoing intense suffering over patients with relatively minor ailments.

Second, we should note that positive consequentialist views, including classical utilitarianism, all happen to share this problem, and actually do so on an additional level. For not only do such views agree that some suffering can be worse than some other suffering, but they further endorse a corresponding claim about pleasure: that some forms of pleasure are more morally important to bring about than others, and that it is morally important to realize greater forms of pleasure in the first place, even when there are no beings who desire these pleasures.

This latter contention about the importance of increasing pleasure is significantly more controversial than the claim that it is important to prevent (worse forms of) suffering. As Daniel Kahneman notes (Mandel, 2018):

> what can confidently be advanced is a reduction of suffering. The question of whether society should intervene so that people will be happier is very controversial, but whether society should strive for people to suffer less — that's widely accepted."

The claim that it is morally important to bring about greater forms of pleasure is an additional premise that proponents of positive consequentialist views must defend, on top of the controversial premise that we have discussed in previous sections: that suffering can be measured against, and outweighed by, pleasure (more critical discussion of this latter premise can be found in Anonymous, 2015, sec. 2.2.12-2.2.13). Positive consequentialists thus need to defend two additional premises, both of which are more controversial than the premise that they share with negative consequentialists.

Common defenses of outweighing

The issues of clarification and justification that face the view that pleasure can always outweigh suffering are rarely addressed in much detail. So far, it seems that the most common way to defend this view has been to present thought experiments that are believed to render it plausible.

Bliss for many

The following thought experiment from Leuven & Visak's critique of Richard Ryder's painist view is an example (Leuven & Visak, 2013, p. 416):

> imagine a population of beings that are all not particularly happy, but are neither suffering. There would be one way of significantly raising the level of welfare of the whole population except one, by causing a mild and brief suffering to the one person. After this brief period of mild suffering this person would continue on his usual welfare level, while the rest of the population would have a really blissful live [sic]. Ryder's theory would dismiss this option, and rather require that everyone keeps muddling on with a more or less neutral level of welfare.

There are many things to say in response to this thought experiment. First, one can argue that a problem with the thought experiment is that the beings who are supposedly free from suffering nonetheless can appear to be in a troubled or disturbed state given the description.

For example, when we read about beings who are "not particularly happy" and who keep "muddling on with a more or less neutral level of welfare", we hardly get associations to beings who feel *perfectly untroubled* and whose conscious states feel *entirely unproblematic*. If we rephrase things in these terms, the thought experiment becomes significantly weaker. Compare the original formulation with: "Ryder's theory would dismiss this option, and rather prescribe that everyone remains in a perfectly untroubled and entirely unproblematic state of consciousness so as to avoid

the creation of suffering." (For more replies along similar lines, see Anonymous, 2015, sec. 2.2.2; Gloor, 2017, sec. 4.2; Vinding, 2020, sec. 2.4.)

Another possible reply is to defend the moral premise that it is (indeed) always wrong to create pleasure for some beings at the price of suffering for others (this moral claim may be considered a natural implication of Epicurean axiologies, yet it is not predicated on such axiological views, cf. Vinding, 2020, 6.4). Defenses of this premise can be found in Ryder, 2001, ch. 2; Vinding, 2020, ch. 3.

Ryder supports his view that "it is always wrong to cause pain to A merely in order to increase the pleasure of B" with various thought experiments, yet his view also follows from some other moral "rules" that he defends (Ryder, 2001, p. 30). For example, "Rule 11": our primary "moral concern should always be with the individual who is the maximum sufferer", which implies that, in Leuven and Visak's thought experiment, we should primarily be concerned with the worst-off person on whom suffering would be imposed (Ryder, 2001, p. 29). (A moral principle similar to this latter moral "rule" of Ryder's has been defended by Joseph Mendola, who holds that our chief moral obligation is to "ameliorate the condition of the worst-off moment of phenomenal experience in the world", Mendola, 1990, p. 86.)

Third, one may reply to the thought experiment from a negative consequentialist perspective concerned with other bads than just suffering. From such a perspective, one may, for example, argue that the beings who are "not particularly happy", yet who also not undergoing any experiential suffering, nonetheless could be in a bad and harmed state. For even if they are not suffering, they may still have a frustrated preference, such as the preference for living a very blissful life. And a negative consequentialist may hold that this frustrated preference, or many such frustrated preferences, could be worse than a single instance of mild suffering, and hence more important to reduce. Note, however, that this response does not rely on any outweighing in the sense of "canceling out" or "making up for" the suffering in question. The mild suffering would still be an uncompensated bad on this view, a bad that is only allowed in order to ameliorate a supposedly greater bad.

World destruction

A related thought experiment sometimes raised in favor of the view that pleasure can outweigh suffering is one that involves world destruction. For example, we might imagine a paradise full of blissful people, and then wonder whether it could really be good to painlessly destroy such a paradise for the sake of preventing a single instance of mild suffering.

A significant problem with this thought experiment is that it brings other issues into play than just that of pleasure outweighing suffering, such as world destruction. Another problem is that it attacks an implication that people with suffering-focused views, and harm-focused views more generally, need not endorse.

Taking a step back, we may start by noting that we could consider a similar thought experiment in which no sentient being has yet been created, but where we may create a paradise for innumerable sentient beings by imposing suffering on a single being. When we phrase the thought experiment in this way, a way that helps us control for status quo bias among other things, it is not clear that we are doing anything wrong by choosing not to create suffering for the sake of creating pleasure. Again, one can argue that it is the opposite — imposing suffering to create pleasure — that would be wrong (Vinding, 2020, ch. 3).

Second, it is worth noting that *all* views that say that experiential ill- and well-being are the only things that matter morally, which includes views such as negative and classical utilitarianism, will imply that we should destroy the world in order to prevent a tiny amount of suffering, provided the "net balance" of ill- and well-being is exactly zero otherwise (Pearce, 2017, "The Pinprick Argument"). Thus, world-destruction objections of this kind arguably count more as an objection against purely welfarist views than against a moral asymmetry between pleasure and pain, or between good and bad things more generally. (For more discussion of world destruction arguments against utilitarian views, see Knutsson, 2021.)

As hinted above, there are many replies available to this world destruction argument that do not rely on the view that positive things can outweigh suffering or other bad things. For example, one can hold that world destruction would result in other bads that are also worth preventing, such as the frustration of preferences, premature death, rights violations, the loss of hard-won knowledge and artifacts, etc.

In this way, one can maintain that it would be wrong to destroy the world to prevent mild suffering without thinking that the mild suffering in question can be "canceled out" or "made up for" by other things. World destruction would, on such a view, be wrong because the alternative to allowing the mild suffering would be even worse. Such a view may also help explain why the case where no sentient being has yet come into existence seems different, since most, if not all, of these other bads would be removed from the equation in that case.[1]

1 I am grateful to Tobias Baumann, Anthony DiGiovanni, and Michael St. Jules for helpful comments.

References

Ajantaival, T. (2021/2022). Minimalist Axiologies. Ungated

Animal Ethics. (2012). Negative consequentialism. Ungated

Anonymous. (2015). Negative Utilitarianism FAQ. Ungated

Bentham, J. (1789/2007). *An Introduction to the Principles of Morals and Legislation*. Dover Publications.

Breyer, D. (2015). The Cessation of Suffering and Buddhist Axiology. *Journal of Buddhist Ethics*, 22, pp. 533-560. Ungated

Fehige, C. (1998). A pareto principle for possible people. In Fehige, C. & Wessels U. (eds.), *Preferences*. Walter de Gruyter. Ungated

Geinster, D. (2012). The Amoral Logic of Anti-Hurt (Modified Negative Utilitarianism). Ungated

Gloor, L. (2016/2019). The Case for Suffering-Focused Ethics. Ungated

Gloor, L. (2017). Tranquilism. Ungated

Harnad, S. (2016). My orgasms cannot be traded off against others' agony. *Animal Sentience*, 7(18). Ungated

Hewitt, S. (2008). Normative Qualia and a Robust Moral Realism (PhD thesis). New York University. Ungated

Knutsson, S. (2016). Measuring happiness and suffering. Ungated

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated

Knutsson, S. (2021). The world destruction argument. *Inquiry*, 64(10), pp. 1004-1023. Ungated

Lazari-Radek, K. & Singer, P. (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford University Press.

Leuven, J. & Visak, T. (2013). Ryder's Painism and His Criticism of Utilitarianism. *J Agric Environ Ethics*, 26, pp. 409-419.

Mandel, A. (2018). Why Nobel Prize Winner Daniel Kahneman Gave Up on Happiness. Haaretz. Ungated

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Mendola, J. (1990). An Ordinal Modification of Classical Utilitarianism. *Erkenntnis*, 33(1), pp. 73-88.

Mill, J. S. (1863/2007). *Utilitarianism*. Dover Publications.

Ord, T. (2013). Why I'm Not a Negative Utilitarian. Ungated

Pearce, D. (2017). *Can Biotechnology Abolish Suffering?* The Neuroethics Foundation. Ungated

Ryder, R. (2001). *Painism: A Modern Morality*. Centaur.

Ryder, R. (2011/2017). *Speciesism, Painism and Happiness: A Morality for the 21st Century*. Andrews UK Ltd.

Ryder, R. (2015). Painism defended. *Think*, 14(41), pp. 47-55.

Scarre, G. (1996). *Utilitarianism*. Routledge.

Schopenhauer, A. (1819/1966). *The World as Will and Representation*, 2 vols. Dover publications.

Sherman, T. (2017). Epicureanism: An Ancient Guide to Modern Wellbeing. MPhil dissertation, University of Exeter. Ungated

Sidgwick, H. (1874/1981). *The Methods of Ethics*. Hackett Pub. Co.

Tomasik, B. (2015/2017). Are Happiness and Suffering Symmetric? Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Wolf, C. (1997). Person-Affecting Utilitarianism and Population Policy. In Heller, J. & Fotion, N. (eds.), *Contingent Future Persons*. Kluwer Academic Publishers. Ungated

Suffering and happiness: Morally symmetric or orthogonal?

Summary

The purported value symmetry between suffering and happiness ought to be questioned and contrasted with alternative views. I here present two asymmetric pictures that collectively cover a broad range of axiological and ethical views. These pictures merit serious consideration.

Introduction

"Some views, most notably standard economic utilitarian views, encourage us to treat bads as negative goods. But with Karl Popper, I'm convinced that this is a very serious mistake."

— Clark Wolf (Wolf, 2019)

"I believe that its insistence on the moral symmetry of happiness and suffering is one reason why many people find utilitarianism hard to take seriously. … this is a great pity [given the strong priority utilitarianism devotes to the reduction of suffering]."

— Jamie Mayerfeld (Mayerfeld, 1996, p. 335)

Our views of the relative moral significance of happiness and suffering matter greatly for our priorities, which renders it crucial that we scrutinize our immediate intuitions and assumptions on this issue. I have argued elsewhere that the notion that happiness can morally outweigh suffering stands in need of clarification and defense (Vinding, 2020b), and presented various arguments against a moral symmetry between happiness and suffering (Vinding, 2020a, part I; see also Ajantaival, 2021/2022).

My aim here is first to briefly review some of the arguments and views that reject a moral symmetry. These views then motivate the alternative pictures, i.e. actual visual models, that I shall propose as more plausible than the symmetric utilitarian picture.

The implications of symmetry

Karl Popper famously criticized the idea that we can treat suffering as "negative pleasure", or pleasure as "negative pain" (Popper, 1945, ch. 9, note 2):

> [A] criticism of the Utilitarian formula 'Maximize pleasure' is that it assumes, in principle, a continuous pleasure-pain scale which allows us to treat degrees of pain as

> negative degrees of pleasure. But, from the moral point of view, pain cannot be outweighed by pleasure, and especially not one man's pain by another man's pleasure.

One may argue that an extrapolation of the implications of the supposed moral symmetry between happiness and suffering gives us reason to agree with Popper's critique.

For example, such a symmetry would imply that it is just as morally important to cause an untroubled person to experience a state of intense happiness as it is to alleviate (similarly) intense suffering. Jamie Mayerfeld argues against this implication with the following thought experiment (Mayerfeld, 1999, p. 133):

> We give surgery patients anesthesia to avert the agony they would feel if they remained conscious. Suppose some drug became available that gave people a joy as intense as the pain averted by anesthesia, and suppose that there were no drawbacks in the consumption of this drug. It seems quite clear to me that the provision of this drug would be less important than the administration of anesthesia.

One thing to note in this context is that it is not obvious what it means to talk about "similarly intense" states of happiness and suffering, respectively (Knutsson, 2016a). Yet, as Mayerfeld hints, even if we grant that we can meaningfully compare happiness and suffering in this way, the implication of symmetry outlined above still conflicts with the plausible moral intuition that the badness of suffering does not compare to the supposed badness of a neutral, untroubled state of consciousness that could have been more pleasurable (Anonymous, 2015, sec. 2.2.14).

Yet perhaps the most damning implication of symmetry is that it can be permissible — indeed obligatory — to impose extreme suffering on a given individual in order to raise the happiness of others, even when those others are already well-off (Vinding, 2020a, ch. 3).

Some have further argued that findings from psychology and neuroscience give us reason to be skeptical of the view that happiness and suffering are relevantly symmetric (Diener & Emmons, 1984; Baumeister et al., 2001, p. 331; Shriver, 2014).

As Adam Shriver writes (Shriver, 2014, abstract):

> Recent results from the neurosciences demonstrate that pleasure and pain are not two symmetrical poles of a single scale of experience but in fact two different types of experiences altogether, with dramatically different contributions to well-being.

Consequently, "ethicists cannot simply assume that what is said about pleasure has similar implications for pain, and vice versa" (Shriver, 2014, p. 13, draft version).

An asymmetry in urgency

A fundamental difference, one may argue, is that suffering is intrinsically problematic, and that it carries an inherent urgency — in the words of Thomas Metzinger, an "urgency of change" (Metzinger, 2017, "Option 4: eliminating the NV-condition"). A neutral state that could have been intensely happy, by contrast, is not problematic, and hence "raising" such unproblematic states toward pleasure carries no corresponding urgency.

Popper expressed a similar view (Popper, 1945, ch. 9, note 2):

> suffering makes a direct moral appeal, namely, the appeal for help, while there is no similar call to increase the happiness of a man who is doing well anyway.

Contentment and the avoidance of preference frustration

Another view that entails a moral asymmetry between happiness and suffering is the view that contentment — i.e. the absence of discomfort and frustrated desires — is what matters, not the intensity of our pleasures.

There are many variations of this view. One is the antifrustrationist view defended by Christoph Fehige, according to which "we don't do any good by creating satisfied extra preferences. What matters about preferences is not that they have a satisfied existence, but that they don't have a frustrated existence." (Fehige, 1998, p. 518).

On this view, pleasure is only good to the extent that it satisfies a frustrated preference, and hence the creation of pleasure is not valuable per se, and may indeed often have no value at all. Suffering, in contrast, *does* imply — and arguably constitutes — a frustrated preference, and its prevention will thus always be valuable on the antifrustrationist view.

Michael St. Jules expresses a similar view of the moral (un)importance of promoting pleasure for its own sake (St. Jules, 2020):

> something only matters if it matters (or will matter) to someone, and an absence of pleasure *doesn't necessarily* matter to someone who isn't experiencing pleasure, and certainly doesn't matter to someone who does not and will not exist, and so we have no inherent reason to promote pleasure. On the other hand, there's no suffering unless someone is experiencing it, and according to some definitions of suffering, [the experience of suffering] necessarily matters to the sufferer.

This is related to the view that "a conscious state is only non-optimal or problematic if this is directly experienced, not if the state doesn't match up in some comparison we make from the outside" (Anonymous, 2015, sec. 2.2.1; Vinding, 2020a, sec. 1.4).

Similar arguments have been made in defense of the Asymmetry in population ethics, which says that we have a moral obligation not to bring miserable lives into existence, yet no corresponding moral obligation to bring about happy lives (Benatar, 1997; 2006; St. Jules, 2019; Frick, 2020).

Another set of views that identify contentment rather than pleasure intensity as the seat of experiential value are various Epicurean and Buddhist views of well-being. These views hold that the most significant determinant of the value of our experiences is the degree to which they are absent of negative components, such as pain, fear, and boredom. Experiential states that are absent of such negative components are deemed optimal, and hence there is no additional value to adding intense pleasure to such an already, according to these views, optimal state (Schopenhauer, 1819, vol I, p. 319; 1851, pp. 41-43; Breyer, 2015; Gloor, 2017; Sherman, 2017; Knutsson, 2019).

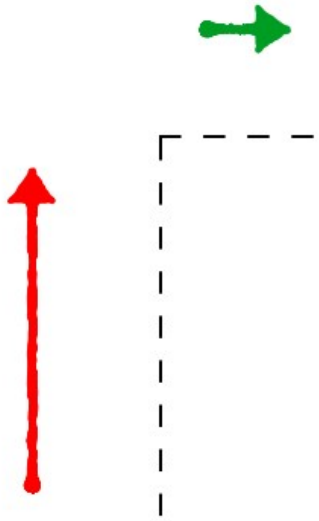
A shared picture: Suffering and happiness as morally orthogonal

In light of the preceding arguments and views, it seems natural to propose a picture according to which happiness and suffering are morally orthogonal — i.e. orthogonal in the notional space of moral importance.

Indeed, such an orthogonal picture captures the essence of most of the views presented above. For example, whether we think increasing pleasure carries a moral urgency and importance that is lexically inferior to the moral importance of reducing suffering (cf. Wolf, 1997, Knutsson, 2016c), or whether we think increasing pleasure carries no moral importance whatsoever, the resulting picture is practically the same: the moral importance of reducing suffering occupies a different, overriding moral dimension than does the moral (un)importance of increasing pleasure.

(For some of the views presented above, e.g. the Asymmetry, this picture may not apply in general, but it will still capture the essence of these views in the case of suffering versus pleasure for future beings who do not currently exist, which is arguably the most relevant case to consider in relation to our priorities.)

The picture we get is roughly the following:



The reduction of suffering is represented with a red arrow that urgently points away from the depths of misery, while the increase of happiness is symbolized with a green arrow that goes sideways: moving along this dimension is fine, but it carries no (comparable) moral urgency. This is the picture we get when we reject the notion that pleasure can morally outweigh suffering.

Alternative picture: "Craving pleasures" as subtly negative

Some views may insist on a more pessimistic representation of the value of pleasure, or at least of certain forms of pleasure.

For example, Simon Knutsson defends a view according to which some kinds of pleasure — so-called kinetic pleasures (pleasures involved in the active pursuit of something) — can have negative value. This stands in contrast to katastematic, or static pleasures, i.e. states of perfect calm and tranquility, which Knutsson contends have neutral value (Knutsson, 2019).

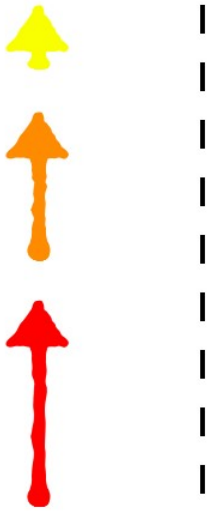
Knutsson follows (Cooper, 2012), who argues that (at least some) kinetic pleasures are "compromised by the stressful state of mind … associated with intense desires"; that such pleasures involve a "frustrated dissatisfaction"; or that they are "mixed with pain" (Cooper 2012, pp. 237-238, as quoted in Knutsson, 2019). These subtly negative components are what render (at least some) kinetic pleasures negative on Knutsson's view.

(Relatedly, modern neuroscience draws a distinction between "wanting" and "liking", and one could argue that the putatively negative components of kinetic pleasures roughly correspond to components of "wanting", Berridge & Robinson, 2016; Tomasik, 2016. One may also relate these

negative experiential components to the Buddhist concept of *kama-tanha*, which roughly means craving for pleasure.)

A similar view is the tranquilist position defended by Lukas Gloor, which holds that "a state of consciousness is negative or disvaluable if and only if it contains a craving for change" (Gloor, 2017, 2.2). On this view, any craving for pleasure is disvaluable, as is any state of pleasure that contains a craving for change.

These views may be represented in the following way:



Pleasures that contain any cravings or subtle frustrations can be thought of as occupying the yellow arrow: they are far from intensely painful, yet they are still mildly negative. Only fully tranquil states are optimal on these views.

Arthur Schopenhauer seemingly also viewed pleasure as something that corresponds to a move along the yellow arrow above (Schopenhauer, 1819, vol I, p. 319):

> [Happiness] is not a gratification which comes to us originally and of itself, but it must always be the satisfaction of a wish. For desire, that is to say, want, is the precedent condition of every pleasure; but with the satisfaction, the desire and therefore the pleasure cease; and so the satisfaction or gratification can never be more than deliverance from a pain, from a want.

Note how Schopenhauer argues that the pleasure itself ceases when satisfaction is attained, and how he appears to see pleasure as inextricably tied to desire ("with the satisfaction, the desire and

therefore the pleasure cease"). Thus, one may argue that pleasure for Schopenhauer roughly amounts to a (suboptimal) kinetic pleasure in the framework defended by Knutsson.

Explaining the appearance of symmetry

The views outlined in the previous section may help explain why happiness and suffering are sometimes considered morally symmetric. On these views, increasing pleasure can appear the moral equivalent of reducing negative states because it in fact often *is*, or at least results in, the reduction of negative states. (And note in this context that words like "unhappy", "unpleasant", and "uncomfortable" invariably refer to states that are negative rather than neutral.)

On these views, what we usually consider a neutral state actually tends to contain bothersome components — worry, stress, boredom, etc. — yet we mostly fail to recognize these discomforts, partly because we are so used to them, and partly because they are shared by everyone (Benatar, 2006, p. 72). Thus, when we experience pleasure in what we naively imagine to be a neutral state, we are in fact, on these views, misinterpreting the relief from such negative components as the attainment of a truly positive state (Gloor, 2017, sec. 2.1). (I say more about this in my essay "A phenomenological argument against a positive counterpart to suffering".)

Resistance to asymmetry due to a non sequitur?

One reason we may feel a pull to affirm a moral symmetry between happiness and suffering is the tacit assumption that the rejection of such a moral symmetry must necessarily be hostile to our continued existence. Yet this need not follow. After all, one can hold that other things besides suffering and happiness matter morally, such as accomplishing one's life goals and living a virtuous life. Pluralist views of this kind are common among academic philosophers. (See e.g. Mayerfeld, 1996, "Life and Death"; Wolf, 1997, sec. VIII; Benatar, 2006, pp. 211-219.)

For example, Jamie Mayerfeld (Mayerfeld, 1999, p. 160) argues that death, for others and ourselves, is bad because

> the successful completion of our projects depends on our staying alive. Other reasons can be added. Most people have a deeply rooted desire to go on living. A proper respect for their autonomy requires that we do not thwart their desire to live.

There are also strong instrumental reasons to favor continued existence, for oneself and others, even if one thinks suffering is the only thing that matters (Vinding, 2020a, sec. 8.1-8.2). For instance, our continued existence is a precondition if we are to reduce suffering for others in effective ways,

which means that the aim of reducing suffering strongly recommends that we secure our continued existence.

Additionally, it is worth noting that a moral symmetry between happiness and suffering seems to carry implausible implications when it comes to the ethics of death and continued existence, suggesting that such a moral symmetry is not in fact a satisfying foundation for our views on these matters.

For example, a purely welfarist view relying on such a moral symmetry would imply that it would be morally right to kill all existing beings in arbitrarily excruciating ways if we could in turn replace them with sufficiently many, sufficiently happy beings. This implication seems considerably more repugnant than the corresponding replacement implications entailed by purely suffering-focused views, as these views do not allow replacements that increase suffering. (For further discussion of these issues, see Knutsson, 2021; Ajantaival, 2022.)

Thus, valuing and preferring continued existence does not in itself constitute a strong reason to endorse a moral symmetry between happiness and suffering, as there are other, arguably more plausible views that can accommodate this intuition, and indeed accommodate it more robustly. And these alternative options are only rendered more plausible, comparatively speaking, by the many reasons reviewed here and elsewhere against the purported moral symmetry between happiness and suffering (see e.g. Mayerfeld, 1996; 1999; Vinding, 2020a, ch. 3).[2]

References

Ajantaival, T. (2021/2022). Minimalist Axiologies. Ungated

Ajantaival, T. (2022). Peacefulness, nonviolence, and experientialist minimalism. Ungated

Anonymous. (2015). Negative Utilitarianism FAQ. Ungated

Baumeister, R. et al. (2001). Bad is stronger than good. *Review of General Psychology*, 5, pp. 323-370.

Benatar, D. (1997). Why It Is Better Never to Come into Existence. *American Philosophical Quarterly*, 34(3), pp. 345-355.

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

Berridge, K. & Robinson, T. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *Am Psychol*, 71(8), pp. 670-679.

2 Thanks to Tobias Baumann, Michael St. Jules, and Rupert McCallum for helpful comments.

Breyer, D. (2015). The Cessation of Suffering and Buddhist Axiology. *Journal of Buddhist Ethics*, 22, pp. 533-560. Ungated

Cooper, J. (2012). *Pursuits of Wisdom: Six Ways of Life in Ancient Philosophy from Socrates to Plotinus*. Princeton University Press.

Diener, E. & Emmons, R. (1984). The Independence of Positive and Negative Affect. *Journal of Personality and Social Psychology*, 47(5), pp. 1105-1117.

Fehige, C. (1998). A pareto principle for possible people. In Fehige, C. & Wessels U. (eds.), *Preferences*. Walter de Gruyter. Ungated

Frick, J. (2020). Conditional Reasons and the Procreation Asymmetry. Ungated

Gloor, L. (2016/2019). The Case for Suffering-Focused Ethics. Ungated

Gloor, L. (2017). Tranquilism. Ungated

St. Jules, M. (2019). Defending the Procreation Asymmetry with Conditional Interests. Ungated

St. Jules, M. (2020). Comment on the goodness of pleasure. Ungated

Knutsson, S. (2015). The 'Asymmetry' and extinction thought experiments. Ungated

Knutsson, S. (2016a). Measuring happiness and suffering. Ungated

Knutsson, S. (2016b). What is the difference between weak negative and non-negative ethical views? Ungated

Knutsson, S. (2016c). Value lexicality. Ungated

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated

Knutsson, S. (2021). The world destruction argument. *Inquiry*, 64(10), pp. 1004-1023. Ungated

Mayerfeld, J. (1996). The Moral Asymmetry of Happiness and Suffering. *Southern Journal of Philosophy*, 34, pp. 317-338.

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Metzinger, T. (2017). Suffering. In Almqvist, K. & Haag, A. (eds.), *The Return of Consciousness*. Axel and Margaret Ax:son Johnson Foundation.

Popper, K. (1945/2011). *The Open Society and Its Enemies*. London: Routledge.

Schopenhauer, A. (1819/1966). *The World as Will and Representation*. 2 vols. New York: Dover.

Schopenhauer, A. (1851/1970). *Essays and Aphorisms*. Harmondsworth, Eng: Penguin Books.

Sherman, T. (2017). Epicureanism: An Ancient Guide to Modern Wellbeing. MPhil dissertation, University of Exeter. Ungated

Shriver, A. (2014). The Asymmetrical Contributions of Pleasure and Pain To Animal Welfare. *Cambridge Quarterly of Healthcare Ethics*, 23(2), pp. 152-162. Ungated draft version

Tomasik, B. (2015). Are Happiness and Suffering Symmetric? Ungated

Tomasik, B. (2016). How Cravings Influence Happiness-vs.-Suffering Trades. Ungated

Vinding, M. (2020a). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Vinding, M. (2020b). On purported positive goods "outweighing" suffering. Ungated

Wolf, C. (1997). Person-Affecting Utilitarianism and Population Policy. In Heller, J. & Fotion, N. (eds.), *Contingent Future Persons*. Dordrecht Boston: Kluwer Academic Publishers. Ungated

Wolf, C. (2019). Panel: Humanity and the Future. Ungated

A phenomenological argument against a positive counterpart to suffering

Various views deny that suffering has a positive counterpart. Proponents of such views often pursue a line of argument that focuses on the prevalence of subtle frustrations and bothersome sensations. That is, when we typically think that we are in a neutral state, and we claim that some pleasure takes us above that neutral state, what we are experiencing is really a subtly bothered and unsatisfied state that becomes (somewhat) relieved of its commonly overlooked unpleasant features (see e.g. Sherman, 2017, pp. 103-107; Gloor, 2017, sec. 2.1; Knutsson, 2022, sec. 5).

This essay will pursue a different line of argument. Rather than focusing on unpleasant states, and arguing for their subtle omnipresence, my aim here is instead to zoom in on the purportedly positive side. I will argue that purportedly positive experiences do not possess any property that renders them genuine opposites of painful and uncomfortable experiences, neither in phenomenological nor axiological terms.

Candidates of positive experiences

I will start by listing a variety of experiential states that are often claimed to be positive. This list is by no means exhaustive, but it still helps to make the discussion that follows more concrete and precise.

Common candidates of positive experiences include **feelings of**:

> Excitement, gratitude, optimism, motivation, contentment, inspiration, enthusiasm, pride, ecstasy, amusement, elation, orgasm, euphoria, lust, hope, cheerfulness, awe, confidence, joy, passion, love, social recognition, social connection, being desired, and being successful.[3]

With this list in place, we can avoid certain pitfalls and misunderstandings. For example, if someone asserts that positive experiences do not exist, many people may intuitively interpret that to mean that experiential states such as excitement and gratitude do not exist. This seems like a trivially false claim, and hence the stated claim about the non-existence of positive experiences is apt to be dismissed.

3 Additional such experiences are listed in Moore, 2004.

So to be clear, I am *not* claiming that the feelings listed above do not exist. What I am claiming, rather, is that none of these feelings or experiential states are phenomenological opposites to suffering and discomfort.

In other words, my core claim is that **there is nothing about the phenomenological nature of these states that render them a positive counterpart to suffering** (beyond the extent to which they are absent of suffering). And since these states are not phenomenological opposites to suffering, they are plausibly not axiological opposites to it either. That is, we should not consider states such as those listed above to be axiological opposites to suffering any more than we should consider, say, experiences of color or sound to be axiological opposites, or counterparts, to suffering.

Unpacking the phenomenological claim that I deny

Intuitively, it might seem like I am making a strong claim. Yet I would argue that it is actually the claim that I am denying that is the stronger one, and I believe this becomes apparent once we carefully unpack the exact nature of that claim.

To say that certain experiences represent a positive counterpart to suffering is not merely to say that the experiences in question are absent of suffering. Instead, the claim is essentially that experiences of suffering fall along one axis of experience, while (purported) positive experiences fall along another axis, where these two axes are anti-directional relative to some neutral point or state space. Suffering has a phenomenological counterpart that in some sense amounts to *anti*-suffering.

When specified in these more precise terms, it is at least not obvious that the existence of such a phenomenological counterpart to suffering is more plausible than is its non-existence. And as we

shall see in the next section, there are indeed good reasons to doubt the existence of such a positive counterpart.

Arguments against a phenomenological counterpart to suffering

A priori reasons to doubt phenomenological dual opposites

We arguably have no a priori reason to think that suffering has a positive counterpart in the way described above. More than that, one could even argue that we have a priori reasons to doubt the dual-axis picture outlined in the previous section. After all, a view of phenomenology that insists on the existence of such an anti-directional double axis of experience seems considerably less simple and less parsimonious than does a view that entails no oppositely directed dimensions of experience.

More generally, one may argue that it is doubtful whether experiential states even *can* have phenomenological opposites. Does this notion even make sense?

For instance, does the experience of the color red have an opposite experience? What could this mean? Perhaps the most plausible candidate pair of phenomenological opposites in the realm of phenomenal color is black and white. Yet even here it is far from obvious whether we have identified an example of phenomenological opposites, rather than states that exhibit a high degree of phenomenological contrast. The same could be said about experiential states that involve loud sounds versus states that involve silence. There is stark contrast, and there is the presence and absence of different properties, but it is doubtful whether there are genuine anti-properties in any meaningful sense (a similar point is made in Heathwood, 2007, p. 27).

Continuing the analogy to phenomenal sound, one could argue that phenomenological anti-suffering makes no more sense than does phenomenological "anti-noise". That is, just as an experience cannot get more silent than absolute silence, an experience cannot be more opposite or anti-directional to suffering than when it is wholly absent of suffering or discomfort.[4]

In general, it seems unclear what it would be like for two different experiential states to be phenomenological dual opposites, and it is likewise unclear whether this notion of phenomenological opposites even makes sense in the first place.

Introspection

In my view, the strongest argument against the existence of a phenomenological counterpart to suffering is that introspection yields no sign of such a counterpart. When we introspectively

4 Thanks to Anthony DiGiovanni for suggesting this extension of the analogy.

examine the proposed candidates of positive experiences, such as those listed above, we do not find that they have any phenomenological properties that render them the dual opposites of suffering, or *anti*-suffering, as it were.

In other words, even if we grant that the notion of phenomenological dual opposites is a coherent one (despite the doubts raised in the previous section), and if we set out to search for the phenomenological dual opposites of suffering via introspective examination, the conclusion, I submit, is that they do not exist.

I do not expect anyone to accept this claim on authority. I encourage readers to pursue this introspective exercise themselves: to search for the phenomenological property (or properties) that would render an experience the dual opposite of suffering. (It is my impression that such a phenomenological property is often tacitly assumed but rarely seriously looked for or scrutinized.)

Note that I am not claiming that the purportedly positive experiences listed above have no properties in common (I take no position on that issue here). Nor am I saying that the experiences listed above cannot be intense or even all-consuming.

One might object that different experiential states are opposites in terms of the respective behaviors that they tend to elicit — e.g. some experiences may motivate us to *approach* ripe berries while other experiences may motivate us to *avoid* moldy ones. But even if we grant that different experiences can in some sense motivate opposite behaviors, this still does not imply that the experiential states in question are dual opposites in phenomenological terms. After all, the experience of wanting and savoring something does not seem to be the dual opposite of the experience of not wanting and actively disliking something else.

Indeed, one could argue that the experiences that motivate us to approach certain things, as well as the experiences that motivate us to avoid certain (other) things, are all ultimately driving us through the force of frustrated desires and unpleasant states. On this view, even drives to attain desired experiences (e.g. sexual and gustatory ones) are ultimately animated by the frustrations and bothersome states that we experience when our desires for these states are not fulfilled (cf. Sherman, 2017, pp. 60-61; Knutsson, 2019; Vinding, 2022, sec. 23).

Likewise, it seems that one can reconceptualize the "approach versus avoidance" framework in comparative terms that dissolve the apparent oppositeness. To take the example of berries, one can think of it in terms of preferences for (experiences of) ripe berries over mild hunger, mild hunger over slightly moldy berries, slightly moldy berries over intense hunger, etc. These latter examples reveal that whether we approach or avoid something is not an absolute matter, but rather dependent on our alternative choices. And in line with the argument made above, one could argue that the

motivating force of our experiential states is ultimately best understood in such comparative terms, where we generally seek to attain states that are less bothered or which have fewer unmet needs (cf. Sherman, 2017, p. 106).[5]

Evidence from psychology and neuroscience

Lastly, there is evidence from psychology and neuroscience that casts doubt on the notion that pain and suffering are opposites of pleasure and other purportedly positive experiences. Evidence from neuroscience is less relevant than the introspective evidence, since the claim that we are concerned with is a phenomenological one, and neuroscience is not directly about phenomenological claims. Yet evidence from neuroscience can plausibly still help inform our views on phenomenology and on the nature of suffering and its purported dual opposites.

Baumeister et al. write the following in a review article (2001, p. 331): "Although laypersons typically regard [pleasant and unpleasant emotions] as opposites, there is some evidence that the two are somewhat independent …"

Likewise, philosopher of psychology and neuroscience Adam Shriver summarizes the evidence in the following way (2014b, abstract):

> Recent results from the neurosciences demonstrate that pleasure and pain are not two symmetrical poles of a single scale of experience but in fact two different types of experiences altogether, with dramatically different contributions to well-being.

(See also Shriver, 2014a; Bain & Brady, 2014; de Boer, 2014, p. 712.)

Why we might believe that a positive counterpart to suffering exists

If experiences such as those listed earlier are not a positive counterpart to suffering in phenomenological terms — as I argue they are not — it is natural to wonder why these experiences are often thought to be such a positive counterpart.

I am aware of two factors that may help explain this belief. One reason might be that we are used to thinking about various phenomena in terms of positive and negative real numbers, and hence we are quick to project such numbers onto our experiences, even if such a conceptual representation might not be supported by careful introspection or other lines of evidence.

Another potential explanation that has been proposed by various authors is that purportedly positive experiences often serve to reduce states of suffering and discomfort, and hence we might confuse this genuine reduction of unpleasant states — which in some sense *is* a case of anti-suffering — for

5 Thanks to Simon Knutsson for making this point about reframing things in comparative terms.

being a positive experience that goes over and above states that are wholly absent of suffering and discomfort.

As Toby Sherman puts it, "pleasure can be remedial, not for particular pains, but for pain-in-general, which is why it often seems to be not remedial at all." (Sherman, 2017, p. 8; see also sec. 11.2.)

Lukas Gloor has expressed a similar view (Gloor, 2017, sec. 2.1):

> When our brain is flooded with pleasure, we temporarily become unaware of all the negative ingredients of our stream of consciousness, and they thus cease to exist. Pleasure is the typical way in which our minds experience temporary freedom from suffering, which may contribute to the view that happiness is the symmetrical counterpart to suffering, and that pleasure, at the expense of all other possible states, is intrinsically important and worth bringing about.

Indeed, one may argue that the two potential explanations reviewed above complement each other: we are accustomed to thinking in terms of real-valued pluses and minuses, and we apparently find introspective support for thinking about our experiences in these terms when we notice that some feelings (e.g. feelings of excitement and gratitude) move us away from the "minuses", which suggests that they are genuine "pluses". But what we miss, Sherman and Gloor might argue, is that these "pluses" only represent *relative* "pluses", toward a smaller or non-existent "minus" (i.e. less suffering and discomfort). They do not take us to an absolute "plus" of phenomenological anti-suffering. (See also Knutsson, 2022, sec. 5.2.)

Axiological implications

As noted earlier, it seems natural to argue that if purportedly positive experiences do not represent a phenomenological counterpart to suffering, then they do not represent an axiological counterpart to suffering either.

An expanded axiological version of the phenomenological argument outlined above lends further support to this view. That is, just as there is nothing about purportedly positive experiences that suggests that they are phenomenological opposites to suffering (in the strong anti-directional sense specified above), nor is there, I submit, anything else about those experiences that suggests that they can outweigh experiences of suffering.

Again, I would encourage readers to introspect and search for any such phenomenological property that would suggest that an experience can axiologically outweigh experiences of suffering.

Of course, references to phenomenological properties are not the only way in which one could attempt to defend the view that purportedly positive experiences can outweigh states of suffering. Yet arguments that rely on phenomenological properties are perhaps among the most obvious arguments that could be made in its favor, and it seems that if we were to establish that there is no phenomenological support for the claim that purportedly positive experiences can outweigh suffering (as I argue there is not), then this would be a significant blow to that claim about outweighing.

Objection: This argument cuts both ways

Perhaps the main objection to my argument is that it cuts both ways: if purportedly positive experiences are not the phenomenological opposite of suffering, then neither is suffering the phenomenological opposite of purportedly positive experiences. So the axiological argument could also be made in the other direction: we have no phenomenological reason to think that suffering can outweigh purportedly positive experiences, and why should we start from the axiological assumption that suffering is worth preventing rather than assuming that purportedly positive experiences are worth creating?

I agree with the first part of this objection: suffering is indeed not the phenomenological opposite of purportedly positive experiences ("is not the opposite of" is clearly a symmetric relation). However, I think there are various arguments that support the axiological starting point that suffering is worth preventing over the starting point that purportedly positive experiences are worth creating.

One such argument is a basic asymmetry between the presence of suffering and the absence of purportedly positive experiences, and a consequent asymmetry between the non-prevention and non-creation of these respective states. That is, the presence of suffering amounts to a problematic state, and hence so does the failure to prevent suffering, whereas the absence of purportedly positive experiences does not amount to a problematic state, and hence neither does its non-creation. The presence of suffering is more plausibly a state worth rectifying than is the absence of purportedly positive experiences, and the absence of purportedly positive experiences is arguably not in any way suboptimal in axiological terms.

Another argument for the same conclusion is that there is a phenomenological asymmetry in the plausibility of these respective claims. In phenomenological terms, experiences of suffering and discomfort feel like they have intrinsic disvalue — or at least they have phenomenal qualities that render it plausible to assign them such disvalue — but purportedly positive experiences do not, on closer inspection, feel like they have intrinsic positive value, or like it is plausible to assign them such value (see e.g. Knutsson, 2021, sec. 3; Knutsson, 2022).

Moreover, even if one thinks that purportedly positive experiences feel like they have some intrinsic value, it still seems plausible that states of suffering have phenomenal features that make them disvaluable in a qualitatively different and overriding way, such that they are not plausibly outweighed by the alleged intrinsic value of purportedly positive experiences (Vinding, 2020, sec. 1.4).

In particular, one could argue that suffering introspectively carries a felt "urgency of change" while purportedly positive experiences do not (cf. Metzinger, 2017, p. 254). And to the extent that purportedly positive experiences do contain such an urgency, one may argue that they are in fact bothersome and suboptimal experiences (Gloor, 2017, sec. 2.2; Knutsson, 2019).

(Additional arguments in favor of a strong axiological asymmetry between suffering and purportedly positive experiences are found in Vinding, 2020, Part I; Ajantaival, 2021/2022.)

Purportedly positive experiences can still be instrumentally positive

An important qualification is that the non-existence of intrinsically positive experiences does not imply the non-existence of *instrumentally* positive experiences. (By "positive experiences", I here specifically mean "suffering-outweighing experiences".)

For instance, states of excitement and gratitude may still have "net positive" value to the extent that they help prevent suffering, such as by relieving suffering in the experiencer or by motivating future actions that reduce suffering. These experiences can thus still be worth actively cultivating and investing in, even if it is ultimately for instrumental reasons. Much like knowledge, they can serve as an important resource for creating a better world.

Concluding reflections on my argument

An alternative view to the one I have defended here is that there exist positive experiences that are phenomenological opposites to suffering, but that those positive experiences do not have corresponding positive value. I suspect that this is the view that many people will think of when they hear a claim such as "positive experiential value does not exist". And that view may seem inconsistent and ad hoc. After all, if we assign negative value to negative experiences, why should we not assign positive value to oppositely directed positive experiences?

What I have tried to argue in this essay is that that view rests on an erroneous foundation. There are no oppositely directed positive experiences in phenomenological terms to begin with (i.e. no phenomenological anti-suffering), and hence there is nothing inconsistent or ad hoc about not assigning corresponding positive intrinsic value to any experiences.

The view I have charted here does not deny the existence of excitement, amusement, awe, etc.; it does not deny the instrumental utility of those states; and it does not posit any ad hoc break between the phenomenological and the axiological level. These features seem worth highlighting, as it appears that a strong axiological asymmetry between suffering and purportedly positive experiences is often deemed implausible precisely because it is thought to entail those non sequiturs.[6]

References

Ajantaival, T. (2021/2022). Minimalist Axiologies. Ungated

Bain, D. & Brady, M. (2014). Pain, Pleasure, and Unpleasure. *Review of Philosophy and Psychology*, 5(1), pp. 1-14. Ungated

Baumeister, R. et al. (2001). Bad is stronger than good. *Review of General Psychology*, 5, pp. 323-370.

de Boer, J. (2014). Scaling happiness. *Philosophical Psychology*, 27(5), pp. 703-718.

Gloor, L. (2017). Tranquilism. Ungated

Heathwood, C. (2007). The Reduction of Sensory Pleasure to Desire. *Philosophical Studies*, 133(1), pp. 23-44.

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated.

Knutsson, S. (2021). The world destruction argument. *Inquiry*, 64(10), pp. 1004-1023. Ungated

Knutsson, S. (2022). Undisturbedness as the hedonic ceiling. Ungated

Metzinger, T. (2017). Suffering. In Almqvist, K. & Haag, A. (eds.), *The Return of Consciousness: A New Science on Old Questions*. Axel and Margaret Ax:son Johnson Foundation.

Moore, A. (2004/2019). Hedonism. The Stanford Encyclopedia of Philosophy. Ungated

Sherman, T. (2017). Epicureanism: An Ancient Guide to Modern Wellbeing. MPhil dissertation, University of Exeter. Ungated

Shriver, A. (2014a). The Asymmetrical Contributions of Pleasure and Pain to Subjective Well-Being. *Review of Philosophy and Psychology*, 5, pp. 135-153. Ungated

Shriver, A. (2014b). The Asymmetrical Contributions of Pleasure and Pain to Animal Welfare. *Cambridge Quarterly of Healthcare Ethics*, 23(2), pp. 152-162.

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

6 For their helpful comments, I thank Tobias Baumann, Anthony DiGiovanni, Simon Knutsson, and Winston Oswald-Drummond.

Vinding, M. (2022). Point-by-point critique of Ord's "Why I'm Not a Negative Utilitarian". Ungated

A thought experiment that questions the moral importance of creating happy lives

Many people have the intuition that extinction would be bad. A problem, however, is that the term "extinction" carries many different connotations, and extinction may be considered bad for many different reasons. For instance, an extinction scenario might be considered bad because it involves frustrated preferences, violations of consent, or lethal violence. Yet extinction scenarios need not involve any of these elements in principle. By considering thought experiments that involve extinction without involving any of the elements listed above, we can get a better sense of what might explain the intuition that extinction would be bad. In this post, I will present a thought experiment that casts doubt on the notion that extinction would be bad or morally objectionable because it would prevent the creation of future happy lives.

Introduction

It is often implied that the worst thing about extinction is that it could prevent a potentially vast number of happy lives from coming into existence (see e.g. Parfit, 1984, pp. 453-454; Lazari-Radek & Singer, 2014, pp. 375-377; Ord, 2020, pp. 43-44). Conversely, the intuitive badness of extinction is sometimes invoked in support of the purported value of creating happy lives (see e.g. Holtug, 2004, pp. 139-140; Mogensen, 2022, p. 11).

The latter line of argument is problematic because the issue of extinction potentially draws other elements into play than that of creating happy lives, e.g. the violation of existing preferences (Knutsson, 2015). To control for such potentially distorting factors, we can devise a thought experiment that excludes these extraneous elements. I will present such a thought experiment in the following section, and I will proceed to argue that this thought experiment questions the value and moral importance of creating new happy lives (for their own sake).

Thought experiment: A world of voluntary non-procreation

Recall the three elements mentioned above: frustrated preferences, violations of consent, and lethal violence. These elements are so commonly associated with extinction that it might seem difficult to imagine extinction scenarios without them. Yet extinction scenarios free of those elements are in fact conceivable (even if they may not be realistic), as illustrated by the following hypothetical world:

A world of voluntary non-procreation

> Imagine a world that consists only of people who have a strong preference not to procreate. These people are not in any way harmed or worse off by their voluntary non-procreation (we can imagine that society is arranged such that everyone is taken care of when they become old, e.g. by insentient robots). Nor do the people in this world regret the fact that their non-procreation will result in extinction; in fact, they are at peace with this outcome, and even prefer it over the alternative. Peaceful extinction through voluntary non-procreation is what everyone in this hypothetical world wants and what everyone considers morally best (or least bad). And this is eventually what happens in this world: in accordance with their own will, everyone refrains from procreating, and extinction eventually occurs without any violence being involved.

(A similar thought experiment is found in Knutsson, 2015.)

Would extinction be bad or morally wrong in the world of voluntary non-procreation?

It is natural to wonder whether the extinction that occurs in the world of voluntary non-procreation is bad, and to further ask whether it is morally wrong for the people in that world not to procreate. At first sight, it is not clear what would be bad or morally wrong about this extinction outcome (compared to alternative scenarios that involve continued procreation). After all, no one has their preferences or their consent violated, nor is anyone subjected to violence of any kind. Moreover, one could argue that extinction would be the least bad outcome in this hypothetical world, both because it would not violate the preferences or the consent of existing people (it would even satisfy their preferences), and because it would prevent all bads for future generations, including their potential suffering, preference frustrations, violence, and death.

Yet proponents of the moral importance of creating happy lives may argue that this hypothetical extinction scenario is extremely suboptimal, assuming that continued procreation could have created new happy lives. And those who assign great moral importance to the creation of happy lives would presumably further argue that the people in the world of voluntary non-procreation are doing something morally wrong, perhaps even something atrociously wrong, if they could have brought trillions of happy beings into existence (Parfit, 1984, pp. 453-454).

But this view appears to have implausible implications. For example, it would seem to imply that the people in the world of voluntary non-procreation are morally obliged to bring happy beings into the world (assuming that they could create such beings). This is arguably an implausible moral

obligation in general — especially if it implies that one should incur significant opportunity costs in terms of reducing suffering (Vinding, 2022). And the obligation seems more implausible still when it goes unanimously against the preferences and moral judgments of all existing beings (e.g. one could think that it becomes less plausible in such a world for contractarian, contractualist, or preference-respecting reasons).

Most damningly, in its stronger consequentialist forms (e.g. classical utilitarianism), the view described above would imply that it would be right to force the people in the world of voluntary non-procreation to procreate, again contrary to their preferences and their consent. Or rather, this implication would follow provided that forceful action is the only way to bring about a large happy future population. Let us stipulate that it *is* the only way, for the sake of argument. Specifically, let us assume that we can push a button that would force people in this hypothetical world to procreate such that they create a vast number of new happy people. (To get a stronger version of this thought experiment, we could assume that the button also forces all existing people to experience intense suffering for a significant fraction of their lives, while also assuming that this suffering will not affect the future generations that they are forced to give rise to.)

This puts consequentialists who endorse the moral importance of creating happy lives in a tricky dilemma. They can either push the button and force the people in this hypothetical world to procreate against their will, which is arguably implausible — at the very least, it seems questionable to say that it would be morally right to force people in this hypothetical world to procreate (especially if doing so would also force these people to endure large amounts of intense suffering). Alternatively, such consequentialists could argue that the moral importance of creating happy lives is not sufficient to override the consent (let alone the intense suffering) of an existing population in which everyone has strong preferences against procreating. Yet if they opt for this second response, it would seem to follow that the moral importance of creating happy beings is not that great, since it means that the creation of countless generations of happy people cannot justify overriding the preferences of a single generation (or in the stronger version of the thought experiment: that it cannot override the preferences and the intense suffering of a single generation).

Either way, the dilemma questions the moral importance of creating happy lives for their own sake.

Caution about “appeals to extinction”

A concluding recommendation is that “appeals to extinction” in favor of the value or moral importance of creating happy lives need to be made with care. That is, when proponents of the moral importance of creating happy lives seek to support their view by appealing to extinction scenarios, they should ideally refer to the scenario described in “a world of voluntary non-

procreation" or a similar scenario that avoids extraneous factors such as preference frustrations and violence (Knutsson, 2015).

When these distorting factors are removed, the "appeal to extinction" in favor of the moral importance of creating happy lives seems to lose much of its force.

References

Holtug, N. (2004). Person-Affecting Moralities. In Ryberg, J. & Tännsjö, T. (eds.), *The Repugnant Conclusion*. Kluwer Law International.

Knutsson, S. (2015). The 'Asymmetry' and extinction thought experiments. Ungated

Lazari-Radek, K. & Singer, P. (2014). *The Point of View of the Universe: Sidgwick and Contemporary Ethics*. Oxford University Press.

Mogensen, A. (2022). The weight of suffering. Ungated

Ord, T. (2020/2021). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Vinding, M. (2022). Reply to Chappell's "Rethinking the Asymmetry". Ungated

Minimalist versions of objective list theories of wellbeing

My colleague Teo Ajantaival is currently writing an essay on minimalist views of wellbeing, i.e. views according to which wellbeing ultimately consists in the minimization of one or more sources of illbeing. My aim in this post is to sketch out a couple of related points about objective list theories of wellbeing.

I should note that these views of wellbeing are not necessarily the ones that I myself consider most plausible, but I still think it is worth highlighting how one can endorse minimalist versions of objective list views, which are arguably the most plausible versions of these views.

Objective list theories of wellbeing

In their typical formulations, objective list theories say that wellbeing consists in having a variety of objective goods in one's life. These purported objective goods could include knowledge, health, virtuous conduct, personal achievements, and autonomy. Note that a key claim of objective list views is that these purported goods contribute independently to a person's wellbeing, and not merely by means of satisfying our desires or improving our hedonic states.

Minimalist versions of objective list theories can be largely equivalent to standard versions of these theories, in the sense that they may include essentially the same list of objective goods, except that these "goods" are construed in terms of the absence of bads. That is, minimalist versions of objective list theories understand wellbeing as consisting in the absence of objective bads, rather than consisting in the presence of objective goods (which do not exist on the minimalist conception of wellbeing).

For example, rather than seeing autonomy as an objective good that can bring our wellbeing above some neutral level, the absence of autonomy is seen as an objective bad that detracts from our wellbeing, placing us below a neutral or unproblematic state of wellbeing; and having full autonomy can at most bring us to an untroubled or unproblematic level of wellbeing.

Similarly, rather than seeing health as an objective good that takes us above a neutral or unproblematic state, the lack of health is seen as an objective bad, and complete health can at most bring us to an untroubled level of wellbeing. Rather than seeing virtue as an objective good that contributes positively to wellbeing, vice is seen as an objective bad that contributes negatively, and virtue may be understood as the mere absence of vice (cf, Kupfer, 2011; Knutsson, 2022, sec. 4). And so on for any other purported objective good.

Harms of premature death

It is worth noting that minimalist versions of objective list views can support the view that premature death is bad, and they can do so in many ways. For not only may these views consider premature death to be bad because it entails many other objective bads (e.g. death would prevent us from completing our life projects), but these views may also see premature death itself as an objective bad. Minimalist objective list views may thus see a far greater harm in death than do more optimistic views of wellbeing.

A possible foundation for a negative utilitarian view

Note also how these minimalist views could be incorporated into a version of utilitarianism that might be more intuitive than most other forms of utilitarianism. That is, minimalist objective list views could form the basis of a negative utilitarian view that says that we ought to minimize illbeing, understood as the minimization of the independent bads that contribute to illbeing.

Such views can avoid many of the counterintuitive implications of classical utilitarianism — e.g. that we should force people to bring about new happy beings in hypothetical worlds where nobody wants to create such beings, even at the price of increasing extreme suffering — while also avoiding the conclusion that early death is always morally best for any individual's own sake in isolation, as implied by some other forms of negative utilitarianism.

Of course, minimalist theories of wellbeing are not tied to any particular view of ethics, but this ethics-related point seems worth stressing since discussions of negative utilitarianism often overlook the possibility of basing utilitarianism on the theories of wellbeing outlined above.

Concluding remarks

My aim in this post has not been to provide arguments in favor of minimalist objective list theories over competing "objective goods" theories of wellbeing. Such arguments could seek to establish that it is more plausible that the purported objective goods found in objective list theories are in fact objective bads to be avoided, or they could seek to establish that purported objective goods only contribute instrumentally to wellbeing by reducing objective bads. Yet such arguments are beyond the scope of this brief post, whose aim has been of a more modest nature, namely to draw attention to a group of minimalist views that is often overlooked.

Minimalist views can be construed in many different ways and can accommodate a wide range of intuitions, which makes them a far richer and more flexible class of views than is commonly

acknowledged. Consequently, it is worth avoiding the common mistake of dismissing all minimalist views with reference to arguments that only apply to a relatively narrow subset of these views.

Clarifying lexical thresholds

Summary

Views that assign lexical disvalue to extreme suffering are often framed and discussed in ways that make such views seem implausible. For example, it may be claimed that lexical views imply that lexicality must abruptly "kick in" at some precise level of painful stimulus, such as when a scorching object reaches a certain temperature.

Yet this is not true. Not only can one defend lexical views without abrupt breaks, but it is also possible to formulate lexical views in terms of pain intensity or the overall disagreeableness of experiential states, as opposed to stimulus intensity. Lexical views are arguably best framed and discussed in terms of such mental states rather than external stimulus. Moreover, the rejection of lexical views can hardly be considered most plausible by default, even if lexical views appear to have counterintuitive implications, since alternative views may rest on premises and imply corollaries that are less plausible, all things considered.

Introduction

If some amount of a given bad is worse than any amount of some other bad, the former is said to be *lexically worse* than the latter, and there is said to be a *value lexicality* between the two bads.

Many have defended the view that certain bads, such as a full day of the most extreme suffering, are lexically worse than comparatively trivial bads, such as a mild headache (see e.g. Mayerfeld, 1999, pp. 178-179; Leighton, 2011, ch. 9; Tomasik, 2013; 2015a; Klocksiem, 2016; Gloor, 2016, sec. II; Vinding, 2020b, ch. 4-5).

This view is often countered with the challenge of explaining how this lexicality can emerge. For example, if we construct a sequence of bads that fall in between the two bads in question, at which point, if any, is the lexicality supposed to kick in?

One may outline such a sequence in the following way (inspired by Tomasik, 2013):

Say that the suffering of burning at 500 degrees Celsius for a full day is claimed to be worse than any amount of experience-moments of merely being uncomfortably hot at 50 degrees Celsius. We can then construct a sequence where the sufferer burns at 490 degrees for 10 days, 480 degrees for 100 days, 470 degrees for 1,000 days, and so on, until we reach 50 degrees for 10^{45} days.

If we grant that each step in this sequence is worse than the previous one (which is, of course, a big "if"), it might seem to follow that the value lexicality claimed at the outset cannot obtain.

But this is not necessarily the case. The example above is deceptive, as it is unclear whether it pertains primarily to *experiences* or to variations in a given *stimulus*. Confusion on this point can make lexical views appear needlessly implausible.

The pitfalls of focusing on stimulus

It is not difficult to see why it is tempting to speak in terms of changes in stimulus when discussing this issue. Such changes are palpable and concrete, and we can often measure them in precise, quantifiable terms.

Nonetheless, such a framing is ultimately misleading. What we care about, and what the lexical views we are here concerned with pertain to, is not really stimulus per se, but rather the *experiences* elicited by the stimulus in question. And the truth is that changes in external stimuli do not necessarily track changes in experiences all that well.

For example, it can be far worse to experience a moderately intense stimulus for a long duration (say, being burned by an iron rod of 100 degrees Celsius) than to be subject to a more intense stimulus for a tiny duration (e.g. being burned by a 200 degree hot object for a split second). Indeed, the latter may not even result in any suffering at all if the duration is sufficiently short.

This also illustrates why the sequence argument outlined in the previous section is problematic. It fails to clearly distinguish 1) innumerable instances of experience-moments that each feel "uncomfortably hot", and 2) the experience of being burned at 50 degrees Celsius for a long duration. In the latter case, the resulting experience could — and all but surely would — eventually get a lot worse than merely "uncomfortably hot".

It is thus wholly consistent to maintain that a small amount of intense suffering is worse than any amount of merely "uncomfortably hot" experiential states, yet to *not* consider a small amount of intense suffering worse than having to go through a very long duration of being exposed to, say, a temperature of 50 degrees Celsius. After all, the latter may eventually give rise to more than just a small amount of intense suffering.

The pitfalls outlined above can be avoided by framing discussions of this issue explicitly in terms of psychological states and evaluations rather than stimulus, and by being clear about the significance of this distinction.

Lexical views based on consent

An example of a view that entails value lexicality between extreme and mild suffering is the view that a single instance of unbearable suffering — i.e. suffering so intense that the sufferer is unable to consent to it — is worse than any amount of mild, bearable suffering (Tomasik, 2015a; Vinding, 2020b, ch. 4-5 and sec. 6.7).

This is a view that centers explicitly on a psychological evaluation. And note how such a view gives rather plausible replies to the sequence argument outlined above. On this view, there is no set point in terms of temperature or duration at which lexicality must kick in. It simply comes down to the psychological state of the sufferer, which in turn depends on many factors, such as the intensity and duration of the noxious stimuli, as well as the overall constitution of the sufferer. Such flexibility seems a desirable feature of a lexical view.

Note also that defenders of this view need not maintain that any sharp steps are found between "wholly bearable" and "wholly unbearable" suffering (although one can hold such a view, cf. Klocksiem, 2016). After all, psychological states and evaluations are commonly fuzzy and tend to manifest in varying degrees. And the same may well apply to evaluations of "unconsentability" and unbearableness in particular: they plausibly come in degrees.

Yet this need not prevent us from drawing a clear distinction between states of suffering that are perfectly bearable versus states that are completely unbearable (to think otherwise is to commit the continuum fallacy). Nor does it prevent us from considering a single instance of unbearable suffering worse than any amount of suffering or discomfort that is wholly bearable.

Consent is just one example of a psychological state or construct on which one can base a plausible suffering-focused lexical view. An alternative option is to phrase such views in terms of the intensity of pain (Klocksiem, 2016), or in terms of distinct experiential components of suffering (Vinding, 2020a). A consent-based view can differ from a view based on pain intensity in that one could hold that a constant pain intensity can be endurable for some time, yet become unbearable eventually (at the level of the sufferer's overall psychological state), in which case lexicality would kick in after a certain duration despite the intensity of the pain, as an experiential component, being the same.

The rejection of lexicality is not plausible by default

Lexical views are often rejected on the grounds that they appear to have some strange and implausible implications. But this is not in itself a good reason to reject lexical views. For beyond the fact that some of the purported implications of such views need not actually follow (e.g. that one

must accept some kind of abrupt break), it seems that alternative views are bound to have some highly counterintuitive implications of their own. For example, many people may find it even more implausible that a sufficiently large amount of mild discomfort could ever be worse than a full day of intense suffering (cf. Tomasik, 2015b).

The view that the disvalue of many states of mild discomfort can be represented with real numbers and added together such that they are worse than a full day of extreme suffering is itself a view that rests on highly non-obvious premises. For example, it assumes that the disvalue of different levels of discomfort and suffering can, in principle, be measured along a cardinal scale that has interpersonal validity, and further assumes that these value entities occupy the same dimension (so to speak) on this notional scale. These premises are "highly controversial and widely rejected" (Knutsson, 2016), and hence they, too, require elaborate justification.

References

Arrhenius, G. & Rabinowicz, W. (2015). Value Superiority. In Hirose, I. & Olson, J. (eds.), *The Oxford Handbook of Value Theory*. Oxford University Press.

Gloor, L. (2016/2019). The Case for Suffering-Focused Ethics. Ungated

Klocksiem, J. (2016). How to Accept the Transitivity of Better Than. *Philosophical Studies*, 173(5), pp. 1309–1334.

Knutsson, S. (2016/2018). Measuring happiness and suffering. Ungated

Knutsson, S. (2021). Many-valued logic and sequence arguments in value theory. *Synthese*, 199, pp. 10793-10825. Ungated

Leighton, J. (2011). *The Battle for Compassion: Ethics in an Apathetic Universe*. Algora Pub.

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Tomasik, B. (2013/2019). Three Types of Negative Utilitarianism. Ungated

Tomasik, B. (2015a/2017). Are Happiness and Suffering Symmetric? Ungated

Tomasik, B. (2015b). A Small Mechanical Turk Survey on Ethics and Animal Welfare. Ungated

Vinding, M. (2020a). Lexical views without abrupt breaks. Ungated

Vinding, M. (2020b). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Lexicality between mild discomfort and unbearable suffering: A variety of possible views

It appears to be a common intuition that no amount of mild discomfort can be worse than extreme suffering (i.e. that extreme suffering is lexically worse than mild discomfort). Yet this view is often considered implausible due to continuum arguments. Such arguments go roughly like the following: For any duration of any state of suffering, there is a slightly less intense state of suffering that would be worse if extended for a sufficiently long duration. Therefore, the argument goes, by continually lowering the intensity and increasing the duration of suffering, we will eventually end up with a large amount of mild discomfort that is indeed worse than the initial state of suffering in the sequence, no matter how intense that initial state may be.

The aim of this essay is to present a number of views that entail value lexicality between mild discomfort and extreme suffering, and which reject the argument outlined above in different ways. My overall point is that continuum arguments are much less compelling than they are often taken to be, and that it is reasonable to challenge standard assumptions that beg the question against lexical views.

A widely accepted lexical view: Lexicality between things that matter and things that don't

When discussing lexical views, it may be helpful to start by considering a form of lexicality that is endorsed by many views, and which is so trivial that it can be easy to overlook that it is in fact an instance of value lexicality.

One way to describe this common form of lexicality is with reference to purely hedonistic axiologies. Standard formulations of these axiologies entail that there is a lexical difference between the value of a hedonically neutral state and a slightly disagreeable state of consciousness, meaning that no amount of hedonically neutral states — including states that involve (non-hedonic) preference violations — will be intrinsically worse than a single state of consciousness that is slightly hedonically disagreeable.

Indeed, a wide range of axiologies imply value lexicality between entities that are assigned a neutral value and entities that are assigned a negative value, regardless of how slight the negative value may be. (Though axiologies often differ strongly as to what should be regarded as having neutral

versus non-neutral value, meaning that different axiologies commonly entail value lexicality between each others' purported bads, as hinted above with the case of non-hedonic preference violations.)

This trivial example of value lexicality serves to highlight two points. First, it shows that value lexicality is not some strange and alien feature that virtually no axiological view endorses, but instead a feature that is entailed by many views, at least in some form.

Second, the example highlights an important point about value lexicality, which is that lexicality can be both abrupt *and* gradual in nature. That is, the step from neutral to non-neutral value may be gradual in that a bad state of consciousness might be bad to just a tiny degree, while still being lexically worse than the slightly less bad state that is a perfectly neutral state. And such gradual yet abrupt lexicality is not only theoretically possible, but also quite plausible and widely endorsed, at least in the case of neutral versus slightly non-neutral states or value entities.

Abrupt but gradual lexical views

One possible reply to the continuum argument against value lexicality between unbearable suffering and mild discomfort is to argue that an abrupt but gradual threshold likewise exists between mildly bad and intensely bad states.

For example, Justin Klocksiem defends an absolute lexical threshold between "discomfort" and "genuine pain", and argues that such a threshold is plausible both in phenomenological terms and because it helps avoid a number of implausible conclusions in value theory (Klocksiem, 2016). In Klocksiem's view, the step from discomfort to genuine pain is an abrupt one (in evaluative terms), but it is gradual in that the intensity of genuine pain still increases in a gradual manner.

One may defend several absolute thresholds of this kind between different types of experiential states. Thus, besides Klocksiem's threshold between "discomfort" and "genuine pain" — and in addition to the widely endorsed threshold between "neutrality" and "discomfort" — one may further defend an abrupt but gradual lexical threshold between, say, "genuine pain" and "unbearable suffering". And perhaps additional thresholds beyond that.

In terms of geometric visualization, one may think of the crossing of each lexical threshold as a tiny step in the direction of a new dimension of experiential disvalue, akin to the step from an "origo state" of perfect neutrality toward the mildest of discomfort. (Such lexically distinct dimensions could, for instance, correspond to the activation of different circuits of aversive experience, or to different kinds or combinations of painful emotions.)

An argument in favor of additional such thresholds, beyond the threshold between neutrality and discomfort, is that it seems a priori implausible that all experiences — even when they are mediated by different neural circuits — must necessarily be ordered along a single uniform axis.

Moreover, in phenomenological and a posteriori terms, one may argue that mild discomfort and unbearable suffering are even more dissimilar in their experiential character than are states of neutrality and discomfort, and hence that it is at least as plausible that mild discomfort and unbearable suffering likewise occupy lexically distinct dimensions of experience. (Though this argument is also compatible with non-abrupt lexical views, which we will explore shortly.)

Disanalogous forms of lexicality?

One might object that the abrupt forms of lexicality outlined in the previous section are fundamentally different from the lexicality between "things that matter and things that don't". After all, in the latter case, we are talking about a difference between states that are perfectly neutral versus states that are not, whereas the forms of lexicality explored in the previous section involve lexicality between states that both (or all) entail some disvalue.

However, the difference between these cases is less significant than our standard unidimensional conceptions of disvalue might suggest. After all, on the views outlined in the previous section, the abrupt difference between lexically distinct states of disvalue is also a zero-to-one difference in an important sense. For example, to take a simple toy model, a way to think about abrupt but gradual lexicality could be that discomfort is mediated by activity in neural circuit *C1*, while genuine pain is mediated by activity in *C2*, and the lexicality between them would then occur when we go from merely having activity in *C1* to having some activity in *C2* as well. There is a similar "nothing-to-something" step in a new dimension. (Again, the views outlined in the previous section are not predicated on any particular claim about neural circuits or the like, but the broader point about seeing lexical thresholds as representing a step into a new dimension of disvalue does apply to all of these views.)

Non-abrupt lexical views

In contrast to the views explored in the previous section, there are lexical views that entail no abrupt thresholds. Several such non-abrupt lexical views have been proposed, two broad examples of which are outlined below.

Diminishing marginal disvalue of bads

One view that has been suggested is to assign diminishing marginal disvalue to instances of the same bad such that the total disvalue converges to a certain limit. This would mean that a large number of identical bads never get to be worse than a single instance of a sufficiently severe bad (see e.g. Carlson, 2000; Rabinowicz, 2003).

An argument against these views might be that it seems implausible and ad hoc to say that the disvalue of adding a given bad should depend on how many similar bads already exist. Yet a possible reply, or version of these views, could hold that the overall disvalue only diminishes in comparison to worse bads, where one may argue that diminishing marginal value is not implausible.

Specifically, one may argue that an additional bad should always add the same amount of disvalue as long as we are comparing the same kind of bad, whereas such linear addition becomes implausible when we compare bads of a different kind, such as suffering of different intensities. For example, one could reasonably argue that adding more instances of mild pain — while bad — is not bad in the same way as is increasing the intensity of pain, which arguably represents an altogether different parameter of badness (Leighton, forthcoming, "The map and the territory"). One can thus turn the objection above on its head and argue that it is more ad hoc and unwarranted to claim that the disvalue of many instances of the same pain must add up linearly in the context of inter-intensity comparisons.

A way to still assign unique and impartial quantities of disvalue to outcomes on these views might thus be to distinguish different kinds of badness, namely "instance badness", "intensity badness", and the more complete "instance + intensity badness" (cf. Leighton, forthcoming). Thus, speaking only in terms of "instance badness", disvalue may plausibly increase linearly as more instances of the same bad are added. Yet in terms of the combined metric of "instance + intensity badness", the added disvalue of the same bad may coherently be diminishing because we are implicitly comparing it to more intense states of suffering (even if these are merely potential states). Again, in value comparisons across different intensities of suffering, one could argue that this addition scheme, although not perfect, is at least more plausible than is a scheme of linear addition that renders many bearable discomforts worse than unbearable torment. (After all, the latter scheme is not plausible by default.)

Rejecting real numbers and strict "better-or-worse" answers

An alternative and in my view more plausible view is to reject that disvalue is best represented with real numbers, and to further reject that bads with non-identical disvalue must be either strictly better

or strictly worse than one another. Instead, one may allow differences in disvalue to be vague or imprecise (cf. Qizilbash, 2005), or allow different bads to have a certain *degree* of worseness relative to each other, where this degree might assume values between 0 to 1 (cf. Knutsson, 2021).

These degrees can also be extended to lexicality itself. For instance, one may hold that lexicality between two different bads is plausible, or true, to the degree 0.3, rather than insisting that the plausibility or truth degree of lexicality must be exactly 0 or 1 (cf. Knutsson, 2021).

Note that these degrees can express either a subjective degree of plausibility that one assigns to lexicality between the bads in question (i.e. the degree to which one subjectively endorses lexicality), or an objective truth degree of lexicality between the bads. The points I make below apply equally to both interpretations, which are equivalent at the purely formal level (cf. Knutsson, 2021, "Introduction"). (I likewise use the terms "truth degree" and "degree of plausibility" equivalently below, without taking a stand on the interpretation issue.)

The views described above allow for more flexible and refined views, and they can entail lexicality between bads without any abrupt thresholds. For example, one may hold that it is plausible to degree 0.7 that some state of suffering, *S0*, is worse than some slightly less intense state of suffering, *S1*, and further hold that it is plausible to degree 0.01 that *S0* is lexically worse than *S1*. And one may then similarly consider it plausible to degree 0.01 that *S1* is lexically worse than a still less intense state of suffering, *S2*, which may in turn have the same relation to *S3*, and so on, all the way up to, say, *S100*.

The combined truth degree of lexicality across such a sequence can obviously be construed in myriad ways. A simple toy model might be to say that the truth degree of lexicality is additive throughout the sequence, such that, for instance, *S0* is lexically worse than *S5* to degree 0.05, while it is lexically worse than *S10* to degree 0.10, etc.

An alternative view would be to say that the truth degree should increase by a certain factor — perhaps a factor of 10. In that case, one could hold that the truth degree of lexicality between *S0* and *S100* is 1 (the upper limit), while also holding that the truth degree of lexicality between any two adjacent states of suffering in this sequence is only 1/(10^100). (And if we were to introduce intermediate states of suffering between *S0* and *S1*, the truth degree of lexicality between them could be smaller still, such that it converges to 0 as we approach exactly the same state of suffering.)

Thus, in concrete terms, these views can coherently endorse a complete value lexicality between states of mild discomfort and unbearable suffering, while not endorsing it between highly similar states of suffering.

There may be various reasons to opt for views that allow for such degrees of plausibility. One reason might simply be that there is little justification for *not* allowing degrees of plausibility or truth. In the absence of any positive justification for a black-or-white picture of better or worse, it seems natural to reject such a restrictive view in favor of a more nuanced range of possible judgments.

A more substantive reason to favor graded evaluations might be that they can feel more apt and precise in real-life attempts to compare similar states of suffering. In particular, it seems likely that people who experience different states of suffering directly would, at least in some cases, find it more accurate to rate the relative badness of these states in terms of degrees rather than in strictly binary terms — and they might even prefer to use a range of degrees (cf. Mayerfeld, 1999, p. 29; Parfit, 2016, p. 113).

Another substantive reason is that some philosophers have defended views that entail that the suffering of the worst-off consciousness-moment always has lexical priority compared to less intense states of suffering (Mendola, 1990; Ryder, 2001, pp. 28-29). These views essentially entail an abrupt lexical threshold at each gradual worsening of suffering with truth degree 1. One can reasonably argue that this truth degree is much too high. Yet conversely, given that sensible people have defended this view, and given that it can appear to have at least *some* degree of plausibility to say that the maximum sufferer deserves overriding priority, one can likewise argue that a truth degree of strictly 0 seems too low, and that a non-zero truth degree such as 0.01 or 1/(10^100) would be more plausible.

Representing disvalue with real numbers: An unexamined assumption?

The discussion above raises important questions concerning how to best represent disvalue. It is commonly assumed (e.g. among utilitarians) that it is plausible to represent disvalue with real numbers. Yet it seems to me that this assumption is often made without much justification, and without acknowledging that there are reasonable alternatives.

In particular, the assumption that we can represent disvalue with real numbers gives rise to many counterintuitive implications, and it seems that much time is spent grappling with those implications, while comparatively little time is spent questioning the initial assumption that gives rise to these issues.

After all, certain frameworks will rule out value lexicality from the outset. For instance, if we assume that the disvalue of any aversive state can be represented with a negative real number, and further assume that total disvalue should increase linearly as we add more such states (also in inter-

intensity comparisons), it follows trivially that sufficiently many states of mild discomfort can be added up to be worse than any state of extreme suffering. But what is not trivial is whether this set of starting assumptions is plausible to begin with.

In this respect, it is worth being aware of potential biases due to certain ways of thinking that have become second nature to us. We are, after all, very much used to thinking in terms of real numbers and standard addition, which is obviously valid in many (other) contexts. One person with 100 dollars does indeed have as much money as 100 people who each have 1 dollar. But suffering that is rated as having an intensity of 100 on an ordinal scale of pain intensity does *not* similarly represent a sum of 100 mild-intensity suffering stacked on top of each other (as a purely descriptive matter), and hence it is not obvious whether the disvalue of many mild states of suffering can legitimately be summed up in this way either, as an evaluative matter (Leighton, forthcoming, "The map and the territory").

Therefore, even if we think that it seems plausible to represent disvalue with real numbers, it seems worth at least being open to the possibility that other views might ultimately be more plausible — not least given that our familiar ways of thinking may bias us toward the use of real numbers and toward prematurely dismissing less familiar alternatives, akin to how strong familiarity with a particular numeral system can make it seem like the obviously "right one".

It would be a shame if we allowed certain ingrained conceptual frameworks to covertly dictate our views of what matters and what is most worth prioritizing.[7]

References

Carlson, E. (2000). Aggregating harms – Should we kill to avoid headaches? *Theoria*, 66(3), pp. 246-255.

Klocksiem, J. (2016). How to Accept the Transitivity of *Better Than*. *Philosophical Studies*, 173(5), pp. 1309-1334.

Knutsson, S. (2021). Many-valued logic and sequence arguments in value theory. *Synthese*, 199, pp. 10793-10825.

Leighton, J. (Forthcoming). *The Tango of Ethics*.

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Mendola, J. (1990). An Ordinal Modification of Classical Utilitarianism. *Erkenntnis*, 33(1), pp. 73-88.

7 For their feedback on this post, I am grateful to Teo Ajantaival, Tobias Baumann, Anthony DiGiovanni, Michael St. Jules, Simon Knutsson, and Winston Oswald-Drummond.

Parfit, D. (2016). Can We Avoid the Repugnant Conclusion? *Theoria*, 82(2), pp. 110-127.

Rabinowicz, W. (2003). Ryberg's Doubts About Higher and Lower Pleasures: Put to Rest? *Ethical Theory and Moral Practice*, 6(2), pp. 231-237.

Ryder, R. (2001). *Painism: A Modern Morality*. Open Gate Press.

Qizilbash, M. (2005). Transitivity and Vagueness. *Economics and Philosophy*, 21(1), pp. 109-131.

Lexical priority to extreme suffering — in practice

Some ethical views grant a lexical priority to the prevention of extreme suffering over mild forms of suffering, meaning that the prevention of extreme suffering takes precedence over the prevention of mild suffering.

Such views have been claimed to have implausible practical implications. For instance, one objection is that such a lexical priority implies that we should neglect all endeavors that do not aim directly at the reduction of extreme suffering. My goal in this post is to reply to a couple of these objections, and to clarify some key aspects regarding how one might think about prioritization in light of lexical views.

Different kinds of de facto lexical views: Theoretical and practical

Before diving into the practical implications, let us start by reviewing some different ways in which one can end up with a de facto lexical view in practice.

Lexicality in theory

Perhaps the most straightforward way to end up with a lexical view is to endorse it directly at the theoretical level. For instance, one may hold that a certain amount of extreme suffering is worse than any amount of mild suffering even in theory. Such views can then be further divided into lexical views that entail abrupt thresholds and lexical views that do not (cf. Klocksiem, 2016; Knutsson, 2016; 2021).

Strong lexicality in practice without lexicality in theory

Alternatively, one may endorse a lexical view in practice without endorsing lexicality at a purely theoretical level. For example, one might hold that the badness of any amount of extreme suffering could be exceeded by a sufficiently large amount of mild suffering *in theory*, while also holding that a single instance of the most extreme suffering is worse than any amount of mild suffering that could ever be physically realized in the accessible universe.

In particular, if one thinks that the badness (or disvalue) of suffering increases superlinearly as its intensity increases, and if one thinks that the worst suffering is sufficiently intense, then one could in effect think that there exists a threshold of suffering intensity such that no physically realizable

amount of sub-threshold suffering could be worse than a single instance of the most extreme suffering (because the disvalue of the latter is so great).[8]

Weaker lexicality in practice: "Lexicality in expectation"

Weaker versions of practical (de facto) lexical views are also possible, and have been endorsed by some thinkers. One such view is what we may call a "lexical in expectation" view, which holds that the *expected* amount of extreme suffering is so large that no physically realizable amount of mild suffering could be worse.

Of course, such a view can take various forms, as one can defend a variety of distributions regarding how much extreme suffering will occur in the future. And these distributions will tend to imply different probabilities assigned to the claim that the amount of extreme suffering in the future (that we can influence) is lexically worse than any physically realizable amount of mild suffering. (It will obviously also vary depending on how we define "mild suffering", as well as on the relative disvalue we assign to different states of suffering.)

Such a view may thus entail that there is, say, a 0.1 percent probability that the amount of extreme suffering in the future could be less bad than the largest amount of mild suffering (e.g. mild headaches) that is physically realizable. Yet note that this would still practically be a lexical view, as it entails that with 99.9 percent probability, no amount of physically realizable mild suffering could be worse than the amount of extreme suffering that will occur in the future. And if we hold the expected amount of extreme suffering up against the *expected* amount of mild suffering rather than what is physically realizable, the probability of "practical extreme suffering dominance" would become much higher still.[9]

Indeed, someone who endorses lexicality at the theoretical level with a significant degree of moral uncertainty could easily end up having a lower credence in lexicality compared to someone who holds the above-mentioned version of the "lexicality in expectation" view with a low degree of moral uncertainty.[10]

---

8 This is assuming that we are talking about suffering that we could causally influence, and assuming the validity of cosmological models that imply that the potential suffering that we can influence is finite. Such assumptions may be wrong, of course. Yet note that one might still endorse strong practical lexicality in an infinite universe, e.g. if one holds that any realistic lower bound of the density of extreme suffering throughout the universe will imply that no amount of physically realizable mild suffering could be worse.

9 A similar point is made in Tomasik, 2013. Philosopher Lars Bergström likewise seems to endorse a form of practical lexicality: "There is nothing realistic that could happen that could counterbalance the bad [e.g. extreme suffering]".

10 Similarly, lexical views that entail gradual thresholds and lexical views that entail abrupt thresholds may ultimately end up being equivalent at the practical level, since empirical uncertainty means that there will be a range of uncertainty as to what qualifies as lexically bad suffering in any case. That is, if one holds that lexicality kicks in abruptly when suffering becomes sufficiently intense, one will still have uncertainty in practice as to which beings experience such "abruptly lexical" suffering, and this range of uncertainty may be roughly equivalent to the range of practical uncertainty that would be implied by a non-abrupt lexical view, even as the latter view would entail a

Practical implications of lexical views

Say we were to accept one of the lexical views outlined above, such that we effectively grant a lexical priority to the reduction of extreme suffering over mild suffering in practice. How, then, should we think about practical ethics and prioritization? Perhaps a good way to address this question is to start by looking at some of the practical objections that have been raised against lexical views.

Devoting resources to a narrow range of endeavors?

One objection against granting a lexical priority to the reduction of extreme suffering over mild suffering is that it implies that we should devote all our resources toward an implausibly narrow range of actions that aim directly at the reduction of extreme suffering (cf. Huemer, 2010, p. 338). Yet there are various reasons why this implication need not — and indeed does not — follow.

A narrow focus need not follow at the theoretical level

It is worth noting that value lexicality between extreme and mild suffering does not imply value lexicality between extreme suffering and other potential bads, such as rights violations or premature death. Hence, granting a lexical priority to the prevention of extreme suffering over mild suffering is compatible with granting a similarly strong priority to the prevention of other potential bads. Furthermore, one might endorse non-consequentialist duties that imply that we should — at least in some cases — pursue other actions than just those that strictly minimize extreme suffering, even if we granted a lexical priority to the reduction of extreme suffering over all other value entities at the axiological level.[11]

A narrow focus does not follow at the practical level

If we disregard the points made in the previous section, and assume that the reduction of extreme suffering is always our sole priority, we still find good *practical* reasons not to devote all our resources toward a narrow range of actions.

For while it is true that the very best ways to reduce extreme suffering *on the margin* will tend to fall within a fairly narrow range of causes, the same is decidedly not true from a broader perspective that includes all the endeavors necessary for humanity as a whole to reduce suffering in effective

graded "priority range" both at the empirical and the evaluative level, cf. Knutsson, 2021; Vinding, 2022b. The empirical uncertainty may be so significant that it causes plausible versions of these respective views to effectively converge in practice

11 In moral philosophy, it is common to distinguish that which has value from that which is morally right or what we have moral duties to do. While some views entail that what is morally right to do is to maximize value or minimize disvalue, other views entail that we have moral duties that do not reflect pure value optimization, cf. Vinding, 2020, sec. 6.4.

ways. After all, if humanity were to change its resource allocation such that it devoted vast amounts of resources to the best causes on the current margin, the marginal analysis would change, and new causes would become more promising on the margin. And if those causes were to be fully covered or even overprioritized, then other things would become more pressing, and so on.

When we look at the totality of endeavors that are necessary for the reduction of extreme suffering — from a broad as opposed to a momentary marginal perspective — we find that they are numerous and diverse indeed. They include the acquisition of knowledge in a wide range of fields, from mathematics to sociology, as well as the skillful application of such knowledge, at every level ranging from grassroots activism to the highest political offices. They also include many endeavors that reduce extreme suffering in rather indirect ways, such as increasing humanity's ability to cooperate, improving humanity's values, and increasing our overall capacity to reduce suffering (Vinding, 2022a, ch. 9).

Moreover, even if we zoom in on a single individual who aspires to reduce extreme suffering as effectively as possible, it is still dubious to claim that such a person should adopt a very narrow focus, for at least two reasons.

First, even if such an individual should ideally focus on a single "most effective cause", there will likely be great empirical uncertainty as to what that cause is, which may warrant broad exploration into a wide range of plausible causes in order to identify that most promising cause (relative to one's talents, motivations, etc.).

Second, if we assume that a given individual had already identified their single "most effective cause", it by no means follows that this single cause will imply a particularly narrow focus. Indeed, competent action within virtually any promising cause — whether it be the abolition of factory farming or the reduction of s-risks due to AI conflicts — will tend to require a wide range of insights and practical implementation skills.

In short, a singular focus on the reduction of extreme suffering does not imply a narrow practical focus (Vinding, 2020, sec. 9.4-9.5).

Ignoring mild suffering?

A related objection is that views that give lexical priority to extreme suffering over mild suffering will imply that we should ignore all (seemingly) mild suffering in practice, which is arguably implausible. Yet there are various reasons why this supposed implication does not follow (Vinding, 2020, sec. 8.11).

First, views that grant a lexical priority to the reduction of extreme suffering still tend to hold that the reduction of mild suffering is valuable when other things are equal. Therefore, if one were wholly uncertain as to the eventual effects on extreme suffering, these views would deem it worthwhile to reduce the mild suffering in question.

It may then be objected that other things are virtually never equal in practice, and hence the impact that our actions have on mild suffering per se should virtually always be completely disregarded in practice. Yet even if we grant that claim, there are still good reasons to reduce mild suffering in practice. One reason is that ignoring mild suffering may condition us to also ignore more intense suffering. In contrast, if we make an active effort to reduce all suffering — including mild suffering — then this likely reinforces a commitment to the reduction of suffering, which suggests that such efforts tend to (slightly) reduce intense suffering in expectation.

Second, and more importantly, there is the point that we face considerable empirical uncertainty. Unlike the theoretical case in which we can simply stipulate that some being experiences mild rather than extreme suffering, the practical reality is that we do not know from the outside which beings are — or shortly will be — experiencing extreme suffering. This means that there is a risk that beings who *appear* to merely experience mild suffering are in fact experiencing extreme suffering. And this is not purely hypothetical.

After all, even in the case of humans, there are many forms of intense suffering that can be difficult to verify based on outward appearances — e.g. a state of severe depression may not look all that bad from the outside; a state of intense paranoia may give little external clues of horror; and, as an extreme case, those who wake up and feel excruciating pain during anesthesia may look wholly unconscious. Yet our uncertainty tends to be much greater in the case of non-human animals, especially when it comes to beings who look less like us, such as birds, fish, and insects. And our uncertainty gets greater still when it comes to new potential forms of sentience.

Hence, when we are confronted with beings who are *most likely* experiencing mild suffering, it still seems right — from our uncertain vantage point — to assign a non-zero probability to the possibility that there are instances of extreme suffering among their experiences. In other words, it seems right to think that there is some amount of extreme suffering *in expectation* among the experiences of those beings.

Our empirical uncertainty thus highlights the importance of thinking in terms of expected value when trying to reduce extreme suffering in practice, and it likewise reveals why lexical views do not entail a discontinuity between (what appears to be) mild suffering and (what appears to be)

extreme suffering at the practical level. Instead, such views entail continuous probabilities — and continuous expected amounts — of extreme suffering among different beings.

Lexical views can thereby end up resembling non-lexical views to some extent, since these continuous probabilities concerning the presence of extreme suffering will tend to render practical priorities more continuous than one might naively assume.[12]

Example: Helping more beings who seem less likely to experience intense suffering

As a toy example, consider a case in which we can either help a billion small beings who, on our best guess, can experience "lexically bad" suffering with 10 percent probability, or we can help a million slightly larger beings whom we believe can experience "lexically bad" suffering with 51 percent probability. (Say that these beings all have equally long lives, and that these respective groups of beings tend to experience their most intense forms of suffering with the same frequency).

While the smaller beings most likely do not experience "lexically bad" states of suffering, whereas the slightly larger beings most likely do, a standard expected value framework would still recommend that we prioritize helping the smaller beings in this hypothetical example. Indeed, such a framework would entail that there is over two orders of magnitude as many instances of "lexically bad" states of suffering — in expectation — among the smaller beings.[13]

This conclusion departs quite radically from a naive decision procedure that would round off the 10 percent credences to zero while rounding the 51 percent credences to 100 (cf. the human tendency to engage in "belief digitization").

To be clear, I am not claiming that the expected value approach outlined above is unassailable, or that it should be our only decision procedure for determining priorities. Indeed, I think there are good reasons not to rely exclusively on this approach. After all, the expected value approach might not be the best way to make decisions when our probability estimates are at a high risk of being unreliable or misguided, which suggests that we may benefit from supplementing our decision procedure with additional heuristics, to help render it more robust to miscalculation (cf. Karnofsky, 2011).

12 One can also wonder whether this practical point might potentially be a distorting factor in people's evaluations of thought experiments such as "torture vs dust specks", since there could from a practical perspective — as opposed to from a purely theoretical perspective — be more extreme suffering (in expectation) in an unfathomably large number of "seemingly mild states of suffering" than in an inconceivably smaller number of "seemingly extreme states of suffering".

13 Note how this line of reasoning may support a focus on reducing intense suffering among insects, given how numerous insects are and given that it hardly seems justified to have extremely low credences in "intense insect suffering". Of course, future beings who may exist in even greater numbers could well warrant greater priority still.

That said, it still seems worth using expected value calculations as a key yardstick — arguably even the main one — in our practical deliberations, and it certainly beats simpler alternatives such as the crude belief digitization approach that simply rounds all credences off to 0 or 100 without any justification.[14]

References

Huemer, M. (2010). Lexical priority and the problem of risk. *Pacific Philosophical Quarterly*, 91(3), pp. 332-351.

Karnofsky, H. (2011). Why we can't take expected value estimates literally (even when they're unbiased). Ungated

Klocksiem, J. (2016). How to accept the transitivity of *better than*. *Philosophical Studies*, 173, pp. 1309-1334.

Knutsson, S. (2016). Value lexicality. Ungated

Knutsson, S. (2019). Lars Bergström on pessimism, ethics, consequentialism, Ingemar Hedenius, and quantifying well-being. Ungated

Knutsson, S. (2021). Many-valued logic and sequence arguments in value theory. *Synthese*, 199, pp. 10793-10825. Ungated

Tomasik, B. (2013). Three Types of Negative Utilitarianism. Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Vinding, M. (2022a). *Reasoned Politics*. Ratio Ethica. Ungated

Vinding, M. (2022b). Lexicality between mild discomfort and unbearable suffering: A variety of possible views. Ungated

14 For helpful comments, I thank Teo Ajantaival, Tobias Baumann, Simon Knutsson, and Winston Oswald-Drummond.

Part II: Replies to Critiques of Suffering-Focused Views

Note on Pummer's "Worseness of nonexistence"

Summary

In "The Worseness of Nonexistence", Theron Pummer makes an interesting argument that suggests that a failure to create new people can be as bad as cutting an existing person's life short. I here briefly sketch out a reply to Pummer that can be made, in some version, from a variety of different views.

Outline of Pummer's argument

The primary aim of Pummer's essay is to defend comparativism, the view that things can be better or worse for merely possible persons. I agree with Pummer that we should accept some version of comparativism — for example, it seems obvious to me that a state of affairs in which a single person is brought into existence only to be tortured for their entire life is worse than a state of affairs in which no individual is brought into existence (non-comparativists cannot say this, as they hold that the two states of affairs are not comparable).

Yet Pummer seeks to establish worseness in the other direction, as he argues that a state of affairs in which a (possible) person does *not* come into existence can be worse than a state of affairs in which this person does come into existence. (Note that various views accept comparativism without accepting the proposed worseness of nonexistence, e.g. Benatar, 1997, 2006; Fehige, 1998; St. Jules, 2019; Frick, 2020; Vinding, 2020, ch. 2).

Pummer's argument is roughly that it seems plausible that death — more precisely, the prevention of future life — can be bad for a being who has just barely come to fully meet all the criteria for being a person. Yet it seems implausible that the badness of preventing future life should disappear completely in the case of a slightly less developed being, i.e. a being that just barely fails to fully meet a given criterion required for personhood. Hence, if we accept that it can be worse for an existent being to have their life cut short, we should also accept that it can be worse to fail to create future life for beings who do not yet exist (Pummer, 2019, sec. 4).

I believe there are several plausible lines of response to this argument that can resist the proposed worseness of nonexistence. One such line of response is to rely on conditional interests.

(Alternatively, one could reply along Epicurean lines by arguing that death is not worse for the person who dies. Arguments for this Epicurean position can be found in Rosenbaum, 1986; Hol,

2019. Note that on an Epicurean view, death can still be bad for instrumental reasons, e.g. due to the loss of the positive roles that a life has for others. And a proponent of the Epicurean view may argue that such instrumental factors are a significant confounder in our evaluations of the supposed worseness of nonexistence for a life in isolation.)

Conditional interests

Conditional interests have been defended as a plausible basis on which to rest our views on population ethics, and for defending the Asymmetry in population ethics (St. Jules, 2019; Frick, 2020). On these views, interests only matter conditional on existing, and hence there is no moral value in creating and satisfying additional interests that otherwise would not have existed (this is related to the antifrustrationist axiology defended in Fehige, 1998).

Gradual emergence of interests

Yet a defense of conditional interests per se is arguably not sufficient to address Pummer's argument, or at least an adapted version of it. For it still seems strange to say that there should be some point at which a being goes from having absolutely no interest in continued existence to suddenly having a very strong such interest. What is further needed to avoid Pummer's discontinuity is that interests emerge gradually.

On such a view, it may be that, say, a human fetus that is three months old has no interest in continued existence, a fetus of four months has a very slight such interest, at five months the interest is somewhat greater, and so on. This view seems in line with the fact that the brain structures that mediate sentience and sentient interests develop gradually (Tawia, 1992). It is also consistent with the widely held view that it is worse to abort a fetus later rather than sooner, e.g. after eight months compared to three, which is supported by various views in ethics and value theory, such as Jeff McMahan's time-relative interest account of the badness of death (McMahan, 2002).

On the view sketched above, there is no point at which an individual's interest in continued existence goes from being wholly unimportant to being as important as, say, a strong interest in continued existence held by a fully developed person; continued existence is only wholly unimportant as long as no rudimentary such interest has developed. Thus, there is no sudden discontinuity, nor any worseness of nonexistence before an interest in continued existence has emerged.

It is worth noting that some views based on conditional interests may imply the rejection of what Pummer calls "weak deprivationism", namely that death can be bad even if there is no desire or preference for continued existence — a premise that I think can reasonably be questioned. Yet conditional-interest views can also be consistent with this premise, as one may hold that the creation of a person creates an interest in continued existence that is not reducible to mere desires or preferences, while still maintaining that no such interest exists for merely possible persons.

Other views with similar replies

One may reply to Pummer's argument in a similar vein based on other views. For example, one may hold that death is an intrinsic harm (as, e.g., W. D. Ross did), and further hold that this harm can vary depending on how maturely developed the being in question is. Such a view, too, would avoid the worseness of nonexistence and the discontinuity implied by Pummer's argument, as the intrinsic harm of death for a being that is (in a morally relevant sense) just barely existent could be very slight, and then increase gradually rather than discontinuously.[15]

15 I'm grateful to Michael St. Jules for useful comments.

Comparing repugnant conclusions: Response to the "near-perfect paradise vs. small hell" objection

Minimalist views of value hold that "the less of a given bad, the better", and further hold that the only form of positive value that exists is the reduction of bads (e.g. unmet needs). Negative utilitarianism is an example of a minimalist view, which specifically says "the less suffering, the better".

An objection sometimes raised against negative utilitarianism and similar minimalist views is that they would (supposedly) imply the wrong choice between the following populations:

1. "Near-perfect paradise": a very large population of extremely happy lives that each contain a slight bad
2. "Small hell": a much smaller population that consists only of maximally hellish lives

Namely, because the large number of slight bads in the first population would make the totality of these bads worse in the aggregate, these views would conclude that the small hell is better than the near-perfect paradise. This seems implausible, and hence — the objection goes — so do these views.

The following are three of the main points I would make in response to this objection.

Only some (of arguably the least plausible) minimalist views imply this conclusion

The objection above assumes a certain view of aggregation (i.e. how we "add up" the bads in question) that is only entailed by certain versions of minimalist views, but not by others. That is, some views of aggregation, called Archimedean or "non-lexical" views, hold that mild instances of a given bad can always be added up so as to be worse than severe instances of this bad, e.g. that a sufficient amount of mild suffering can be "added up" to be worse than extreme suffering.

In my view, such non-lexical views of aggregation are highly implausible, and I think we have strong reasons to reject such views (see e.g. Vinding, 2020, ch. 4). Indeed, I think the objection above is itself a good reason to reject such accounts of aggregation, and to instead favor a view that gives supreme (i.e. lexical) priority to severe bads.

Such lexical views are commonly acknowledged by those who raise the objection above, but it seems that such views are often gestured at as though they are much more problematic than non-lexical views of aggregation. In other words, non-lexical views often seem presented as though they are the most plausible versions of minimalist views (and hence that the objection above is quite devastating to minimalist views in general), whereas I would argue, again, that such views are among the *least* plausible minimalist views (and hence that the objection is not at all devastating to minimalist views in general). In any case, it seems to me that the plausibility of non-lexical views of aggregation is often assumed without adequate justification.

Yet for the rest of this post, I will set aside the (im)plausibility of non-lexical views of aggregation, and simply grant such a view for the sake of argument. What can be said in response to the objection above from the perspective of non-lexical minimalist views? And do non-lexical minimalist views seem more or less plausible than other non-lexical views, such as classical utilitarianism?

What is the bad in question?

An important aspect to clarify is what exactly the relevant bad is. For example, if a view says "the less suffering, the better", and defines suffering as "a negative overall state of experience" (Mayerfeld, 1999, pp. 14-15), it is important that we do not confuse this bad with other supposed bads. So to not miss their mark, objections that are targeted at this view should invoke *this* particular bad, rather than something else, and be carefully formulated so as to not describe this bad in terms that can too easily be interpreted as something else.

For instance, one formulation of the objection above claims that non-lexical negative utilitarianism would favor "arbitrary amounts of torture in order to destroy sufficiently many lives that combine **one pinprick each** [emphasis added] with otherwise blissful and fulfilling immortal lives of rich experience and activity."

But this formulation is potentially confusing, since a pinprick is a form of stimulus, and hence a pinprick need not imply an overall negative state of experience, and can easily be interpreted in a way that involves no such state of experience. The objection would thus be more clear and to the point if it replaced "one pinprick" with "one mildly negative overall state of experience" or the like.

Another formulation of the objection above is phrased in terms of "lives of all-but-perfect bliss, … each enduring an episode of trivial discomfort or suffering (e.g. a pin-prick, waiting a queue for an hour)". Yet even this formulation is potentially confusing, despite being phrased partly in experiential terms, such as "trivial discomfort". For if we speak in terms of a classical utilitarian terminology, "an episode of trivial discomfort", and even "an episode of trivial suffering", could be

misinterpreted to mean that one for a brief moment moves from, say, “100 units” of pleasure to 99 or 90 units of pleasure — or some other, less intense state of pleasure (cf. “suffering” in the sense of suffering a loss of something). And if misinterpreted in this way, the objection again fails to pertain to minimalist views centered on overall negative experiential states.

If one were to illustrate a truly negative dip in numerical terms on a classical utilitarian framework, it would amount to something like:

> +100, +100, −1, +100, +100, …

Such a dip may feel intuitively unrealistic (e.g. because of the buffer effect happiness can have on ordinary sources of discomfort), and the dip might also not intuitively conform to the description of being “trivial” (because of the big absolute difference on this classical utilitarian framework). For these reasons, too, it seems worth being exceptionally clear and precise in how this objection is stated.

In sum, it is important that we do not confuse a negative experiential state with a certain kind of stimulus, or with a mere dip — even a large dip — in something else, such as pleasure. This clarification alone may render the objection above somewhat less implausible (compared to more equivocal versions of the objection). After all, on an empty-individualist framing that sees each consciousness-moment as a distinct person, this clearer formulation renders it apparent that the vast, “near-perfect-paradise” population in fact includes a vast population of person-moments that *do* experience an overall negative state, and who thus *are* genuine victims of sorts.

The corresponding implication of offsetting views is more repugnant

Having made this clarification, we can proceed to ask whether non-lexical minimalist views, such as non-lexical negative utilitarianism, are more or less plausible than analogous offsetting views — i.e. views that entail that a sufficient amount of purported goods can outweigh any bad — such as non-lexical classical utilitarianism.

Specifically, we can ask whether the objection above, directed at minimalist views, is more or less devastating than the corresponding objection to offsetting views:

> **Creating Hell to Please the Blissful**:
>
> Say we have a utopia with a vast population that is maximally blissful for their entire lives, with the exception of a brief moment in which they each experience a *slightly* less than maximally intense state of pleasure (e.g. they consistently experience a stipulated

maximum of "100 units" of pleasure, except for a brief moment in which they experience only 99 units of pleasure).

Non-lexical classical utilitarianism (and similar offsetting views) imply that it would be good to add a smaller population of maximally hellish lives to this population provided that it fully maximizes the pleasure of the (sufficiently) vast population of near-maximally pleasurable lives (because the vast increase in total pleasure would outweigh these maximally hellish lives according to such views).

This implication seems implausible, and hence so do these non-lexical offsetting views.

I would argue that this objection is far more devastating for non-lexical offsetting views than is the corresponding objection against non-lexical minimalist views. After all, in the example raised against minimalist views centered on the reduction of suffering, the vast and mostly happy population does — as clarified in the previous section — include countless (mildly) negative experiences, and thus in a sense includes countless victims (i.e. mildly afflicted consciousness-moments). While helping these countless victims by replacing them with a small hell may seem implausible (in my view unacceptably implausible), it nonetheless seems *less* implausible than does the addition of a small hell to a condition that contains no suffering and no victims, for the frivolous purpose of increasing the pleasure of a vast number of almost maximally pleasurable consciousness-moments.

As Anthony DiGiovanni notes, "misery is still reduced" in the example raised against minimalist views (assuming the non-lexical account of aggregation, that is), which arguably renders this example far less repugnant than the case of creating hell to please the blissful (cf. Vinding, 2020, ch. 3).[16]

16 I am grateful to Teo Ajantaival for useful feedback on this post.

Reply to Gustafsson's "Against Negative Utilitarianism"

This post is a reply to Johan Gustafsson's draft paper "Against Negative Utilitarianism". Gustafsson acknowledges that for many common objections raised against negative utilitarianism (NU), there are corresponding objections that can be raised against classical utilitarianism (CU) (see e.g. Knutsson, 2021a). Hence, as he writes, "these objections have little force when we assess the relative merits of Classical and Negative Utilitarianism" (Gustafsson, 2022, p. 1).

The aim of Gustafsson's paper is to present novel counterexamples against NU that have no analogues in the case of CU. My aim in this post is to show that CU does face such analogous counterexamples, and that these counterexamples are worse than those facing NU. I also argue that views that give overriding importance to the reduction of extreme suffering seem uniquely plausible in light of the counterexamples reviewed here.

Gustafsson's main counterexample: Bliss versus Torture

The main counterexample Gustafsson raises against NU is the following (p. 3):

> *Bliss versus Torture*
>
> You have a choice between the following outcomes, where the same people live for the same duration:
>
> *A* Everyone gets a century of pure bliss followed by a pinprick.[17]
>
> *B* Someone gets a century of torture, and everyone else gets a century of no pleasure and no pain.
>
> Given a large enough population, you ought to choose *B* over *A* according to Negative Utilitarianism.

Gustafsson continues (pp. 3-4):

> Yet *A* seems more choice-worthy than *B* on, basically, any plausible moral metric: *A* is overwhelmingly in everyone's subjective interest (given, as seems plausible, that everyone strongly prefers ending up in *A* to ending up in *B*). *A* is more equal than *B*. The worse-off are better off in *A* than in *B*. There is less torture in *A* than in *B*. And so on.

17 The pinprick in this thought experiment should ideally be replaced with something like a "micro pain", cf. the various pitfalls of focusing on stimuli rather than experiential states.

Moreover, it seems that Bliss versus Torture lacks an analogue for Classical Utilitarianism.

Favoring the most plausible choice by accident?

I will present analogous counterexamples against CU in the next section. But first, I think it is worth clarifying that the counterexample raised above does not apply to all versions of NU. In particular, many lexical versions of NU would maintain that *B* is worse than *A*, because these views hold that the suffering caused by the torture in *B* is worse than any amount of pinpricks. (Gustafsson also provides a counterexample against such lexical versions of NU, which I will reply to below.)

In other words, Gustafsson's counterexample only applies to versions of NU that make certain (highly non-trivial) assumptions about aggregation, which many proponents of NU — and similar suffering-focused views — would firmly reject (see e.g. Gurney, 1887, ch. 4; Mendola, 1990; Ryder, 2001; Leighton, 2011).

Indeed, I think Gustafsson's counterexample derives its force primarily from the repugnance of choosing a century of torture in order to avoid many minor pains — an implication that is also shared by CU when those bads are pitted against each other in isolation (cf. Gustafsson, 2022, p. 2).

A proponent of views that give lexical priority to extreme suffering could thus argue that CU only happens to give the most plausible answer by accident in Gustaffson's example, because the addition of bliss in *A* happens to align the CU choice with these lexical views. And a proponent of such views could then further argue that if we remove the bliss in the example above — such that *A* involves that "everyone gets a century of no pleasure and no pain, followed by a minor pain" — we see that non-lexical versions of both CU and NU are less plausible than views that assign lexical disvalue to extreme suffering (because the non-lexical views would choose the century of torture, i.e. *B*, in that case).[18]

At any rate, it is not correct when Gustafsson writes that his counterexample "cannot, plausibly, be blocked by clinging to the intuition that evil and suffering have greater moral import than goodness and happiness" (Gustafsson, 2022, p. 3). Those who endorse versions of NU that give overriding priority to extreme suffering can indeed hold on to that intuition while avoiding the choice of torturous suffering over many minor pains.

18 A proponent of CU may opt for a lexical version of CU instead, but such a view is hardly more plausible if it still allows torment to be created for the sake of bliss.

Counterexamples to classical utilitarianism

I think there is not only one, but many counterexamples to CU that are similar to Gustafsson's *Bliss versus Torture* example. Below are two quite similar ones.

Counterexample I: Torture for Micro Pleasures

> *Torture for Micro Pleasures*
>
> Assume that a micro pleasure has the same intensity as a micro pain, such that a classical utilitarian would say that these respective states cancel each other out. You have a choice between the following outcomes, where the same people live for the same duration:
>
> *A1* Everyone gets a century of hedonic neutrality followed by a micro pain.
>
> *B1* Someone gets a century of torture, and everyone else gets 50 years of micro pains plus 50 years of micro pleasures, followed by two micro pleasures.
>
> Given a large enough population, you ought to choose *B1* over *A1* according to classical utilitarianism.

One could here make the same claims that Gustafsson makes regarding his counterexample, about how *A1* seems more choice-worthy than *B1* on any plausible moral metric. In particular, it is trivial to see that *A1* is more equal than *B1*, the worse-off are better off in *A1* than in *B1*, and there is less torture in *A1* than in *B1*. And one could argue that *A1* is in everyone's interest, and that most people probably would prefer to live in *A1* rather than *B1*.

Justification for this latter claim can be found in descriptive research on people's preferences regarding tradeoffs between suffering and happiness (or pain and pleasure). Such research finds that people tend to endorse a significant asymmetry between happiness and suffering, especially as far as lives in potential worlds are concerned (Caviola et al., 2022; Contestabile, 2022, sec. 4).

Indeed, one can argue that this example against CU is even stronger than Gustafsson's example against NU, since we are here ultimately allowing torture for the sake of micro pleasures. That is, while allowing torture for the sake of reducing micro pains seems implausible and repugnant (at least to many people), it is arguably *more* implausible and *more* repugnant to allow torture for the sake of creating micro pleasures (cf. Vinding, 2020, ch. 3; 2021).

Brief reply from Gustafsson

In a footnote, Gustafsson brings up essentially the same thought experiment as the one I raised above, and attempts to briefly argue that it is not analogous to his counterexample against NU.

Before proceeding to Gustafsson's argument, I should say that I find it odd that he only devotes a single footnote to the discussion of potentially analogous counterexamples against CU, seeing that his principal claim is that the counterexample he presents against NU has no analogue in the case of CU. Such a claim would seem to warrant elaborate discussion of potential analogues, to convincingly show that they are in fact *not* genuine analogues. As I will argue below, his cursory remarks in his current draft do not succeed in showing this.

Here is what Gustafsson writes in response to the counterexample (p. 4):

> But, if the micro pleasures really have the corresponding intensity to the pinpricks, they should outweigh the pinpricks in [*B1*] (otherwise they wouldn't have the corresponding intensity according to Classical Utilitarianism). If you don't find that intuitive, you may be imagining micro pleasures of too low intensity. Once the intensity is imagined correctly, [*B1*] should then be in most people's subjective interest (or, at least, [*A1*] would not be overwhelmingly in their interest).

This reply is unsatisfactory, since it simply begs the question in favor of CU. After all, a negative utilitarian could similarly argue that a micro pain cannot be outweighed by a micro pleasure (because otherwise the micro pain wouldn't be a genuine micro pain according to NU). That micro pleasures can outweigh micro pains is a key premise that needs to be established, and hence it cannot simply be asserted like this.

More glaring, however, is that Gustafsson's reply wholly ignores the most problematic aspect of this counterexample against CU, namely that it allows micro pleasures to outweigh a century of torturous suffering, which seems uniquely repugnant, and seems to render this counterexample worse than the one he raises against NU (cf. Vinding, 2020, ch. 3; 2021).

Counterexample II: Torture for Everyone for Micro Pleasures

The following counterexample against CU is even more serious, and hence arguably stronger still compared to the counterexample raised by Gustafsson against NU.

> *Torture for Everyone for Micro Pleasures*
>
> You have a choice between the following outcomes, where the same people live for the same duration:

A2 Everyone gets a century of hedonic neutrality.

B2 Someone gets a century of torture and everyone else gets 50 years of torture (i.e. intense suffering) and 50 years of "correspondingly" intense pleasure, followed by one micro pleasure.

Given a large enough population, you ought to choose *B2* over *A2* according to classical utilitarianism.[19]

Again, one can make all the same claims that Gustafsson makes in relation to his counterexample against NU: *A2* is more equal than *B2*, the worse-off are better off in *A2* than in *B2*, and there is less torture in *A2* than in *B2* — in fact, there is far less torture, since *everyone* gets tortured for at least 50 years in *B2*, whereas nobody even suffers in *A2*. Thus, in terms of the "least torture" criterion that Gustafsson mentions in his paper (a highly important criterion, in my view), the difference is much greater in this case than in the counterexample that Gustafsson presents against NU.

It likewise seems plausible to claim that *A2* is in everyone's subjective interest, and that most people would strongly prefer to end up in *A2* rather than *B2*. After all, not only do people generally endorse a significant evaluative asymmetry for "equally intense" pleasure and pain, but this asymmetry seems even more pronounced in the case of intense suffering versus intense pleasure (Caviola et al., 2022, p. 6). Indeed, in one small informal survey (n=99), roughly 45 percent said that they would not endure just one minute of intense suffering for *any* number of happy years added to their life (Tomasik, 2015).

In everyone's subjective interest?

Suppose that a proponent of CU wanted to argue that these counterexamples against CU are not fully analogous to, or not as devastating as, the counterexample that Gustafsson provides against NU. How could a proponent of CU do this?

One strategy might be to focus on Gustafsson's claim that, in the counterexample he provides, "*A* is overwhelmingly in everyone's subjective interest (given, as seems plausible, that everyone strongly prefers ending up in *A* to ending up in *B*)" (Gustafsson, 2022, p. 3).

19 One might object to the name of this thought experiment, since CU is not merely trading torture for micro pleasures in this case, but also for intense pleasure. However, it is still the case that the micro pleasures are what ends up tipping the scale such that CU favors torturing everyone. Besides, this is hardly a fundamental point of contention by the lights of CU (at least in its non-lexical forms), since one could in any case reframe the thought experiment such that CU would torture everyone purely for the sake of micro pleasures — by replacing the intense pleasure with micro pleasures and by shortening the torture by a sufficient duration that is in turn replaced by micro pleasures.

After all, if we take the case of *Torture for Everyone for Micro Pleasures*, one could argue that it is not clear that *everyone* would strongly prefer to end up in *A2* rather than *B2*. It seems likely that at least *some* people would choose *B2* (e.g. if they were asked in a survey).[20]

Yet the same objection can be raised against Gustafsson's original claim. Indeed, an important class of axiological views would say that the individuals in *B* who experience absolutely no pain in their entire lives are better off than all the individuals in *A*, who each experience both bliss and (one) pain. Thus, Epicureans and many Buddhists would hold that the pain-free lives are in fact better, and that the absence of pain (and other problematic states) is in some sense the highest bliss (see also Schopenhauer, 1819; 1851; Fehige, 1998; Geinster, 1998; 2012; Gloor, 2017; Ajantaival, 2021/2022; Knutsson, 2022).

In other words, Gustafsson's claim that "*A* is overwhelmingly in everyone's subjective interest" is rejected by (what many consider) reasonable axiological views, and adherents of these views would likewise deny that "everyone strongly prefers ending up in *A*". In particular, proponents of these axiological views could argue that *A* is not in the subjective interest of the experience-moments that undergo the pain of a pinprick, and then further maintain that these dispreferred experience-moments are not outweighed by pleasure that occurs in other experience-moments. (And those who endorse these views might in any case say that they personally would prefer a hedonically neutral life over any life that contains even the smallest amount of pain or suffering.)

A proponent of CU might instead argue for a weaker claim, namely that *most people* would strongly prefer to end up in *A* over *B* (in Gustafsson's *Bliss versus Torture* example). And one could then further argue that even if most people would also prefer to end up in *A2* over *B2* (in *Torture for Everyone for Micro Pleasures*), it still seems likely that the majority that favors *A* over *B* is significantly larger than the majority that favors *A2* over *B2*.

This claim is, of course, quite uncertain. After all, 50 years of torture is quite a lot to sign up for, and we have reason to expect that the vast majority of people would prefer to not endure 50 years of intense suffering in order to gain 50 years of ("similarly") intense pleasure plus a micro pleasure (cf. Tomasik, 2015; Caviola et al., 2022; and this informal survey).

Yet even if we grant that the "greater majority" claim above is true, what would follow? This "greater majority in one case than the other" criterion seems quite strange and ad hoc, and it is not clear why we should consider it particularly relevant. After all, most people could be wrong about what is good for them, or about what the experiences they choose to undergo will in fact end up feeling like.

20 Note, however, that the informal survey mentioned above tentatively suggests that only a small minority of people would choose *B2*, perhaps around 1 percent or less, Tomasik, 2015.

Moreover, we can reasonably ask why this (seemingly ad hoc) "greater majority" criterion should be given greater importance than other criteria, including the "less torture" criterion mentioned by Gustafsson, on which the difference between *A2* and *B2* is vastly greater compared to the difference between *A* and *B*.

All in all, the second counterexample raised against CU above does not appear to be weaker than, nor meaningfully disanalogous from, the counterexample Gustafsson raises against NU. On the contrary, one can argue that it is considerably stronger overall, not least since it forces CU to accept torture for everyone for the sake of micro pleasures, while Gustafsson's counterexample against NU "only" implies torture for one individual, and only for the sake of reducing (supposedly) greater suffering (in the form of micro pains) — not for the sake of creating pleasure whose absence causes no problem (cf. Vinding, 2020, ch. 3).

Objection against "critical-level lexical NU"

Gustafsson concedes that his counterexample against NU is not applicable to views that grant lexical priority to the reduction of intense suffering. Yet against such views, he presents another counterexample (p. 6):

> *Bliss and Severe Pain versus Almost-Severe Pains*
>
> You have a choice between the following outcomes, where the same people live for the same duration:
>
> *C* Everyone gets a century of pure bliss followed, for someone, by the briefest, least severe pain that counts as severe.
>
> *D* Everyone gets a century of pain that is just slightly less severe than the critical level in severity.
>
> On the lexical view, we ought to choose *D*, no matter how large the population is.

There are basically two kinds of replies to this counterexample, which are sketched out below.

Reply from the perspective of abrupt lexical views

Someone who endorses an abrupt threshold at which lexicality kicks in would argue that it is not, in fact, implausible to accept the conclusion presented by Gustafsson above, as long as we are careful to clarify what the scenarios above entail.

In particular, one could argue that no number of mild discomforts could ever be worse than even one instance of genuine pain, and maintain that a sharp threshold exists between these respective

states (Klocksiem, 2016). So if we adopt this perspective, the choice above can be equivalently rephrased in the following way:

> *C* Everyone gets a century of pure bliss followed, for someone, by the briefest, least severe pain that counts as a genuine pain.
>
> *D* Everyone gets a century of mild discomfort that is just slightly less severe than the least severe pain that counts as a genuine pain.

A proponent of lexicality who maintains that *C* is worse than *D* would have a number of things to say in response to the following claims made by Gustafsson:

> There is a lot more suffering in *D* than in *C*. While the pain in *C* is worse in severity, it is only slightly worse in severity than the pains in *D* which last much longer and afflict an arbitrarily large number of people. The difference in severity between the pain in *C* and the pains in *D* is, we can assume, barely perceptible.

The claim that "there is a lot more suffering in *D* than in *C*" is somewhat vague, and adherents of abrupt lexical views may disagree with it in a number of ways. First, they might disagree that *D* contains any genuine suffering at all, since they might hold that the bad states found in *D* fall short of qualifying as suffering proper (i.e. they may prefer to reserve this term for states that are worse than mere discomfort).

Second, even if a proponent of this lexical view did consider the bad states in *D* to amount to suffering, they would still reject the claim that there is more suffering in *D* than in *C* in the most relevant sense. That is, while there is a greater *duration* of suffering in *D* than in *C* (if we grant that there is suffering in both), there is nevertheless more suffering in *C* than in *D* in terms of the *disvalue* of the suffering, according to the abrupt lexical view.

Gustafsson's other claim — about how the bad state in *C* is just barely worse than those in *D* — can be met with the reply that virtually all views entail some form of lexicality between one state and a barely worse state. For instance, a classical utilitarian would hold that a barely perceptible pain carries greater evaluative significance than any number of hedonically neutral states, including hedonically neutral states and lives that contain features that other axiological views consider intrinsically bad, such as extra-experiential preference frustration, bad motives, bad acts, etc.[21]

21 Such putative non-hedonic bads would presumably all be granted some degree of "expected disvalue" or "choice-relevance" by someone who favors CU within a framework of moral uncertainty (cf. MacAskill et al., 2020), which means that there is a significant parallel between views that integrate CU within a moral uncertainty framework and views that entail lexicality between different bads at the object-level. That is, while CU would entail a lexical value difference between the mildest hedonic state and any non-hedonic state, where the latter has absolutely zero value, this zero-to-one difference between hedonic and non-hedonic states would not persist within an all-things-considered choice-worthiness framework if one assigns non-zero credence to non-hedonistic views. Hence, if one prefers to maintain value lexicality between hedonic and non-hedonic bads (as opposed to accepting tradeoff ratios

In other words, virtually all standard views in value theory entail what we may call abrupt but gradual lexical thresholds, whereby the tiniest change implies value lexicality, and hence this feature is not unique to views that imply lexicality between different bads. (One might object that the generic form of lexicality entailed by most views between neutral and non-neutral states is not comparable to lexicality between different bads; a brief reply to that objection is found here.)

Reply from the perspective of non-abrupt lexical views

The second and perhaps more important reply is to point out that Gustafsson appears to overlook non-sharp lexical views, which his objection does not apply to. That is, one can hold that a single instance of extreme suffering is worse than arbitrarily many mild states of suffering, while also maintaining that there is no sharp lexical threshold between them (see e.g. Knutsson, 2016a; 2016b; 2021b; Vinding, 2022).

Gustafsson has thus not established that it is implausible to endorse views that give lexical priority to extreme suffering. Indeed, such views arguably stand as the most plausible ones in light of the various counterexamples reviewed above.[22]

Appendix: Reply to the rest of Gustafsson's footnote

In his earlier-mentioned footnote, Gustafsson replies to another potential counterexample against CU. I do not find this counterexample nearly as strong as the ones I have provided above, which makes it a bit of a distraction relative to the stronger counterexamples, especially *Torture for Everyone for Micro Pleasures*. But it might nevertheless be worth replying to Gustafsson's objections to this weaker counterexample. (The following counterexample against CU is one of a number of counterexamples I have suggested to Gustafsson.)

Here is what Gustafsson writes (p. 4):

> Magnus Vinding suggests a possible analogue, where, in *A'*, everyone gets a century of hedonic neutrality filled with a very large amount of non-hedonic goods followed by a pinprick and, in *B'*, someone gets a century of torture and everyone else gets a century of hedonic neutrality followed by two pinpricks and four micro pleasures (pleasures corresponding in intensity to a pinprick). Given a sufficiently large population, you ought to choose *B'* according to Classical Utilitarianism.

between them or the like), one would also end up with lexicality between value entities that each have non-zero "expected" disvalue. And if lexicality between different bads is a critical problem for views that entail such lexicality directly at the object-level, it would presumably also be a problem for views that arrive at the same conclusion within a moral uncertainty framework.

22 Gustafsson also raises objections against what he calls Weak NU (Gustafsson, 2022, pp. 7-8). These objections are similar to the ones I have addressed here in my reply to Toby Ord's "Why I'm Not a Negative Utilitarian".

> But this counter-example is disanalogous. It introduces, in addition to pleasures and pains, a third element, namely non-hedonic goods. And, if those goods are good for people, it would merely motivate a switch to a version of Classical Utilitarianism where these non-hedonic goods also contribute to well-being rather than a switch to Negative Utilitarianism.

First, even if this thought experiment introduces an element that goes beyond pleasures and pains, it still implies a highly implausible choice for CU (in its classical form focused on pleasure and pain), and it still fulfills all the same criteria that Gustafsson listed under his counterexample against NU. So the counterexample still appears to speak strongly against (traditional) CU, and it thus seems to merit a response from proponents of (traditional) CU, degree of analogy notwithstanding.

Second, as Anthony DiGiovanni notes, one could argue that there is an important parallel in terms of how CU construes the value of (putative) non-hedonic goods and how NU construes the value of bliss. That is, CU can acknowledge that non-hedonic goods — such as knowledge, relationships, autonomy, etc. — are highly valuable in practice, due to their effects on hedonic states. Yet in abstract thought experiments, CU requires us to ignore these secondary effects, and to only count hedonic states. This theory-practice distinction is why it can seem so counterintuitive that CU would prefer to forgo the creation of arbitrarily many insights, relationships, freedoms, etc. in order to create a single micro pleasure (or so a proponent of CU may argue). Similarly, NU can acknowledge that bliss may be valuable in practice, due to its various positive roles, while nevertheless denying that bliss has any intrinsic positive value. And this, a proponent of NU could argue, is why it can seem counterintuitive to forgo the creation of any amount of bliss for the sake of avoiding the tiniest of pain.

Third, we could simply remove the non-hedonic value entities in the counterexample above, whereby we would still get an analogous counterexample against CU (though one that is still far weaker than *Torture for Everyone for Micro Pleasures*, which I think should be the primary focus of the discussion). In particular, we would have a counterexample in which CU allows torture for the sake of micro pleasures, and in which minimalist axiologies would say that everyone is better off in *A'* compared to *B'*.

Finally, it is not the case that any presumed significance of non-hedonic value entities necessarily motivates a move to other forms of CU. Such non-hedonic value entities could just as well motivate a move to other forms of NU — i.e. broader harm-focused versions of NU — according to which the absence of certain non-hedonic value entities is bad (even if it causes no pain), and where the

presence of these value entities amounts to a less bad state (cf. Fehige, 1998; Benatar, 2006, ch. 2; Knutsson, 2016a).[23]

References

Ajantaival, T. (2021/2022). Minimalist axiologies. Ungated

Anonymous. (2015). Negative Utilitarianism FAQ. Ungated

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

Breyer, D. (2015). The Cessation of Suffering and Buddhist Axiology. *Journal of Buddhist Ethics*, 22, pp. 533-560. Ungated

Caviola, L. et al. (2022). Population ethical intuitions. *Cognition*, 218, 104941. Ungated

Contestabile, B. (2022). Is There a Prevalence of Suffering? An Empirical Study on the Human Condition. Ungated

DiGiovanni, A. (2021a). Tranquilism Respects Individual Desires. Ungated

DiGiovanni, A. (2021b). A longtermist critique of "The expected value of extinction risk reduction is positive". Ungated

Fehige, C. (1998). A pareto principle for possible people. In Fehige, C. & Wessels U. (eds.), *Preferences*. Walter de Gruyter. Ungated

Geinster, D. (1998). Negative Utilitarianism – A Manifesto. Ungated

Geinster, D. (2012). The Amoral Logic of Anti-Hurt (Modified Negative Utilitarianism). Ungated

Gloor, L. (2017). Tranquilism. Ungated

Gurney, E. (1887). Tertium quid: Chapters on various disputed questions. Kegan Paul, Trench, & Co. Ungated

Gustafsson, J. (2022). Against Negative Utilitarianism. Ungated

Klocksiem, J. (2016). How to accept the transitivity of *better than*. *Philosophical Studies*, 173(5), pp. 1309-1334.

Knutsson, S. (2016a). Thoughts on Ord's "Why I'm Not a Negative Utilitarian". Ungated

Knutsson, S. (2016b). Value lexicality. Ungated

23 For helpful feedback, I am grateful to Teo Ajantaival, Tobias Baumann, Anthony DiGiovanni, Simon Knutsson, and Winston Oswald-Drummond. I also wish to thank Johan Gustafsson for engaging in a dialogue about his paper.

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated

Knutsson, S. (2021a). The World Destruction Argument. *Inquiry*, 64(10), pp. 1004-1023. Ungated

Knutsson, S. (2021b). Many-valued Logic and Sequence Arguments in Value Theory. *Synthese*, 199, pp. 10793-10825. Ungated

Knutsson, S. (2022). Undisturbedness as the hedonic ceiling. Ungated

Leighton, J. (2011). *The Battle for Compassion: Ethics in an Apathetic Universe*. Algora.

MacAskill, W. et al. (2020). *Moral Uncertainty*. Oxford University Press. Ungated

Mendola, J. (1990). An Ordinal Modification of Classical Utilitarianism. *Erkenntnis*, 33, pp. 73-88.

Ryder, R. (2001). *Painism: A Modern Morality*. Open Gate Press.

Schopenhauer, A. (1819/1909). *The World as Will and Representation*. Kegan Paul, Trench, Trübner & Co.

Schopenhauer, A. (1851/1973). *Essays and Aphorisms*. Penguin.

Tomasik, B. (2015). A Small Mechanical Turk Survey on Ethics and Animal Welfare. Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Vinding, M. (2021). Comparing repugnant conclusions. Ungated

Vinding, M (2022). Lexicality between mild discomfort and unbearable suffering: A variety of possible views. Ungated

Reply to Chappell's "Rethinking the Asymmetry"

My aim in this post is to respond to the arguments presented in Richard Yetter Chappell's "Rethinking the Asymmetry". Chappell argues against the Asymmetry in population ethics, which roughly holds that the addition of bad lives makes the world worse, whereas the addition of good lives does not make the world better (other things being equal).

"Awesome Lives"

To refute the Asymmetry, Chappell relies on the following claim as a core premise:

> **Awesome Lives:** It is (intrinsically) good or desirable that Awesome Lives come to exist. (Chappell, 2017, p. 168)

Chappell defines an awesome life as "one that exhibits a very high quality of life, along whatever dimensions you take to be normatively relevant" (Chappell, 2017, p. 168).

He continues:

> Awesome Lives is, I think, intuitively highly plausible. When we think about what makes for a good state of affairs, the quality of life for the sentient beings contained therein is surely a (if not the) primary factor. A world full of awesome, flourishing lives is (intuitively) better than a world that lacks these good lives. (Chappell, 2017, p. 168)

This intuition is, of course, rejected by many views, including all views that belong to the broader category of minimalist axiologies — i.e. views centered on the alleviation of bads, which include certain Buddhist axiologies, as well as axiologies inspired by Epicureanism (see also Schopenhauer, 1819; 1851; Benatar, 1997; 2006; Fehige, 1998; Gloor, 2017; Knutsson, 2022b).

Furthermore, those who endorse minimalist axiologies may have plausible explanations as to why Awesome Lives can *seem* intuitively plausible, even to those who ultimately favor minimalist axiologies.

For instance, we may intuitively feel that it is good to bring "awesome lives" into existence, not because we endorse such a thing as positive intrinsic value, but instead because our intuitions fail to respect the radical assumption of "other things being equal" that (counterintuitively) is supposed to rule out all positive externalities.

That is, we may be inclined to endorse the creation of new "awesome lives" chiefly because of the positive roles that these lives would (intuitively) have for others. And such positive external effects may be why we rightly intuit that there is such a thing as positive lives, even if there is no such thing as positive intrinsic value, nor such a thing as positive lives in total isolation.

If we reframed Awesome Lives in terms that made it unmistakably clear that the lives in question have no effects on their surroundings — such as by positing that they are isolated matrix lives — the plausibility of Awesome Lives may be considerably reduced for many people. (To highlight the difference between "awesome lives" that have beneficial effects on others versus "awesome lives" that are thought to be worthwhile for their own sake, I will refer to the latter as "intrinsically awesome lives".)

Another class of views that would reject Chappell's Awesome Lives thesis are views centered on conditional interests. Such views hold that it is good that individuals have a high quality of life and that they have their interests and preferences satisfied *conditional on their existence*, while also maintaining that the addition of new such lives and interests does not make an outcome better, other things being equal (see e.g. St. Jules, 2019; Frick, 2020).

Do the "intrinsically awesome lives" contain suffering or other bads?

An important question to clarify is whether the "intrinsically awesome lives" contain significant bads. After all, Chappell's Awesome Lives thesis is seemingly meant to apply to real-world lives in the world of today — not to purely hypothetical or future utopian lives. Specifically, when Chappell writes about these lives having "a very high quality of life", it seems that he refers to "a very high quality of life" by contemporary standards (Chappell, 2017, p. 168).

Yet the reality is that even the best lives contain significant bads, including (for the most part) significant suffering. And when we consider all these unfortunate aspects of the "intrinsically awesome lives" — e.g. their heartbreaks, losses, sufferings, failures, and eventual death — it becomes even less clear that it is, on the whole, intrinsically good or desirable that such lives come to exist, especially when the absence of these lives causes no problem.

In particular, one could argue that even if the "intrinsically awesome lives" do contain positive intrinsic value, this positive value still cannot outweigh all the worst parts of these lives, such as their most intense suffering, their death, or their worst moral failures (some axiological views hold that the latter also contribute directly to people's wellbeing, see e.g. Hurka, 2001; Knutsson, 2022a).

A Distant Realm

Chappell notes that massive investments are required to create an "intrinsically awesome life", whereas no investment is required to avoid creating a miserable life, and he argues that this practical asymmetry is one of the main explanations of asymmetric intuitions in population ethics.

In support of his claim, Chappell presents the following example:

> **A Distant Realm:** You learn that a new colony of awesome, happy, flourishing people will pop into existence in some distant, otherwise-inaccessible realm, unless you pluck and eat a particular apple. (Chappell, 2017, p. 170)

He continues:

> It strikes me as intuitively clear that you have good reason, in this case, to refrain from plucking and eating the particular apple in question. This suffices to refute the Asymmetry – we *can* have moral reason to bring good lives into existence (or refrain from preventing their existence, which I take to amount to much the same thing in this context). (Chappell, 2017, p. 170)

But again, the question raised in the previous section needs to be raised here as well: Do these "awesome, happy, flourishing people" experience bads that are similar to those experienced by people who have a "very high quality of life" in our world, e.g. significant suffering, loss, death, etc?

If they do not experience such bads, we should be clear that our evaluation of the creation of these lives has limited relevance to procreative decisions that concern beings in the real world who do experience such bads. And one may further argue that the creation of these lives could in any case never be better than their non-creation (for their own sake), even if these lives were intrinsically perfect in every way (cf. Schopenhauer, 1819; 1851; Benatar, 1997; 2006; Fehige, 1998; Breyer, 2015; Gloor, 2017; St. Jules, 2019; Frick, 2020; Ajantaival, 2021/2022; Knutsson, 2022b).

If the lives in question *do* contain significant bads — meaning that we focus on a version of the thought experiment that does have real-world relevance — then one could argue that we have no compelling reason to refrain from plucking and eating the apple. Indeed, one could argue that we have strong reasons in *favor* of plucking the apple, seeing that it would prevent all the significant bads that would be entailed by these lives (e.g. their frustrated preferences, losses, sufferings, failures, and eventual deaths), while the non-creation of the distant realm would cause no problem whatsoever.

Opportunity costs in terms of reducing suffering

Another important question is whether the creation of "intrinsically awesome lives" can ever be justified given the massive opportunity costs it would involve. That is, even if we disregard the suffering and other bads entailed by the "intrinsically awesome lives" themselves, and even if we grant that it can be good to bring "intrinsically awesome lives" into existence for their own sake, it may still be unjustifiable to prioritize the creation of "intrinsically awesome lives" given the opportunity costs in terms of wretched lives and suffering that one could otherwise have prevented (cf. Rachels, 2014; Benatar, 2020).

For example, one may hold that extreme suffering and extremely bad lives can never be outweighed by the addition of "intrinsically awesome lives", even if the latter are thought to be good in isolation (cf. Wolf, 1996; 1997; 2004; Mayerfeld, 1999, p. 178). On a purely consequentialist framework, this would mean that we should devote our resources toward the prevention of extreme suffering and extremely bad lives over the creation of "intrinsically awesome lives".

Chappell offers no arguments as to why we should think that the addition of "intrinsically awesome lives" can justify or outweigh the — always very real — opportunity cost of failing to prevent extreme suffering and extremely bad lives, and he has therefore not established that the creation of "intrinsically awesome lives" (for their own sake) can ever be justified in practice.

Of course, such claims may lie beyond the scope of Chappell's paper, but we should in any case be clear that his argument has limited practical significance. Specifically, it is worth being clear that Chappell provides no case against a practical Asymmetry according to which we can never justify creating "intrinsically awesome lives" at the expense of failing to prevent extreme suffering and extremely bad lives. Such a practical Asymmetry seems both highly reasonable and wholly unaffected by the arguments provided by Chappell.[24]

References

Ajantaival, T. (2021/2022). Minimalist axiologies. Ungated

Benatar, D. (1997). Why It Is Better Never to Come into Existence. *American Philosophical Quarterly*, 34(3), pp. 345-355. Ungated

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

24 For helpful comments, I am grateful to Teo Ajantaival, Anthony DiGiovanni, Simon Knutsson, Winston Oswald-Drummond, and Michael St. Jules.

Benatar, D. (2020). Famine, Affluence, and Procreation: Peter Singer and Anti-Natalism Lite. *Ethical Theory and Moral Practice*, 23, pp. 415-431.

Breyer, D. (2015). The Cessation of Suffering and Buddhist Axiology. *Journal of Buddhist Ethics*, 22, pp. 533-560. Ungated

Chappell, R. (2017). Rethinking the Asymmetry. *Canadian Journal of Philosophy*, 47(2), pp. 167-177.

Fehige, C. (1998). A pareto principle for possible people. In Fehige, C. & Wessels U. (eds.), *Preferences*. Walter de Gruyter. Ungated

Frick, J. (2020). Conditional Reasons and the Procreation Asymmetry. *Philosophical Perspectives*, 34(1), pp. 53-87. Ungated

Gloor, L. (2017). Tranquilism. Ungated

Hurka, T. (2001). *Virtue, Vice, and Value*. Oxford University Press.

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated

Knutsson, S. (2022a). Pessimism about the value of the future and the welfare of present and future beings based on their acts and traits. Ungated

Knutsson, S. (2022b). Undisturbedness as the hedonic ceiling. Ungated

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Rachels, S. (2014). The immorality of having children. *Ethical Theory and Moral Practice*, 17(3), pp. 567-582.

Schopenhauer, A. (1819/1909). *The World as Will and Representation*. Kegan Paul, Trench, Trübner & Co.

Schopenhauer, A. (1851/1973). *Essays and Aphorisms*. Penguin.

St. Jules, M. (2019). Defending the Procreation Asymmetry with Conditional Interests. Ungated

Wolf, C. (1996). Social Choice and Normative Population Theory: A Person Affecting Solution to Parfit's Mere Addition Paradox. *Philosophical Studies*, 81, pp. 263-282.

Wolf, C. (1997). Person-Affecting Utilitarianism and Population Policy. In Heller, J. & Fotion, N. (eds.), *Contingent Future Persons*. Kluwer Academic Publishers. Ungated

Wolf, C. (2004). O Repugnance, Where Is Thy Sting? In Tännsjö, T. & Ryberg, J. (eds.), *The Repugnant Conclusion*. Kluwer Academic Publishers. Ungated

Comments on Mogensen's "The weight of suffering"

Andreas Mogensen's paper "The weight of suffering" presents an interesting argument in favor of the axiological position that "there exists some depth of suffering that cannot be compensated for by any measure of well-being" — a position he calls "LTNU" (Mogensen, 2022, abstract). Mogensen then proceeds to explore how one might respond to that argument and thereby reject LTNU.

My aim in this post is to raise some critical points in response to this paper. As a preliminary note, I should say that I commend Mogensen for taking up this crucial issue regarding the weight of suffering, and for exploring it in an open-ended manner.

"The greatest cost of accepting LTNU"

Mogensen writes (p. 12):

> the greatest cost of accepting LTNU is surely that it appears to support the desirability of human extinction or the extinction of all sentient life (Crisp, 2021).

This seems to be the main objection that Mogensen raises against LTNU. Yet two points are worth making in response to that objection.

First, LTNU need not imply the all-things-considered desirability of human extinction (and how extinction occurs can also be relevant, but more on this below).

Second, the objection rests on the assumption that extinction is bad, which one may reasonably disagree with (especially if extinction happens through, say, voluntary non-procreation).

I will elaborate on these two points in turn.

LTNU need not imply the desirability of extinction

Axiological reasons

As Mogensen observes, "LTNU does not presuppose consequentialism, nor any other theory of normative ethics. It is a fragment of a population axiology" (p. 13). And this axiological principle may be combined with other axiological principles that would imply that extinction is bad, especially if it involves everyone getting killed.

After all, LTNU (as Mogensen defines it) only says something about the relative value of extremely bad lives and lives that (purportedly) have positive welfare levels. It does not say anything about the relative value of extremely bad lives and other potential bads, such as death, murder, rights violations, etc.[25]

In particular, one may hold that (purported) positive goods can never outweigh extremely bad lives, but that other bads, such as those mentioned above, are of comparable disvalue to extremely bad lives. (David Benatar appears to hold roughly such a view, in that he seems to consider death bad in itself, Benatar, 2006, pp. 211-221.)

Indeed, on some preference-based views, one need not even posit additional axiological principles beyond LTNU, since the preference frustration entailed by (involuntary) extinction may overall imply even worse welfare levels compared to non-extinction (depending on how one construes these views, as well as how bad the non-extinction scenario would be).

This also highlights the importance of specifying how the extinction in question occurs. After all, if every individual in a given world were to decide not to procreate, and thus voluntarily bring about extinction, then this need not involve potential bads such as murder, rights violations, or severe preference frustrations. In contrast, scenarios involving violent destruction *would* involve such potential bads. The former extinction scenario would likely be regarded as far less bad by most people, and perhaps even all-things-considered good if it prevents vast amounts of unbearable suffering.

Empirical reasons

Beyond the axiological reasons listed above, there are also empirical reasons why LTNU need not imply the desirability of human extinction. For instance, one may hold the empirical belief that humanity is likely to soon abolish the biology of suffering in all sentient life, in which case human extinction might be very bad by the lights of LTNU.

And even if one is less optimistic about such an abolitionist prospect, one may still believe that continued human existence would on the whole tend to reduce extreme suffering, such as by reducing the number of wild animals, or — more speculatively — by causing less suffering than would an alien civilization in humanity's stead (Knutsson, 2021, sec. 4). (Note that I am not claiming that humanity is most likely to reduce suffering overall, but merely that it is unclear as an empirical matter what humanity's net effect on suffering will be, and hence it is unclear whether LTNU alone would support human extinction.)

25 The name "LTNU" — an abbreviation for "lexical threshold negative utilitarianism" — is thus quite unfortunate, since this axiological principle does not imply utilitarianism of any kind.

Of course, the points above do not apply to the extinction of all sentient life. But even here, there are important qualifications to be made. First, the extinction of all sentient life would not be optimal (by the lights on LTNU) if sentient life re-emerges later in worse ways (cf. Knutsson, 2021, sec. 4). Hence, we must be careful to distinguish a merely temporary extinction of sentient life from permanent extinction. After all, the potential impossibility of permanent extinction could imply the all-things-considered undesirability of temporary extinction. For example, one might hold an optimistic view of the future of humanity that implies that scenarios involving temporary extinction followed by a re-emergence of sentience (which might involve hundreds of millions of years of wild-animal suffering on many planets) would be worse than scenarios without temporary extinction, e.g. if one believes that humanity will soon abolish suffering for good throughout the accessible universe.

Moreover, even if we are talking about permanent extinction, one may still accept LTNU without necessarily believing that the extinction of all sentient life is on the whole desirable. In particular, if one combines LTNU with one or more of the axiological views outlined in the previous section — e.g. axiological views that assign significant disvalue to killings or to frustrated preferences — and if one holds an optimistic view of humanity's ability to prevent suffering and other bads in the future, then one could maintain that non-extinction would overall be preferable to permanent extinction.

Would extinction be bad?

The intuition that extinction would be bad may be questioned in a variety of ways, and it seems worth separating that intuition from related yet distinct intuitions.

For instance, the badness of extinction might be conflated with the potential badness of death and murder, yet extinction need not involve murder, and it need not involve more death than non-extinction (in fact, earlier extinction would likely involve less death overall; the same point is made in Bergström, 2022, sec. VIII).

Additionally, there may be good reasons to doubt our intuitions about the badness of extinction (through voluntary non-procreation, say). One reason is that we might for evolutionary reasons have what Thomas Metzinger calls an existence bias, which strongly biases our intuitions to favor continued existence at almost any price.

Likewise, one may reasonably question our intuitions about the badness of an empty world. That is, we may intuitively feel like an empty world would be a horrifying prospect, but does such an intuition stand up to scrutiny? Some have argued that it does not, or at least that it might not (see

e.g. Benatar, 2018; Crisp, 2021). In particular, some have argued that there is nothing bad or even suboptimal about an empty world (Ajantaival, 2022).

Miscellaneous comments

Suffering in Omelas: Far from the worst suffering

In his discussion of LTNU, Mogensen uses the miserable child in *The Ones Who Walk Away from Omelas* as an example of someone who endures intense suffering (i.e. a child who is locked up in isolation and who barely gets enough food to survive). Yet focusing on this example is arguably unfair to LTNU. That is, compared to how bad suffering can get, this is an exceedingly mild example (even as it is horrific in absolute terms). It would be more fair to focus on more extreme examples of suffering when discussing the plausibility of LTNU.

Indeed, a proponent of LTNU could place the misery threshold at a level of suffering that is much more severe than the misery experienced by the child in Omelas, in which case discussions of Omelas-level suffering have limited bearing on the plausibility of LTNU.

A thought experiment involving destruction

Mogensen writes the following in his discussion of Omelas-level suffering (p. 11):

> If the child is imagined as inhabiting a faroff country, and if the boundless and generous contentment of Omelas is imagined as independent of her suffering, except in that it would have to be destroyed in the process of working to spare her from her misery, then I don't find I have the same reaction as before [i.e. that the child should be spared from suffering].

Yet this talk about destruction is potentially misleading (when raised as an argument against LTNU). As noted above, one may endorse other axiological principles that render such destruction bad overall, without thinking that happy lives can outweigh extremely miserable lives. So this case does not clearly pit the value of happy lives against the disvalue of miserable lives, and hence it does not clearly serve to question the plausibility of LTNU.

Additionally, one can criticize this framing for potentially appealing to a status quo bias, which could be avoided by instead considering the issue from a neutral starting point where no beings exist, and where we are contemplating whether it would be better to create these beings or not. (I have made a similar critique of a similar framing here and here.)

“Disturbing implications”

Finally, I find it a bit unfortunate that the paper repeatedly refers to the “disturbing implications” of LTNU. First of all because it is not specified what these “disturbing implications” are exactly; and the one implication that *is* specified (that extinction would be desirable) is not one that strictly follows from LTNU, nor one that is shown to be disturbing (e.g. such a claim may plausibly be questioned in the case of voluntary extinction).

Moreover, I suspect that this framing can potentially distort the paper’s examination, since a key question of the paper is whether LTNU is all-things-considered more or less plausible (or “disturbing”) than its rejection. Yet the problematic implications of the *rejection* of LTNU are not characterized as disturbing, even though one could argue that such implications are even more disturbing.

References

Ajantaival, T. (2022). Peacefulness, nonviolence, and experientialist minimalism. Ungated

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

Benatar, D. (2018). Is Extinction Bad? Ungated

Bergström, L. (1978/2022). The consequences of pessimism. Translated by Simon Knutsson. Ungated

Crisp, R. (2021). Would extinction be so bad? Ungated

Knutsson, S. (2021). The world destruction argument. *Inquiry,* 64(10), pp. 1004-1023. Ungated

Metzinger, T. (2017). Benevolent Artificial Anti-Natalism (BAAN). Ungated

Mogensen, A. (2022). The weight of suffering. Ungated

Critique of MacAskill's "Is It Good to Make Happy People?"

In *What We Owe the Future*, William MacAskill delves into population ethics in a chapter titled "Is It Good to Make Happy People?" (Chapter 8). As he writes at the outset of the chapter, our views on population ethics matter greatly for our priorities, and hence it is important that we reflect on the key questions of population ethics. Yet it seems to me that the book skips over some of the most fundamental and most action-guiding of these questions. In particular, the book does not broach questions concerning whether any purported goods can outweigh extreme suffering — and, more generally, whether happy lives can outweigh miserable lives — even as these questions are all-important for our priorities.

The Asymmetry in population ethics

A prominent position that gets a very short treatment in the book is the Asymmetry in population ethics (roughly: bringing a miserable life into the world has negative value while bringing a happy life into the world does not have positive value — except potentially through its instrumental effects and positive roles).

The following is, as far as I can tell, the main argument that MacAskill makes against the Asymmetry (p. 172):

> If we think it's bad to bring into existence a life of suffering, why should we not think that it's good to bring into existence a flourishing life? I think any argument for the first claim would also be a good argument for the second.

This claim about "any argument" seems unduly strong and general. Specifically, there are many arguments that support the intrinsic badness of bringing a miserable life into existence that do not support any intrinsic goodness of bringing a flourishing life into existence. Indeed, many arguments support the former while positively denying the latter.

One such argument is that the presence of suffering is bad and morally worth preventing while the absence of pleasure is not bad and not a problem, and hence not morally worth "fixing" in a symmetric way (provided that no existing beings are deprived of that pleasure).[26]

A related class of arguments in favor of an asymmetry in population ethics is based on theories of wellbeing that understand happiness as the absence of cravings, preference frustrations, or other

26 Further arguments against a moral symmetry between happiness and suffering are found in Mayerfeld, 1999, ch. 6; Vinding, 2020, sec. 1.4 & ch. 3.

bothersome features. According to such views, states of untroubled contentment are just as good — and perhaps even better than — states of intense pleasure.[27]

These views of wellbeing likewise support the badness of creating miserable lives, yet they do not support any supposed goodness of creating happy lives. On these views, intrinsically positive lives do not exist, although *relationally* positive lives do.

Another point that MacAskill raises against the Asymmetry is an example of happy children who already exist, about which he writes (p. 172):

> if I imagine this happiness continuing into their futures—if I imagine they each live a rewarding life, full of love and accomplishment—and ask myself, "Is the world at least a little better because of their existence, even ignoring their effects on others?" it becomes quite intuitive to me that the answer is yes.

However, there is a potential ambiguity in this example. The term "existence" may here be understood to either mean "de novo existence" or "continued existence", and interpreting it as the latter is made more tempting by the fact that 1) we are talking about already existing beings, and 2) the example mentions their happiness "continuing into their futures".[28]

This is relevant because many proponents of the Asymmetry argue that there is an important distinction between the potential value of continued existence (or the badness of discontinued existence) versus the potential value of bringing a new life into existence.

Thus, many views that support the Asymmetry will agree that the happiness of these children "continuing into their futures" makes the world better, or less bad, than it otherwise would be (compared to a world in which their existing interests and preferences are thwarted). But these views still imply that the de novo *creation* (and eventual satisfaction) of these interests and preferences does not make the world better than it otherwise would be, had they not been created in the first place. (Some sources that discuss or defend these views include Singer, 1980; Benatar, 1997; 2006; Fehige, 1998; Anonymous, 2015; St. Jules, 2019; Frick, 2020.)

A proponent of the Asymmetry may therefore argue that the example above carries little force against the Asymmetry, as opposed to merely supporting the badness of preference frustrations and other deprivations for already existing beings.[29]

27 On some views of wellbeing, especially those associated with Epicurus, the complete absence of any bothersome or unpleasant features is regarded as the highest pleasure, Sherman, 2017, p. 103; Tsouna, 2020, p. 175. Psychologist William James also expressed this view, James, 1901.

28 I am not saying that the "continued existence" interpretation is necessarily the most obvious one to make, but merely that there is significant ambiguity here that is likely to confuse many readers as to what is being claimed.

29 Moreover, a proponent of minimalist axiologies may argue that the assumption of "ignoring all effects on others" is so radical that our intuitions are unlikely to fully ignore all such instrumental effects even when we try to, and hence we may be inclined to confuse 1) the relational value of creating a life with 2) the (purported) intrinsic

Questions about outweighing

Even if one thinks that it is good to create more happiness and new happy lives all else equal, this still leaves open the question as to whether happiness and happy lives can outweigh suffering and miserable lives, let alone extreme suffering and extremely bad lives. After all, one may think that more happiness is good while still maintaining that happiness cannot outweigh intense suffering or very bad lives — or even that it cannot outweigh the worst elements found in relatively good lives. In other words, one may hold that the value of happiness and the disvalue of suffering are in some sense orthogonal (cf. Wolf, 1996; 1997; 2004).

As mentioned above, these questions regarding tradeoffs and outweighing are not raised in MacAskill's discussion of population ethics, despite their supreme practical significance.[30] One way to appreciate this practical significance is by considering a future in which a relatively small — yet in absolute terms vast — minority of beings live lives of extreme and unrelenting suffering. This scenario raises what I have elsewhere (sec. 14.3) called the "Astronomical Atrocity Problem": can the extreme and incessant suffering of, say, trillions of beings be outweighed by any amount of purported goods? (See also this short excerpt from Vinding, 2018.)

After all, an extremely large future civilization would contain such (in absolute terms) vast amounts of extreme suffering in expectation, which renders this problem frightfully relevant for our priorities.

---

positive value contained within that life in isolation — especially since the example involves a life that is "full of love and accomplishment", which might intuitively evoke many effects on others, despite the instruction to ignore such effects.

30 MacAskill's colleague Andreas Mogensen has commendably raised such questions about outweighing in his essay "The weight of suffering", which I have discussed here.
Chapter 9 in MacAskill's book does review some psychological studies on intrapersonal tradeoffs and preferences (see e.g. p. 198), but these self-reported intrapersonal tradeoffs do not necessarily say much about which *inter*personal tradeoffs we should consider plausible or valid. Nor do these intrapersonal tradeoffs generally appear to include cases of extreme suffering, let alone an entire lifetime of torment (as experienced, for instance, by many of the non-human animals whom MacAskill describes in Chapter 9). Hence, that people are willing to make intrapersonal tradeoffs between everyday experiences that are more or less enjoyable says little about whether some people's enjoyment can morally outweigh the intense suffering or extremely bad lives endured by others. (In terms of people's self-reported willingness to experience extreme suffering in order to gain happiness, a small survey (n=99) found that around 45 percent of respondents would not experience even a single minute of extreme suffering for any amount of happiness; and that was just the intrapersonal case — such suffering-for-happiness trades are usually considered less plausible and less permissible in the interpersonal case, cf. Mayerfeld, 1999, pp. 131-133; Vinding, 2020, sec. 3.2.)
Individual ratings of life satisfaction are similarly limited in terms of what they say about *intra*personal tradeoffs. Indeed, even a high rating of momentary life satisfaction does not imply that the evaluator's life itself has overall been worth living, even by the evaluator's own standards. After all, one may report a very high quality of life yet still think that the good part of one's life cannot outweigh one's past suffering. It is thus rather limited what we can conclude about the value of individual lives, much less the world as a whole, based on people's momentary ratings of life satisfaction.
Finally, MacAskill also mentions various improvements that have occurred in recent centuries as a reason to be optimistic about the future of humanity in moral and evaluative terms. Yet it is unclear whether any of the improvements he mentions involve genuine positive goods, as opposed to representing a reduction of bads, e.g. child mortality, poverty, totalitarian rule, and human slavery (cf. Vinding, 2020, sec. 8.6).

MacAskill's chapter does discuss the Repugnant Conclusion at some length, yet the Repugnant Conclusion does not explicitly involve any tradeoffs between happiness and suffering,[31] and hence it has limited relevance compared to, for example, the Very Repugnant Conclusion (roughly: that arbitrarily many hellish lives can be "compensated for" by a sufficiently vast number of lives that are "barely worth living").[32]

Indeed, the Very Repugnant Conclusion and similar such "offsetting conclusions" would seem more relevant to discuss both because 1) they do explicitly involve tradeoffs between happiness and suffering, or between happy lives and miserable lives, and because 2) MacAskill himself has stated that he considers the Very Repugnant Conclusion to be the strongest objection against his favored view, and stronger objections generally seem more worth discussing than do weaker ones.[33]

Popular support for significant asymmetries in population ethics

MacAskill briefly summarizes a study that surveyed people's views on population ethics. Among other things, he writes the following about the findings of the study (p. 173):

> these judgments [about the respective value of creating happy lives and unhappy lives] were symmetrical: the experimental subjects were just as positive about the idea of bringing into existence a new happy person as they were negative about the idea of bringing into existence a new unhappy person.

While this summary seems accurate if we only focus on people's responses to one specific question in the survey (cf. Caviola et al., 2022, p. 9), there are nevertheless many findings in the study that suggest that people generally do endorse significant asymmetries in population ethics.

Specifically, the study found that people on average believed that considerably more happiness than suffering is needed to render a population or an individual life worthwhile, even when the happiness and suffering were said to be equally intense (Caviola et al., 2022, p. 8). The study likewise found that participants on average believed that the ratio of happy to unhappy people in a population must be at least 3-to-1 for its existence to be better than its non-existence (Caviola et al., 2022, p. 5).

---

31 Some formulations of the Repugnant Conclusion do involve tradeoffs between happiness and suffering, and the conclusion indeed appears much more repugnant in those versions of the thought experiment.

32 One might object that the Very Repugnant Conclusion has limited practical significance because it represents an unlikely scenario. But the same could be said about the Repugnant Conclusion (especially in its suffering-free variant). I do not claim that the Very Repugnant Conclusion is the most realistic case to consider. When I claim that it is more practically relevant than the Repugnant Conclusion, it is simply because it does explicitly involve tradeoffs between happiness and (extreme) suffering, which we know will also be true of our decisions pertaining to the future.

33 For what it's worth, I think an even stronger counterexample is "Creating hell to please the blissful", in which an arbitrarily large number of maximally bad lives are "compensated for" by bringing a sufficiently vast base population from near-maximum welfare to maximum welfare.

Another relevant finding is that people generally have a significantly stronger preference for smaller over larger unhappy populations than they do for larger over smaller happy populations, and the magnitude of this difference becomes greater as the populations under consideration become larger (Caviola et al., 2022, pp. 12-13).

In other words, people's preference for smaller unhappy populations becomes stronger as population size increases, whereas the preference for larger happy populations becomes less strong as population size increases, in effect creating a strong asymmetry in cases involving large populations (e.g. above one billion individuals). This finding seems particularly relevant when discussing laypeople's views of population ethics in a context that is primarily concerned with the value of potentially vast future populations.[34]

Moreover, a pilot study conducted by the same researchers suggested that the framing of the question plays a major role for people's intuitions (Caviola et al., 2022, "Supplementary Materials"). In particular, the pilot study (n=172) asked people the following question:

> Suppose you could push a button that created a new world with X people who are generally happy and 10 people who generally suffer. How high would X have to be for you to push the button?

When the question was framed in these terms, i.e. in terms of creating a new world, people's intuitions were radically more asymmetric, as the median ratio then jumped to 100-to-1 happy to unhappy people, which is a rather pronounced asymmetry.[35]

In sum, it seems that the study that MacAskill cites above, when taken as a whole, mostly finds that people on average do endorse significant asymmetries in population ethics. I think this documented level of support for asymmetries would have been worth mentioning.

(Other surveys that suggest that people on average affirm a considerable asymmetry in the value of happiness vs. suffering and good vs. bad lives include the Future of Life Institute's Superintelligence survey (n=14,866) and Tomasik, 2015 (n=99).)

34 Some philosophers have explored, and to some degree supported, similar views. For example, Derek Parfit wrote (Parfit, 1984, p. 406): "When we consider the badness of suffering, we should claim that this badness has no upper limit. It is always bad if an extra person has to endure extreme agony. And this is always just as bad, however many others have similar lives. The badness of extra suffering never declines." In contrast, Parfit seemed to consider it more plausible that the addition of happiness adds diminishing marginal value to the world, even though he ultimately rejected that view because he thought it had implausible implications, Parfit, 1984, pp. 406-412. See also Hurka, 1983; Gloor, 2016, sec. IV; Vinding, 2020, sec. 6.2. Such views imply that it is of chief importance to avoid very bad outcomes on a very large scale, whereas it is relatively less important to create a very large utopia.

35 This framing effect could be taken to suggest that people often fail to fully respect the radical "other things being equal" assumption when considering the addition of lives in *our* world. That is, people might not truly have thought about the value of new lives in total isolation when those lives were to be added to the world we inhabit, whereas they might have come closer to that ideal when they considered the question in the context of creating a new, wholly self-contained world. (Other potential explanations of these differences are reviewed in Contestabile, 2022, sec. 4; Caviola et al., 2022, "Supplementary Materials", pp. 7-8.)

The discussion of moral uncertainty excludes asymmetric views

Toward the end of the chapter, MacAskill briefly turns to moral uncertainty, and he ends his discussion of the subject on the following note (p. 187):

> My colleagues Toby Ord and Hilary Greaves have found that this approach to reasoning under moral uncertainty can be extended to a range of theories of population ethics, including those that try to capture the intuition of neutrality. When you are uncertain about all of these theories, you still end up with a low but positive critical level [of wellbeing above which it is a net benefit for a new being to be created for their own sake].

Yet the analysis in question appears to wholly ignore asymmetric views in population ethics. If one gives significant weight to asymmetric views — not to mention stronger minimalist views in population ethics — the conclusion of the moral uncertainty framework is likely to change substantially, perhaps so much so that the creation of new lives is generally not a benefit for the created beings themselves (although it could still be a net benefit for others and for the world as a whole, given the positive roles of those new lives).

Similarly, even if the creation of unusually happy lives would be regarded as a benefit from a moral uncertainty perspective that gives considerable weight to asymmetric views, this benefit may still not be sufficient to counterbalance extremely bad lives,[36] which are granted unique weight by many plausible axiological and moral views (cf. Mayerfeld, 1999, pp. 114-116; Vinding, 2020, ch. 6).[37]

References

Ajantaival, T. (2021/2022). Minimalist axiologies. Ungated

Anonymous. (2015). Negative Utilitarianism FAQ. Ungated

Benatar, D. (1997). Why It Is Better Never to Come into Existence. *American Philosophical Quarterly*, 34(3), pp. 345-355. Ungated

Benatar, D. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

Caviola, L. et al. (2022). Population ethical intuitions. *Cognition*, 218, 104941. Ungated; Supplementary Materials

36 Or at least not sufficient to counterbalance the substantial number of very bad lives that the future contains in expectation, cf. the Astronomical Atrocity Problem mentioned above.

37 Further discussion of moral uncertainty from a perspective that takes asymmetric views into account is found in DiGiovanni, 2021.

Contestabile, B. (2022). Is There a Prevalence of Suffering? An Empirical Study on the Human Condition. Ungated

DiGiovanni, A. (2021). A longtermist critique of "The expected value of extinction risk reduction is positive". Ungated

Fehige, C. (1998). A pareto principle for possible people. In Fehige, C. & Wessels U. (eds.), *Preferences*. Walter de Gruyter. Ungated

Frick, J. (2020). Conditional Reasons and the Procreation Asymmetry. *Philosophical Perspectives*, 34(1), pp. 53-87. Ungated

Future of Life Institute. (2017). Superintelligence survey. Ungated

Gloor, L. (2016). The Case for Suffering-Focused Ethics. Ungated

Gloor, L. (2017). Tranquilism. Ungated

Hurka, T. (1983). Value and Population Size. *Ethics*, 93, pp. 496-507.

James, W. (1901). Letter on happiness to Miss Frances R. Morse. In *Letters of William James*, *Vol. 2* (1920). Atlantic Monthly Press.

Knutsson, S. (2019). Epicurean ideas about pleasure, pain, good and bad. Ungated

MacAskill, W. (2022). *What We Owe The Future*. Basic Books.

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Sherman, T. (2017). Epicureanism: An Ancient Guide to Modern Wellbeing. MPhil dissertation, University of Exeter. Ungated

Singer, P. (1980). Right to Life? Ungated

St. Jules, M. (2019). Defending the Procreation Asymmetry with Conditional Interests. Ungated

Tomasik, B. (2015). A Small Mechanical Turk Survey on Ethics and Animal Welfare. Ungated

Tsouna, V. (2020). Hedonism. In Mitsis, P. (ed.), *Oxford Handbook of Epicurus and Epicureanism*. Oxford University Press.

Vinding, M. (2018). *Effective Altruism: How Can We Best Help Others?* Ratio Ethica. Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Wolf, C. (1996). Social Choice and Normative Population Theory: A Person Affecting Solution to Parfit's Mere Addition Paradox. *Philosophical Studies*, 81, pp. 263-282.

Wolf, C. (1997). Person-Affecting Utilitarianism and Population Policy. In Heller, J. & Fotion, N. (eds.), *Contingent Future Persons*. Kluwer Academic Publishers. Ungated

Wolf, C. (2004). O Repugnance, Where Is Thy Sting? In Tännsjö, T. & Ryberg, J. (eds.), *The Repugnant Conclusion*. Kluwer Academic Publishers. Ungated

Reply to the "evolutionary asymmetry objection" against suffering-focused ethics

An objection that is sometimes raised against suffering-focused ethics is that our intuitions about the relative value of suffering and happiness are skewed toward the negative for evolutionary reasons, and hence we cannot trust our intuition that says that the reduction of suffering is more valuable and more morally important than the creation of happiness. My aim in this post is to reply to this objection.

Stating the "evolutionary asymmetry objection" in brief

The argument I will respond to goes roughly as follows: For biologically evolved creatures such as humans, the reproductive costs of losses (e.g. deadly injury) are typically greater than the reproductive gains of successes (e.g. sex). This asymmetry is plausibly reflected in our experiences such that we tend to feel and value suffering (as an intrinsic negative) more strongly than we feel and value pleasure (as an intrinsic positive). Yet we should not expect such an asymmetry to be found at the level of *possible* states of suffering and pleasure. Instead, we should expect the best possible pleasure and the worst possible suffering to be equally intense, and we should therefore expect there to be an axiological and moral symmetry between them. Or at least we should expect our better informed selves to endorse such an axiological and moral symmetry (e.g. if we were fully acquainted with the utmost extremes of pleasure and suffering).

Furthermore, the experiences of future beings need not be subject to the evolutionarily contingent asymmetries found in the experiences of biological beings, and hence we should be far more optimistic about the value of the best future pleasures, and about the total amount of future pleasure, than we are intuitively inclined to be.[38]

Questioning the first three premises of the "evolutionary asymmetry objection"

A straightforward way to reply to the objection outlined above is by questioning some of its individual premises.

38 An argument along these lines is found in Shulman, 2012, though note that there are important differences between Shulman's argument and the argument that I have outlined here. For example, Shulman is primarily talking about energy-efficiency rather than intensity or "best versus worst", as highlighted in Knutsson, 2017.

Are gains and losses asymmetric in terms of human reproductive fitness?

First, one could question whether gains and losses are in fact asymmetric in evolutionary terms for humans. In particular, one could argue that the difference between "being average" and "winning big" (e.g. by gaining great power and having a disproportionate number of offspring) is larger than the difference between "being average" and "losing big" (e.g. being dead) in terms of human reproductive fitness.

However, in my view, this is not a wholly convincing argument, since having a disproportionate number of offspring would tend to require a continual investment (even if only in terms of sexual investment), whereas dying would be a one-time event that would always be highly costly. And this would arguably be even more true for someone who was very powerful and biologically prolific — the reproductive cost of a single "death event" would still tend to be much greater than the reproductive gain of a single "sexual event" (or a similar "success event"), as the former would preclude many instances of the latter.

So it seems that steering clear of the worst outcomes moment-to-moment likely was more important than was attaining the best, in terms of the impact that individual events had on the reproductive fitness of our ancestors, even if there are counterarguments that limit the expected magnitude of this asymmetry.

Are reproductive gains tracked by pleasures as "intrinsic positives"?

Another premise that one could question is whether the above-mentioned asymmetry between reproductive gains and losses is necessarily reflected or "tracked" by an experiential asymmetry in suffering versus pleasure. Specifically, is it plausible to claim that humans are motivated to avoid reproductive losses by pains as "intrinsic negatives" while we are motivated to achieve reproductive gains by pleasures as "intrinsic positives"?

An alternative model could be that we are motivated by various desires (or felt cravings, or needs, etc.) that animate both the avoidance of reproductive losses *and* the pursuit of reproductive gains (cf. drive reduction theory). On this model, there could still be an experiential asymmetry in the sense that the desire to avoid death or severe bodily harm may tend to feel stronger than does the desire to find, say, a high-fitness partner. (One can, of course, dispute whether that is actually true of human desires.) Yet such an experiential asymmetry need not imply any motivation to achieve intrinsically positive states, as opposed to a motivation to attain a more relieved or less needful state.

Should we expect the best possible pleasure and the worst possible suffering to be equally intense?

The desire-based model of motivation outlined above may also question whether the best and the worst possible states are equally intense. That is, if one holds that we are chiefly motivated to satisfy (more or less bothersome) needs and desires, rather than to attain positive pleasures, one may likewise argue that there are no pleasures "on the other side" of a perfectly content state. Similarly, some Epicurean views of pleasure hold that the complete absence of pain is the "limit of pleasure" (Sherman, 2017, p. 103). On such views, it makes little sense to say that the best possible pleasure and the worst possible suffering are "equally intense".

In general, regardless of whether one endorses any of the views of motivation and pleasure outlined above, one can reasonably question whether the respective intensities of pleasure and suffering are commensurable, i.e. whether they are measurable by the same standard (cf. Knutsson, 2016).

Yet even if we grant that suffering and pleasure do have commensurable intensities, and if we further grant that the best possible pleasures and the worst possible suffering are equally intense, it still does not follow that there is an axiological or moral symmetry between them. I think we have strong reasons to reject such symmetries, as I will try to explain in the next section.

What justifies a moral symmetry?

My main reply to the "evolutionary asymmetry objection" is to ask what justifies the acceptance of any kind of moral symmetry or moral outweighing between happiness and suffering to begin with, even at the level of currently accessible states that are claimed to be "similarly intense".

I would argue that there is nothing about experiential states of happiness (e.g. excitement, gratitude, amusement, etc.) that render them a truly positive counterpart to suffering, neither in phenomenological nor axiological terms (Vinding, 2022).[39] Instead, it seems plausible that there are no experiential states above wholly undisturbed mental states (Sherman, 2017; Gloor, 2017; Knutsson, 2022). (For an elaborate defense of minimalist axiologies that reject any supposed symmetry between happiness and suffering, see Ajantaival, 2021/2022.)

In moral terms, there is the argument that the reduction of suffering is morally urgent, whereas there is arguably no moral urgency, let alone a similar moral urgency, in "correcting" a neutral absence of happiness (Vinding, 2020, sec. 1.4). Unlike the presence of suffering, the absence of happiness does

39 Note that one could think that suffering and pleasure have commensurable intensities without thinking that pleasure is a positive counterpart to suffering. One may think that pleasure and suffering can be comparably intense in orthogonal (rather than anti-directional) experiential dimensions, as it were.

not seem morally problematic, which means that failing to create happiness (that nobody needs) is akin to a victimless "crime".

More generally, there are various arguments for the moral principle that it is wrong to create happiness at the price of suffering, and that happiness can never morally outweigh suffering (Vinding, 2020, ch. 3). These arguments include thought experiments in which the supposed moral symmetry between happiness and suffering would imply that it is morally right to torture some beings for the pleasure of others.

In my experience, proponents of a moral symmetry between happiness and suffering rarely address this implication, despite it being perhaps the most problematic implication of such a moral symmetry. And when the objection does get discussed, the main reply is often that one should not commit such torture in practice, or that the thought experiment is unrealistic (see e.g. Lazari-Radek & Singer, 2017, ch. 4). Yet this reply obscures the fact that a general moral symmetry indeed would entail this implication in theory, and even if we were to grant that the scenario is unrealistic, it still appears to be a highly implausible theoretical implication.

More than that, the reply obscures the fact that tradeoffs like these *are* realistic in terms of whether we prioritize preventing extreme suffering or whether we prioritize creating new happy beings, and thereby allow more extreme suffering to occur by omission. (Whether we ourselves impose or merely allow the occurrence of the extreme suffering in question does not ultimately matter according to strict consequentialist versions of the moral symmetry, e.g. the version defended in Lazari-Radek & Singer, 2017.) And note that a vast future indeed would contain large amounts of extreme suffering in expectation, even if we avoid the worst risks of astronomical future suffering, or s-risks (Vinding, 2020, sec. 14.3).

In sum, the objection outlined above seems to assume that suffering can be morally outweighed by pleasure, but it does not provide any justification for this premise. Yet that, to my mind, is the key premise that needs to be defended, especially by addressing its most problematic implications.[40]

Should we be humble given our narrow and potentially skewed range of experience?

A proponent of the "evolutionary asymmetry objection" might argue that our narrow and potentially skewed range of experience should make us humble and uncertain in our inferences regarding the moral (a)symmetry between happiness and suffering, and they might further argue that this

40 A common argument in defense of this premise is that most people are willing to accept tradeoffs between suffering and pleasure in everyday life. I have replied to that argument here. See also DiGiovanni, 2019; Vinding, 2020, sec. 2.4.

uncertainty should push us toward the symmetric view. Yet this objection seems problematic for a couple of reasons.

First, it seems to overlook that we do have significant data to draw on from our current range of experience, and one may argue that we have good grounds for skepticism about the relevance of the "evolutionary asymmetry objection" based on this data. As one author put it (Anonymous, 2015):

> Our current pleasure/pain-intensity ranges may be negatively skewed for evolutionary reasons, but this doesn't provide a strong argument for people who are able to experience the most intense current pleasures and milder current pains and are convinced that there's an asymmetry.

Anthony DiGiovanni makes a similar point regarding experiences of which he has first-hand knowledge (DiGiovanni, 2021b):

> The problem is that when I try to compare apples to apples in terms of intensity of experience, I still don't see how happiness (or complexity, knowledge, beauty, whatever) could compete with suffering for moral priority. I have the same intuition when I consider cases where the intensity of the purportedly positive experience is quite clearly higher.

I myself draw the same introspective conclusion (Vinding, 2022). And so does meditator Roger Thisdell, who reports having experienced many states of unusually intense bliss, e.g. "bliss trips, jhanas, 5-MeO, MDMA, staring into the eyes of a lover without insecurities, laughing fits", yet he still argues that "pleasure as a positive ... does not exist" (Gómez-Emilsson, 2021). Indeed, many traditions that have developed practices of careful introspection appear to have converged on similar asymmetric conclusions regarding the nature of pleasure and phenomenal value (see e.g. Contestabile, 2014; Breyer, 2015; Vinding, 2020, sec. 8.14).

Second, the objection above overlooks that the point regarding humility and uncertainty cuts both ways. That is, just as we are generally far from experiencing the greatest pleasure, we are likewise far from having experienced the worst suffering, and it is not clear whether most of us have been "closer" to the worst suffering than to the greatest pleasure. It is therefore questionable whether uncertainty in light of our narrow range of experience should ultimately push us further toward a symmetric view.

Indeed, if an Epicurean view of pleasure is correct or most plausible, our experiences will tend to be much further from the most intense suffering than from the most intense pleasure (i.e. the complete absence of all disturbances, Sherman, 2017, p. 103). The same would be true on the broadly similar

views of pleasure defended by Arthur Schopenhauer, Eduard von Hartmann, and others. Thus, if one gives some weight to these views, and if one were to update one's beliefs about the extremes of pleasure and suffering based on the various views that have been defended regarding the nature of pleasure and suffering (views that, crudely speaking, tend to range from "symmetric" to "strongly negatively asymmetric"), one could argue that we should overall expect our notional "distance" from the most intense suffering to be greater than our distance from the most intense pleasure.

After all, just as one might argue that we should not a priori privilege asymmetric views of the nature of pleasure and suffering, one might similarly argue that we should not privilege symmetric views either when reasoning about these matters from an uncertain perspective, least of all when many alternative views have been defended.[41]

Finally, the humility objection still does not address how the absence of any hypothetical state of bliss could be morally problematic or morally urgent to address if no existing being feels a need for it, let alone how its absence could be as morally problematic or urgent as the presence of a state of suffering. The non-problematic nature of the absence of a given state seems independent of the nature or intensity of that state (Vinding, 2020, sec. 1.4).

Counterbiases

The "evolutionary asymmetry objection" essentially claims that our assessments concerning the relative moral value of happiness and suffering are biased, and that we would endorse a moral symmetry if only we controlled for this bias, or at least we would get closer to endorsing a moral symmetry.[42] Yet if we grant that such "evolutionary biasing" is possible, one could plausibly argue that we have various biases in the other direction as well, and that we are in fact overall strongly biased toward endorsing a moral symmetry — not necessarily in terms of assuming equal intensities, but in terms of assuming that pleasure can morally outweigh suffering in the first place.

First, as Thomas Metzinger has argued, it seems plausible that we have a strong "existence bias" that pushes us to favor existence at virtually any price (Metzinger, 2017):

> I claim that our deepest cognitive bias is "existence bias", which means that we will simply do almost anything to prolong our own existence. For us, sustaining one's existence is the default goal in almost every case of uncertainty, even if it may violate

41 Note that this point about uncertainty regarding the nature of pleasure is different from moral uncertainty. Indeed, one could argue that each of these forms of uncertainty independently push toward asymmetric views, given the range of views that have been defended at these respective levels. Specifically, in addition to arguments regarding asymmetries in the nature of pleasure and pain, there are independent moral arguments that further support giving greater priority to the reduction of suffering (Vinding, 2020, sec. 1.5, sec. 6.2; MacAskill et al., 2020, p. 185).

42 For a discussion of what it might mean to be biased in our assessments of moral values, see Vinding, 2020, sec. 7.1. One could, for instance, define a moral bias as "a factor that influences our moral reflections in a way that we would not endorse if we were more fully informed".

> rationality constraints, simply because it is a biological imperative that has been burned into our nervous systems over millennia.

This is related to Robert Daoust's claim that humans tend to have strong survivalist intuitions and values that frequently override welfarist concerns (see also Vinding, 2020, sec. 7.11).

In the context of our appraisals of the supposed moral symmetry between happiness and suffering, one could argue that our existence bias and survivalist intuitions plausibly bias us toward endorsing a moral symmetry between happiness and suffering. That is, a moral symmetry conforms much better with our existence bias and our survivalist intuitions than does a moral asymmetry that favors the reduction of suffering, and hence these evolved intuitions plausibly push us strongly toward accepting the desired symmetry. And if Metzinger and Daoust are right about the strength of our existence bias, then this putative bias in favor of embracing symmetry may well be stronger than the purported "evolutionary asymmetry bias" against symmetry.

Another potentially biasing factor is that we are used to thinking in terms of positive and negative numbers in just about every sphere of life. Consequently, we might be inclined to reflexively assign positive and negative numbers to different experiences when trying to represent their value, even if this conceptual move may not be the most plausible way to represent value on reflection.

Likewise, one can speculate that we are inclined to project positive value onto those things that tend to reduce pain and frustration — e.g. things that have positive roles in the alleviation of suffering — while overlooking that this seemingly intrinsic positive value ultimately has its basis in the reduction of suffering and unmet needs. After all, it would be quite demanding if we were to unpack this positive value in terms of its relational roles, and hence the "intrinsic positive value" framing may be more practically efficient and adaptive, even if the notion of intrinsic positive value might not be axiologically plausible on a deeper analysis. (Arguments along these lines are pursued in Ajantaival, 2021/2022.)

All in all, if we grant that we can be biased in our reflections on values, it is far from clear that we are more biased to reject a moral symmetry than we are biased to endorse it, including — and perhaps especially — when we consider potential biases that relate to our evolutionary origin. Indeed, it seems that one could reasonably argue that we are overall more biased to endorse rather than to reject a moral symmetry between happiness and suffering. (For a review of other potential biases, see Vinding, 2020, ch. 7.)

Disproportionally intense suffering may persist in the future

Another relevant consideration is the empirical point that states of suffering might continue to be more intense than are states of pleasure (if we grant, for the sake of argument, that states of suffering and pleasure have commensurable intensities). After all, overridingly intense suffering appears to have been adaptive for biological beings in the past, which suggests that this pattern could also be adaptive for beings in future scenarios that involve similar processes of competition and evolution, even if those scenarios involve advanced technologies.

In other words, if the future will be highly competitive (as seems fairly likely), and if disproportional intensities of suffering confer adaptive advantages in competitive environments (as seems to have been the case historically), then future beings seem likely to also be motivated by disproportional intensities of suffering — perhaps even maximal intensities of suffering among beings designed to be maximally motivated.

To be clear, I am not saying that the future is guaranteed to resemble the past, but merely that the range of experiences that motivate sentient beings today does represent some evidence regarding the range of experiences that we should expect to be prevalent in sentient beings in the future. This consideration means that the "evolutionary asymmetry objection" gives us less reason to be optimistic about the future than one would be if one ignored the likelihood of competitive futures and the seemingly adaptive role of disproportionally intense suffering. (See also "Beware underestimating the probability of very bad outcomes".)

Reply to excerpts from Shulman, 2012

The following are a couple of excerpts from Shulman's "Are Pain and Pleasure Equally Energy-Efficient?", along with my replies to them. Shulman's essay seems to be the most cited exposition of (something in the ballpark of) the argument that I am critiquing here, and hence I find it worth replying to key parts of that essay.

Shulman first defines two quantities, namely "hedons per joule" of the state of matter that produces the most pleasure per unit of energy (which he calls H), and "dolors per joule" of the state of matter that produces the most pain per unit of energy (which he calls D). (As Simon Knutsson has stressed, it is important to be clear that these respective quantities are a measure of energy-efficiency and not intensity; after all, the pleasure intensity of the state of matter that produces the most pleasure per unit of energy could in principle be extremely low.)

Shulman then proceeds to write:

> By symmetry, my default expectation would be that H=D.

In addition to questions about the nature of the "hedons" and "dolors" invoked in these respective quantities, it is natural to ask what justifies the assumption of symmetry (i.e. "by symmetry"). This key premise does not seem justified in Shulman's essay. And given that there are many arguments against thinking of pleasure and suffering in symmetric terms (including at the level of phenomenology), it seems that at least a minimal defense is required (see e.g. Sherman, 2017; Vinding, 2022; Knutsson, 2022).

Shulman again:

> In humans, the pleasure of orgasm may be less than the pain of deadly injury, since death is a much larger loss of reproductive success than a single sex act is a gain. But there is nothing problematic about the idea of much more intense pleasures, such that their combination with great pains would be satisfying on balance.

A key question in this context is what is meant by "satisfying on balance". In particular, what does "satisfying on balance" mean when some beings, or individual consciousness-moments, declare their experiences to be so bad that nothing could ever compensate for them?

In other words, we should be clear that the criterion according to which any state of suffering can be offset by other states, such that the totality is "satisfying on balance", is not a criterion that is in agreement with all the beings or consciousness-moments involved (Tomasik, 2015; Vinding, 2020, ch. 4; DiGiovanni, 2021a).

A criterion that admits of such a "satisfying balance" must forcefully override the preferences and value assessments of the worst-off beings and consciousness-moments who declare their experiences to be unbearable and unoutweighable by any purported positive goods. In contrast, abstaining from the creation of happiness (that is not needed or desired by existing beings) does not violate the preferences of anyone (DiGiovanni, 2021a).[43]

References

Ajantaival, T. (2021/2022). Minimalist Axiologies. Ungated

Anonymous. (2015). Negative Utilitarianism FAQ. Ungated

Breyer, D. (2015). The Cessation of Suffering and Buddhist Axiology. *Journal of Buddhist Ethics*, 22, pp. 533-560. Ungated

43 For helpful comments, I thank Teo Ajantaival, Tobias Baumann, Jesse Clifton, Emery Cooper, Anthony DiGiovanni, Simon Knutsson, Winston Oswald-Drummond, and Brian Tomasik.

Contestabile, B. (2014). Negative Utilitarianism and Buddhist Intuition. *Contemporary Buddhism*, 15(2), pp. 298-311. Ungated

DiGiovanni, A. (2019). Trading for Happiness: In Defense of Suffering-Focused Value Theory. Ungated

DiGiovanni, A. (2021a). Tranquilism Respects Individual Desires. Ungated

DiGiovanni, A. (2021b). Empirical Asymmetries Do Not Explain Suffering-Focused Intuitions. Ungated

Gloor, L. (2017). Tranquilism. Ungated

Gómez-Emilsson, A. (2021). A Conversation with Roger Thisdell about Classical Enlightenment and Valence Structuralism. Ungated

Knutsson, S. (2016). Measuring happiness and suffering. Ungated

Knutsson, S. (2017). Reply to Shulman's "Are Pain and Pleasure Equally Energy-Efficient?". Ungated

Knutsson, S. (2022). Undisturbedness as the hedonic ceiling. Ungated

Lazari-Radek, K. & Singer, P. (2017). *Utilitarianism: A Very Short Introduction*. Oxford University Press.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty*. Oxford University Press. Ungated

Metzinger, T. (2017). Benevolent Artificial Anti-Natalism (BAAN). Ungated

Sherman, T. (2017). Epicureanism: An Ancient Guide to Modern Wellbeing. MPhil dissertation, University of Exeter. Ungated

Shulman, C. (2012). Are Pain and Pleasure Equally Energy-Efficient? Ungated

Tomasik, B. (2015). Are Happiness and Suffering Symmetric? Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Vinding, M. (2022). A phenomenological argument against a positive counterpart to suffering. Ungated

Reply to the scope neglect objection against value lexicality

Some views hold that no amount of mild discomfort can be worse than a single instance of extreme suffering (i.e. they endorse value lexicality between extreme suffering and mild discomfort). An objection to such views is that they are biased by scope neglect — our tendency to disregard the number of affected beings in our evaluations of a problem. Since we cannot comprehend the badness of a vast amount of mild discomfort, the objection goes, we cannot trust our intuitive assessment that extreme suffering is worse than any amount of mild discomfort. My aim in this brief post is to reply to this objection.

Scope neglect vs. intensity neglect

A problem with the scope neglect objection is that we plausibly have biases in the opposite direction as well, and it is not clear whether those biases are any weaker than is our scope neglect in these evaluations. Indeed, one could argue that the biases in the opposite direction are much stronger overall.

In particular, we have an empathy gap that means that we are unable to understand just how intense and bad extreme suffering actually is, especially when we ourselves are experiencing a state of mind that is relatively untroubled (Vinding, 2020, sec. 7.4). And while large numbers can be difficult to comprehend, one could argue that we do at least have *some* rough understanding of what they are and how they work. Likewise, we understand what mild discomfort is like, and we know that states of mild discomfort feel quite bearable no matter how many of them there are.

In contrast, some people who have undergone extreme suffering report that the badness of such suffering is wholly beyond comprehension for those who are spared from it. As torture victim Jacobo Timerman said about the pain he experienced during torture: “It is a pain without points of reference, revelatory symbols, or clues to serve as indicators” (as quoted in Mayerfeld, 1999, p. 42; see also p. 38).

Hence, if one invokes scope neglect as an objection to value lexicality between mild discomfort and extreme suffering, it seems that one needs to explain why scope neglect is more of a distorting factor than is our empathy gap and the “intensity neglect” and “badness neglect” that it plausibly gives rise to.

Scope neglect applied beyond its limits?

Studies on scope neglect tend to show that the amount of money that people will donate to help a group of beings (who are all afflicted by the same ill) does not meaningfully increase as the number of afflicted beings increases, and in some cases the willingness to help even decreases (Desvousges et al., 1992; Kogut & Ritov, 2005; Cameron & Payne, 2011).

Such a donation pattern is in tension with the plausible moral premise that a greater number of beings afflicted by the same ill merit greater help than do fewer beings afflicted by the same ill. Yet note how this moral premise has limited relevance to the question of whether large amounts of mild discomfort can be added up to be worse than extreme suffering — i.e. comparisons that involve very different intensities of suffering, or different kinds of bads more generally. The psychological studies on scope neglect do not explore such comparisons, nor do they demonstrate that commonsense evaluations of tradeoffs between very different kinds of bads are implausible.

To be sure, the fact that we display a scope neglect in our evaluations of similar bads may be a reason to think that our evaluations of different kinds of bads might also be influenced by scope neglect. But again, even if we grant that scope neglect exerts a significant influence in such comparisons, it still does not follow that this distorting factor results in a judgment that *overall* underestimates the badness of many instances of mild discomfort compared to the badness of extreme suffering. And the points reviewed in this section suggest that scope neglect may be a weaker and less relevant factor than one might otherwise have expected, i.e. if one did not take its conceptual and empirical background properly into account.

Analogies to other lexical views

To further question the strength and relevance of scope neglect as an objection to value lexicality, it may be helpful to consider some other examples of lexical views.

For instance, consider an axiological view according to which extreme suffering is lexically worse than ugly art, even though ugly art is itself bad (according to that view). Such a view is not wholly fanciful, since some philosophers hold that art has final value (Stang, 2012, p. 271). And *if* art has final value (or disvalue), it seems plausible that its value is always eclipsed by the disvalue of extreme suffering (cf. Mayerfeld, 1999, p. 196; Vinding, 2020, p. 86).

We could likewise take an example that involves lexicality between different experiences. In particular, we could consider a version of the above-mentioned view in the experiential realm: one may hold that extreme suffering is lexically worse than any hedonically neutral experience of ugly

art, even though hedonically neutral experiences of ugly art are themselves bad (according to that view).

Is scope neglect to blame?

If someone endorses one of these alternative lexical views, is it plausible to argue that they endorse lexicality because of scope neglect? Do they simply fail to appreciate the collective disvalue of very large amounts of ugly art or experiences of ugly art? A hypothetical proponent of these views may object that they do not. They may argue that there is a stark qualitative difference between the badness of (experiences of) ugly art and extreme suffering. And they may further argue that this qualitative difference is not changed or overridden by the addition of more (experiences of) ugly art. The latter can never gain a badness that is equivalent to the badness of unbearable suffering.

My point with these analogies is that proponents of value lexicality between extreme suffering and mild discomfort could make the same argument, in that they may contend that there is a similar qualitative difference between the badness of mild discomfort and the badness of extreme suffering.

In particular, one may argue that there is nothing implausible about thinking that states of mild discomfort are bad in a qualitatively different way than are states of extreme suffering, given that states of extreme suffering *feel* qualitatively different, and given that they are likely mediated by different or additional brain circuits. And just like in the case of (experiences of) ugly art, one can reasonably argue that the qualitative difference in badness between mild discomfort and extreme suffering cannot be overridden by simply adding more instances of mild discomfort. Such addition does not render the merely uncomfortable truly horrific at any point.

Theoretical evaluations vs. practical decisions: An important distinction

Finally, it is worth highlighting the difference between 1) the evaluations that we make in hypothetical thought experiments, and 2) the decisions that we would make in real-world scenarios involving empirical uncertainty. After all, one may give opposite answers to similar dilemmas depending on which of these two kinds of assessments we are concerned with.

Specifically, someone who endorses strong lexicality between extreme suffering and mild discomfort would say that no amount of mild discomfort could be worse than a single instance of extreme suffering in the purely hypothetical case. Yet in the practical case, where we introduce uncertainty regarding what other beings experience, a proponent of strong lexicality need not — and arguably should not — maintain that a single state that *appears* to involve extreme suffering is worse than any number of states that *appear* to merely involve mild discomfort. The reason, in short, is that the empirical uncertainty means that a large number of states that *appear* to only

involve mild discomfort also involve some amount of extreme suffering in expectation. And hence for a sufficiently large number of states that *appear* to only involve mild discomfort, the expected amount of extreme suffering among those states will be larger than the expected amount of extreme suffering in a single state that appears to involve extreme suffering.

This point is important for a couple of reasons. First, it is important because it may distort our evaluations of the plausibility of value lexicality. In particular, if we fail to make it clear that we are considering a purely hypothetical thought experiment that involves absolutely no uncertainty, we may in turn fail to control for real-world intuitions that implicitly track uncertainty, and which intuit — with practical validity — that it would be highly risky to create inconceivably vast numbers of states that appear to only involve mild discomfort. Yet such intuitions about practical uncertainty should not distort our views in the purely theoretical case.

Second, the point is important because it illustrates how scope neglect plausibly *is* an important factor, albeit chiefly in the practical case. That is, if we were to make a real-world decision, it seems that scope neglect could well lead us to intuitively underestimate the expected amount of extreme suffering found among a vast number of states that *appear* to merely involve mild discomfort. Yet this perspective on the relevance of scope neglect in practice is wholly consistent with value lexicality between mild discomfort and extreme suffering at the theoretical level.

This is why we need to be careful to clarify whether we are talking about a hypothetical case involving no uncertainty versus a practical case that inevitably involves great uncertainty.[44]

References

Cameron, C. & Payne, B. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *J Pers Soc Psychol*, 100(1), pp. 1-15.

Desvousges, W. et al. (1992/2010). *Measuring nonuse damages using contingent valuation: An experimental evaluation of accuracy*. RTI Press. Ungated

Kogut, T. & Ritov, I. (2005). The singularity of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes,* 97, pp. 106-116.

Mayerfeld, J. (1999). *Suffering and Moral Responsibility*. Oxford University Press.

Stang, N. (2012). Artworks are not valuable for their own sake. *The Journal of Aesthetics and Art Criticism*, 70(3), pp. 271-280.

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

44 Thanks to Winston Oswald-Drummond for bringing this objection to my attention.

Part III: Practical Issues

Why altruists should be cooperative

Summary

There are many reasons to adopt a cooperative approach to altruism. A cooperative approach can enable positive-sum compromises, make people more willing to join our efforts, and promote collaboration with others toward shared ends. Last but not least, greater cooperation can help reduce some of the main risk factors for s-risks.

Definition

What I mean by a "cooperative approach" in this context includes both common decency — i.e. being friendly and respectful toward others — as well as being willing to strike compromises with people who hold different values. These two notions of "cooperative" are distinct, yet closely related. For example, being friendly toward others is often a prerequisite for gainful compromises.

Gains from compromise

Agents with different values can often achieve mutual gains if they are willing to compromise, also known as gains from moral trade.

For example, rather than engaging in zero-sum competition to achieve 100 percent of one's aims, it may be possible for competing factions to engage in a positive-sum compromise that enables both sides to achieve 80 percent of their respective aims, whereas they might otherwise have achieved far less (e.g. if they engaged in zero-sum competition).

The potential gains from compromise represent one of the many reasons to promote compromise and to try to be considerate of other people's values in our deliberations. Brian Tomasik has explored some ways to promote compromise here.

Movement building

People will likely be less inclined to join altruistic efforts or movements if these are associated with uncooperative and unfriendly attitudes. In contrast, a movement thoroughly imbued with cooperativeness and friendliness invites people in, and makes potential contributors more willing to be associated with that movement. After all, most people will probably feel that it reflects more

positively on them to be involved with a project that explicitly endorses cooperation compared to being associated with projects that stand for the opposite.

This latter consideration is important given how concerned most people are, quite rationally, about the signaling effects of their behavior. A cooperative approach can mean the difference between newcomers deciding to contribute to or oppose the efforts of a given movement.

Cooperation also benefits altruistic movements internally: it enables better collaboration and helps foster a healthy environment in which individuals can remain sustainably motivated and productive.

Cooperation with others toward shared ends

Many of the aims we care about are also shared by other people, even if other people do not prioritize those aims quite as highly as we do. For example, everyone can agree that the worst s-risks would be worth avoiding. Likewise, most people can agree that it is worth preventing intense suffering if it can be achieved at a trivial cost, and would thus be willing to give at least weak support toward this end.

This considerable degree of agreement on values represents a vast potential resource that a cooperative approach can help us capitalize on. Conversely, a hostile approach — e.g. being shaming and unfriendly — risks pushing people away, and thus risks throwing away this vast potential resource. (I say a bit more on this in Vinding, 2020, sec. 10.2.)

Emphasizing the substantial common ground among us instead of focusing mainly on our disagreements seems a good strategy for advancing our shared aims.

Avoiding conflicts

There are some reasons to think that the most worrying s-risks are agential in nature, and most agential s-risks likely result from conflicts or animosity of some kind.

Specifically, increased polarization, hatred, and retribution may be among the main risk factors for agential s-risks. How to best mitigate these risk factors is an open question, yet it is probably helpful if altruists adopt and endorse a cooperative and conciliatory approach.

In contrast, if altruists are needlessly provocative and antagonistic, this increases the risk that altruistic values will be the target of ill will and revenge. This is dangerous in a world where most agents do not care enough to strongly protect the value entities that altruists care about.

Why don't we cooperate?

One reason we may fail to cooperate is that we are biased against it. For example, a drive to signal commitment to one's favorite cause may push one toward posturing behavior that berates those who diverge even slightly from the supposedly ideal path. In other words, there often is a conflict between 1) trying to *appear* sincerely dedicated to one's cause, and 2) doing that which is strategically optimal. And our hidden motives will often pull us toward the former.

Tribal biases may also play a role: it lies deep in human psychology to be opposed to the (perceived) outgroup, and to conspicuously showcase opposition to the outgroup in a way that is visible to one's ingroup. This can prevent us from engaging in positive-sum compromises with other groups, and may impede effective collaboration toward shared ends.

(Note that this kind of tribal dynamic is often just as strong among agents who agree on values yet who disagree on empirical matters; here too, it is necessary that we strike compromises and suppress primitive impulses, Vinding, 2020, 10.4.)

There may, of course, also be genuinely good reasons not to cooperate (further) in many cases — for example, when other agents are aggressive, or if our level of cooperation has become harmfully overaccommodating (Tomasik, 2014; Vinding, 2020, sec. 10.2). Additionally, there may be biases that push our behavior toward too much cooperation, such as drives toward social conformity and fear of confrontation.

There is a balance to be struck between standing up for one's principles on the one hand, and being cooperative on the other. The considerations outlined above do not suggest that we should stop standing by our principles or cease to defend victims of extreme suffering. We most certainly should not. But the considerations reviewed here do suggest that we will be better able to prevent extreme suffering if we pursue principled compassion through a cooperative approach.[45]

References

Baumann, T. (2018). A typology of s-risks. Ungated

Baumann, T. (2019). Risk factors for s-risks. Ungated

Baumann, T. (2020). Common ground for longtermists. Ungated

Ord, T. (2015). Moral Trade. *Ethics*, 126, pp. 118-138. Ungated

45 For helpful comments, I'm grateful to Michael Aird, Tobias Baumann, Anthony DiGiovanni, and Rupert McCallum.

Schubert et al. (2017). Considering Considerateness: Why communities of do-gooders should be exceptionally considerate. Ungated

Tomasik, B. (2013a). Gains From Trade Through Compromise. Ungated

Tomasik, B. (2013b). Possible Ways to Promote Compromise. Ungated

Tomasik, B. (2014). Reasons to Be Nice to Other Value Systems. Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

Suffering-focused ethics and the importance of happiness

It seems intuitive to think that suffering-focused moral views imply that it is unimportant whether people live fulfilling lives. Yet the truth, I will argue, is in many ways the opposite — especially for those who are trying to reduce suffering effectively with their limited resources.

Personal sustainability and productivity

One reason in favor of living fulfilling lives is that we cannot work to reduce suffering in sustainable ways otherwise. Indeed, not only is a reasonably satisfied mind a precondition for sustainable productivity in the long run, but also for our productivity on a day-to-day basis, which is often aided by a strong passion and excitement about our work projects. Suffering-focused ethics by no means entails that excitement and passion should be muted.

Beyond aiding our productivity in work-related contexts, a strong sense of well-being also helps us be more resilient in the face of life's challenges — things that break, unexpected expenses, unfriendly antagonists, etc. Cultivating a sense of fulfillment and a sound mental health can help us better handle these obstacles as well.

Signaling value

This reason pertains to the social rather than the individual level. If we are trying to create change in the world, it generally does not help if we ourselves are miserable. People often decide whether they want to associate with (or distance themselves from) a group of people based on perceptions of the overall wellness and mental health of its adherents. And this is not entirely unreasonable, as these factors arguably do constitute *some* indication of the practical consequences of associating with the group in question.

If failing to prioritize our own well-being has bad consequences in the bigger picture, such as scaring people away from joining our efforts to create a better future, then this failure is not recommended by consequentialist suffering-focused views.

To be clear, my point here is not that suffering-focused agents should be deceptive and try to display a fake and inflated sense of well-being (such deception would likely have many bad consequences). Rather, the point is that we have good reasons to cultivate *genuine* physical and mental health, both for the sake of our personal productivity *and* our ability to inspire others.

A needless hurdle to the adoption of suffering-focused views

A closely related point has to do with people's evaluations of suffering-focused views more directly (as opposed to the evaluations of suffering-focused communities and individuals). People are likely to judge the acceptability of a moral view based in part on the expected psychological consequences of its adoption — will it enable me to pursue the lifestyle I want, to maintain my social relationships, and to seem like a good and likeable person?

Indeed, modern moral and political psychology suggests that these social and psychological factors are strong determinants of our moral and political views, and that we usually underestimate just how much these "non-rationalist" factors influence our views (see e.g. Haidt, 2012, part III; Tuschman, 2013, ch. 22; Simler, 2016; Tooby, 2017).

This is then another good reason to seek to both emphasize and exemplify the compatibility of suffering-focused views and a healthy and fulfilling life. Again, if failing in this regard tends to prevent people from prioritizing the reduction of suffering, then a true extrapolation of suffering-focused views will militate against such a failure, and instead recommend a focus on cultivating an invitingly healthful state of mind.

In sum, there is no inherent tension between living a healthy and fulfilling life and at the same time being committed to reducing the most intense forms of suffering.

Moral circle expansion might increase future suffering

Expanding humanity's moral circle such that it includes all sentient beings seems among the most urgent and important missions before us. And yet there is a significant risk that such greater moral inclusion might in fact end up increasing future suffering. As Brian Tomasik notes:

> One might ask, "Why not just promote broader circles of compassion, without a focus on suffering?" The answer is that more compassion by itself could increase suffering. For example, most people who care about wild animals in a general sense conclude that wildlife habitats should be preserved, in part because these people aren't focused enough on the suffering that wild animals endure. Likewise, generically caring about future digital sentience might encourage people to *create* as many happy digital minds as possible, even if this means also increasing the risk of digital suffering due to colonizing space. Placing special emphasis on reducing suffering is crucial for taking the right stance on many of these issues.

Indeed, many classical utilitarians do include non-human animals in their moral circle, yet they still consider it permissible, indeed in some sense morally good, that we bring individuals into existence so that they can live "net positive lives" and we can eat them (I have argued that this view is mistaken, almost regardless of what kind of consequentialist view one assumes). And some even seem to think that most lives on factory farms might plausibly be such "net positive lives". A wide circle of moral consideration clearly does not guarantee an unwillingness to allow large amounts of suffering to be brought into the world.

More generally, there is a considerable number of widely endorsed ethical positions that favor bringing about larger rather than smaller populations of the beings who belong to our moral circle, at least provided that certain conditions are met in the lives of these beings. And many of these ethical positions have quite loose such conditions, which implies that these views can easily permit, and even demand, the creation of a lot of suffering for the sake of some (supposedly) greater good.

Indeed, the truth is that even a view that requires an enormous amount of happiness to outweigh a given amount of suffering might still easily permit the creation of large amounts of suffering, as illustrated by the following consideration (quoted from the penultimate chapter of my book on effective altruism):

> consider the practical implications of the following two moral principles: 1) we will not allow the creation of a single instance of the worst forms of suffering [...] for any amount of happiness, and 2) we will allow one day of such suffering for ten years of the most sublime happiness. What kind of future would we accept with these respective principles? Imagine a future in which we colonize space and maximize the number of sentient beings that the accessible universe can sustain over the entire course of the future, which is probably more than 10^30. Given this number of beings, and assuming that these beings each live a hundred years, principle 2) above would appear to permit a space colonization that all in all creates more than 10^28 years of [the worst forms of suffering], provided that the other states of experience are sublimely happy. This is how extreme the difference can be between principles like 1) and 2); between whether we consider suffering irredeemable or not. And notice that even if we altered the exchange rate by orders of magnitude — say, by requiring 10^15 times more sublime happiness per unit of extreme suffering than we did in principle 2) above — we would still allow an enormous amount of extreme suffering to be created; in the concrete case of requiring 10^15 times more happiness, we would allow more than 10,000 billion years of [the worst forms of suffering].

This highlights the importance of thinking deeply about which trade-offs, if any, we find acceptable with respect to the creation of suffering, including extreme suffering.

The considerations above concerning popular ethical positions that support larger future populations imply that there is some probability — a seemingly low yet still significant probability — that a more narrow moral circle may in fact lead to less future suffering for the morally excluded beings (e.g. by making efforts to bring these beings into existence, on Earth and beyond, less likely).

Implications

In spite of this risk, I still consider generic moral circle expansion quite beneficial in expectation. Yet it seems less beneficial, and significantly less robust (with respect to the goal of reducing extreme suffering) than does the promotion of suffering-focused values. And it seems less robust and less beneficial still than does the twin-track strategy of focusing on both expanding our moral circle *and* deepening our concern for suffering. Both seem necessary yet insufficient on their own. If we deepen concern for suffering without broadening the moral circle, our deepened concern risks failing to pertain to the vast majority of sentient beings. On the other hand, if we broaden our moral circle without deepening our concern for suffering, we may end up allowing the beings within our moral circle to endure enormous amounts of suffering.

On fat-tailed distributions and s-risks

Summary

It is sometimes suggested that since the severity of many kinds of moral catastrophes (e.g. wars and natural disasters) fall along a power-law distribution, efforts to reduce suffering should focus on "a few rare scenarios where things go very wrong". While this argument appears quite plausible on its face, it is in fact a lot less obvious than it seems at first sight. Specifically, a fat-tailed distribution need not imply that a single or even a few sources of suffering account for most future suffering in expectation, let alone that we should mostly prioritize a single or a few sources of suffering.

Introduction

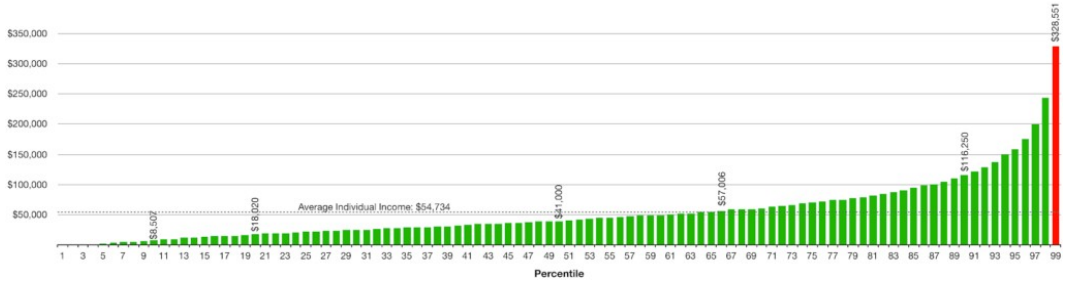
In his post Is most expected suffering due to worst-case outcomes?, Tobias Baumann explores how skewed the distribution of future sources of suffering might be. His conclusion, in short, is that worst-case outcomes may well dominate, but that it is unclear *to what degree* we should expect future suffering to be concentrated in worst-case outcomes.

My aim in this post is not to shed further light on this matter. Instead, my aim is to clarify some key points concerning what follows and doesn't follow if we *do* assume that future (expected) suffering conforms to a highly fat-tailed distribution. In addition, I will outline some reasons to give considerable weight to a broad approach to s-risk reduction as opposed to focusing on a few narrow risks.

Observed distributions

Perhaps the best place to start is to look at a couple of real-world examples of fat-tailed distributions.

First, consider income distribution in the US in 2019:



This is a fat-tailed distribution, with the top one percent of earners (the red bar) making far more than the rest. Still, the income of the richest one percent is "only" about six percent of the total income. The one percent are far richer than the rest, but their income is nowhere near accounting for 50 percent of total income. Top seven percent of earners make 25 percent of the wealth, while the top 20 percent earn roughly 50 percent of the total income.

Another example with a significantly heavier tail is wealth distribution in the United States: in 2007, the top one percent possessed 35 percent of the total wealth, the next four percent owned 27 percent, while the bottom 60 percent had less than five percent of total US wealth.

Implication: Less priority to the low-end

What would be the upshot if future sources of (expected) suffering followed any of these distributions?

Perhaps the most obvious implication is found, not at the crowded end, but rather toward the bottom of the distribution. For example, in the first two distributions above, almost negligibly little wealth is found among those who have the least. So if one were to tax people efficiently, it would make sense to largely ignore the bottom 20 percent of the income distribution, and the bottom 60 percent of the wealth distribution. Sure, there would still be *some* wealth to gain there, but if we had severely limited resources, the effort would hardly be worth it.

Likewise, in terms of our analogy to future sources of suffering, it would make sense to pay less attention to the sources of suffering — or future scenarios, if we conceptualize in these terms — that contain relatively little suffering in expectation, as there is *comparatively* little suffering to reduce there.

Yet note that there is a long way from this conclusion to the claim that we should focus almost exclusively on a small space of possible scenarios or sources of suffering at the other end of the distribution.

Fat tails can still be wide

As the examples above show, the fact that a distribution is fat-tailed does not necessarily imply that a bulk of the distribution is found in a tiny sliver. For instance, in the case of the US income distribution, one would have to include the top 20 percent of earners in order to cover 50 percent of the total income — a much broader range than just the top one percent.

Even in the heavily skewed case of wealth inequality, we would still need to go markedly beyond the top one percent of earners in order to cover 50 percent, and we must go beyond the top ten percent if we are to cover three quarters of the total wealth.

Moreover, the top one percent of a distribution like this is usually rather diverse, which leads us to the next point.

The top one percent is not a narrow set

Even if we grant that potential sources of future suffering follow a power-law distribution similar to wealth distribution in the US — and we should remember that it is quite uncertain whether it does — it does not follow that "most of the expected suffering comes from a few rare scenarios where things go very wrong".

As a case in point, the wealthiest one percent in the US is a diverse group of people, at least in terms of how they acquired their wealth (e.g. from many different industries), and they are also a rather large group *in absolute terms* (one percent of the US population is still more than three million people). Similarly, the worst one percent of scenarios or sources of future suffering could be a rather diverse set, with a substantial number of different sources of suffering.

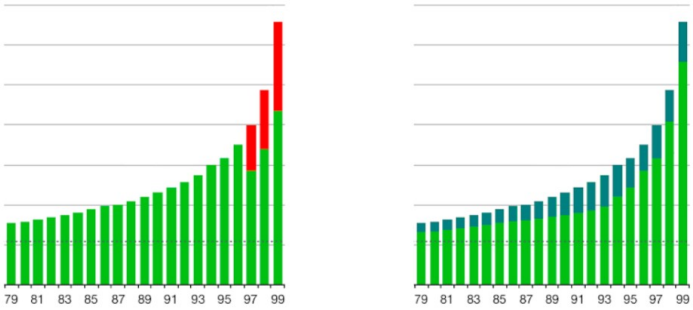
To illustrate this point, consider a real-world example that is even more extreme than any of those we have seen above, namely the frequency of words. It is a commonly observed pattern that words fall along a very fat-tailed distribution, conforming to what is known as Zipf's law. For instance, in the Brown Corpus of American English text, the most frequent word ("the") accounts for around seven percent of all words, while the second most frequent word ("of") accounts for 3.5 percent. Only 135 words — 0.27 percent — account for half of the words.

And yet 135 is still a good deal more than "a few". Indeed, focusing on preventing 135 different sources of suffering (e.g. different types of agential and incidental s-risks) hardly qualifies as a particularly narrow focus in an absolute sense, although it sure is quite narrow in a relative sense, when compared to the *entire space* of possible sources of suffering. We should thus be careful not to confuse such relative narrowness with absolute narrowness. The absolute claim that we should focus on a few specific scenarios is far stronger and much less warranted than a relative claim of the sort that we should focus on, say, the top three percent.

A broad focus may still be ideal given fat tails

A fat-tailed distribution of future sources of suffering does not necessarily imply that we should focus exclusively, or even primarily, on a small class of worst-case outcomes, such as the top three percent. It could well be that we can reduce more suffering by focusing on a much broader scope.

The figures below illustrate this point:



The figure on the left illustrates our impact given a strategy where we devote all our marginal resources toward reducing the sources of suffering found in the upper three percentiles (which is still a rather broad class), while the figure on the right illustrates our impact given a strategy where we target many different sources of suffering. The red and blue colorings represent expected suffering reduced, with the broad strategy on the right reducing significantly more suffering in expectation, even as it is less effective among the very top percentiles. (The impact on both figures is exaggerated for illustrative purposes; realistically, we should not assume that our marginal actions will have this much of an effect on expected global outcomes.)

My point here is not that the broad strategy in fact does reduce the most suffering, but simply that it *may* do so, and hence that even a very fat-tailed distribution does not by itself imply that we should pursue the narrow strategy on the left over the broader one on the right.

After all, it could be the case that there are common actions we can take that help reduce suffering from a broad range of sources (such as by increasing the willingness and resources devoted toward their reduction), and there might likewise be low-hanging fruit to reap in each of the respective classes of causes and interventions.

Note that movement building plausibly accomplishes both: it potentially targets many different sources of suffering (indirectly), and may enable us to pick low-hanging fruit with respect to many different sources of future suffering —not to mention that it can help us gain greater clarity of the risk landscape, which leads us to another set of reasons to favor a broad approach.

Epistemic reasons to favor a broad approach

The fact that we have a lot of uncertainty about the distribution of future sources of suffering is an additional reason to favor a broader, more robust approach. For not only can movement building and broad research on prioritization give us a better sense of the general *shape* of the distribution of future suffering, it should also help us get a better sense of which exact sources of suffering that are most worrisome, as well as how we can best target those sources of suffering. The greater our uncertainty is on these matters, the greater is the risk that our favored priorities and interventions are misguided, and the more important it is that we get more people to help us update our views.

A very narrow focus is especially risky given vast uncertainty, since there is then a greater probability that we are spending most of our resources on something that is suboptimal. Inspired by the illustrations above, we may think of it as drawing among a hundred different tickets, aiming to draw the single longest one based on our best knowledge. And if our "best knowledge" relies strongly on speculation, the probability that the ticket we draw is not in fact the longest one will be quite large. It is, after all, quite possible that the very worst sources of future suffering are not among those we currently consider most worrisome — they could be among those that we currently give little weight, or indeed be unknown unknowns.

Note that the epistemic reasons in favor of a broad approach listed here are quite independent of the reasons listed in the previous section — i.e. the fact that we may take actions that have a positive influence on many different sources of suffering, e.g. movement building, and the fact that we may pick low-hanging fruit in the prevention of many different risks. (Those points would apply even if we knew which sources of suffering are most worrisome.) In combination, these distinct reasons provide a rather strong case for giving at least considerable weight to a broad and robust approach to reducing s-risks.

Beware biases

The tension between narrow and broad approaches underscores the importance of being aware of the biases that might influence our assessments of these matters. These include our tendency toward narrow framing in general, and belief digitization in particular: our inclination to focus purely on the single hypothesis we consider most plausible, and to give insufficient weight to hypotheses we consider less likely. This bias plausibly pulls us toward a narrow approach.

Beyond that, the fact that our brains did not evolve to consider complicated and uncertain questions concerning global prioritization renders it plausible that we generally underestimate the extent of our uncertainty on such matters (Vinding, 2020, sec. 9.2). In contrast, we seem to have little reason

to think that we err to a similar degree in the opposite direction, toward overemphasizing our uncertainty. Indeed, we are generally prone to overconfidence bias, which is an additional reason to expect us to underestimate the extent of our uncertainty, and to be overly confident about our current priorities.

Antinatalism and reducing suffering: A case of suspicious convergence

Two positions are worth distinguishing. One is the view that we should reduce (extreme) suffering as much as we can for all sentient beings. The other is the view that we should advocate for humans not to have children.

It may seem intuitive to think that the former position implies the latter. That is, to think that the best way to reduce suffering for all sentient beings is to advocate for humans not to have children. My aim in this brief essay is to outline some of the reasons to be skeptical of this claim.

Suspicious convergence

Lewis, 2016 warns of "suspicious convergence", which he introduces with the following toy example:

> **Oliver: …** Thus we see that donating to the opera is the best way of promoting the arts.
>
> **Eleanor:** Okay, but I'm principally interested in improving human welfare.
>
> **Oliver:** Oh! Well I think it is *also* the case that donating to the opera is best for improving human welfare too.

The general point is that, for any set of distinct altruistic aims or endeavors we may consider, we should be a priori suspicious of the claim that they are perfectly convergent — i.e. that directly pursuing one of them also happens to be the very best thing we can do for achieving the other. Justifying such a belief would require good, object-level reasons. And in the case of the respective endeavors of reducing suffering and advocating for humans not to procreate, we in a sense find the opposite, as there are good reasons to be skeptical of a strong degree of convergence, and even to think that such antinatalist advocacy might *increase* future suffering.

The marginal impact of antinatalist advocacy

A key point when evaluating the impact of altruistic efforts is that we need to think at the margin: how does our particular contribution change the outcome, in expectation? This is true whether our aims are modest or maximally ambitious — our actions and resources still represent but a very small fraction of the total sum of actions and resources, and we can still only exert relatively small pushes toward our goals.

Direct effects

What, then, is the marginal impact of advocating for people not to have children? One way to try to answer this question is to explore the expected effects of preventing a single human birth. Antinatalist analyses of this question are quick to point out the many harms caused by a single human birth, which must indeed be considered. Yet what these analyses tend not to consider are the harms that a human birth would prevent.

For example, in his book *Better Never to Have Been*, David Benatar writes about "the suffering inflicted on those animals whose habitat is destroyed by encroaching humans" (p. 224) — which, again, should definitely be included in our analysis. Yet he fails to consider the many births and all the suffering that would be *prevented* by an additional human birth, such as due to its marginal effects on habitat reduction ("fewer people means more animals"). As Brian Tomasik argues, when we consider a wider range of the effects humans have on animal suffering, "it seems plausible that encouraging people to have fewer children actually causes an *increase* in suffering and involuntary births."

This highlights how a one-sided analysis such as Benatar's is deeply problematic when evaluating potential interventions. We cannot simply look at the harms prevented by our pet interventions without considering how they might lead to *more* harm. Both things must be considered.

To be clear, the considerations above regarding the marginal effects of human births on animal suffering by no means represent a complete analysis of the effects of additional human births, or of advocating for humans not to have children. But they *do* represent reasons to doubt that such advocacy is among the very best things we can do to reduce suffering for all sentient beings, at least in terms of the direct effects, which leads us to the next point.

Long-term effects

Some seem to hold that the main reason to advocate against human procreation is not the direct effects, but rather its long-term effects on humanity's future. I agree that the influence our ideas and advocacy efforts have on humanity's long-term future are plausibly the most important thing about them, and I think many antinatalists are likely to have a positive influence in this regard by highlighting the moral significance of suffering (and the relative *in*significance of pleasure).

But the question is why we should think that the best way to steer humanity's long-term future toward less suffering is to argue for people not to have children. After all, the space of possible interventions we could pursue to reduce future suffering is vast, and it would be quite a remarkable

coincidence if relatively simple interventions — such as advocating for antinatalism or veganism — happened to be the very best way to reduce suffering, or even *among* the very best ways.

In particular, the greatest risk from a long-term perspective is that things somehow go awfully wrong, and that we counterfactually greatly increase future suffering, either by creating additional sources of suffering in the future, or by simply failing to reduce existing forms of suffering when we could. And advocating for people not to have children seems unlikely to be among the best ways to reduce the risk of such failures — again since the space of possible interventions is vast, and interventions that are targeted more directly at reducing these risks, including the risk of leaving wild-animal suffering unaddressed, are probably significantly more effective than is advocating for humans not to procreate.

Better alternatives?

If our aim is to reduce suffering for all sentient beings, a plausible course of action would be to pursue an open-ended research project on *how* we can best achieve this aim. This is, after all, not a trivial question, and we should hardly expect the most plausible answers to be intuitive, let alone obvious. Exploring this question requires epistemic humility, and forces us to contend with the vast amount of empirical uncertainty that we are facing.

I have explored this question at length in Vinding, 2020, as have other individuals and organizations elsewhere. One conclusion that seems quite robust is that we should focus mostly on avoiding bad outcomes, whereas comparatively suffering-free future scenarios merit less priority. Another robust conclusion is that we should pursue a pragmatic and cooperative approach when trying to reduce suffering (see also Vinding, 2020, ch. 10) — not least since future conflicts are one of the main ways in which worst-case outcomes might materialize, and hence we should generally strive to reduce the risk of such conflicts.

In more concrete terms, antinatalists may be more effective if they focus on defending antinatalism for wild animals in particular. This case seems both easier and more important to make given the overwhelming amount of suffering and early death in nature. Such advocacy may both have more beneficial near-term and long-term effects, being less at risk of increasing non-human suffering in the near term, and plausibly being more conducive to reducing worst-case risks, whether these entail spreading non-human life or simply failing to reduce wild-animal suffering.

Broadly speaking, the aim of reducing suffering would seem to recommend efforts to identify the main ways in which humanity might cause — or prevent — vast amounts of suffering in the future, and to find out how we can best navigate accordingly. None of these conclusions seem to support

efforts to convince people not to have children as a particularly promising strategy, though they likely do recommend efforts to promote concern for suffering more generally.

Priorities for reducing suffering: Reasons not to prioritize the Abolitionist Project

I discussed David Pearce's Abolitionist Project in Chapter 13 of my book on *Suffering-Focused Ethics*. The chapter is somewhat brief and dense, and its main points could admittedly have been elaborated further and explained more clearly. This post seeks to explore and further explain some of these points.

A good place to start might be to highlight some of the key points of agreement between David Pearce and myself.

- First and most important, we both agree that minimizing suffering should be our overriding moral aim.
- Second, we both agree that we have reason to be skeptical about the possibility of digital sentience — and at the very least to not treat it as a foregone conclusion — which I note from the outset to flag that views on digital sentience are unlikely to account for the key differences in our respective views on how to best reduce suffering.
- Third, we agree that humanity should ideally use biotechnology to abolish suffering throughout the living world, provided this is indeed the best way to minimize suffering.

The following is a summary of some of the main points I made about the Abolitionist Project in my book. There are four main points I would emphasize, none of which are particularly original (at least two of them are made in Brian Tomasik's Why I Don't Focus on the Hedonistic Imperative).

I.

Some studies suggest that people who have suffered tend to become more empathetic. This obviously does not imply that the Abolitionist Project is infeasible, but it does give us reason to doubt that abolishing the capacity to suffer in humans should be among our main priorities at this point.

To clarify, this is not a point about what we should do in the ideal, but more a point about where we should currently invest our limited resources, on the margin, to best reduce suffering. If we were to focus on interventions at the level of gene editing, other traits (than our capacity to suffer) seem more promising to focus on, such as increasing dispositions toward compassion. And yet

interventions focused on gene editing may themselves not be among the most promising things to focus on in the first place, which leads to the next point.

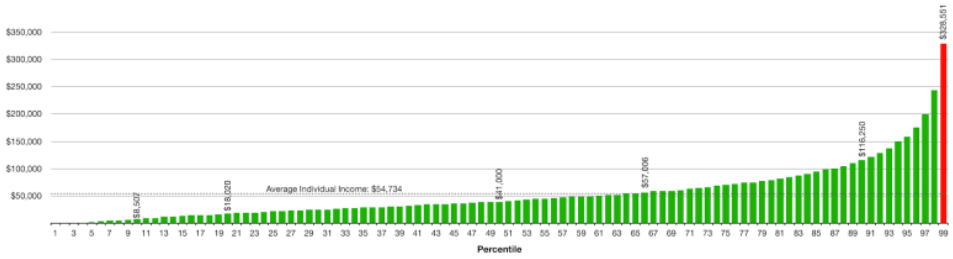
II.

For even if we grant that the Abolitionist Project should be our chief aim, at least in the medium term, it still seems that the main bottleneck to its completion is found not at the technical level, but rather at the level of humanity's values and willingness to do what would be required. I believe this is also a point that David and I mostly agree on, as he has likewise hinted, in various places, that the main obstacle to the Abolitionist Project will not be technical, but sociopolitical. This would give us reason to *mostly* prioritize the sociopolitical level on the margin — especially humanity's values and willingness to reduce suffering. And the following consideration provides an additional reason in favor of the same conclusion.

III.

The third and most important point relates to the distribution of future (expected) suffering, and how we can best prevent worst-case outcomes. Perhaps the most intuitive way to explain this point is with an analogy to tax revenues: if one were trying to maximize tax revenues, one should focus disproportionately on collecting taxes from the richest people rather than the poorest, simply because that is where most of the money is.

The visual representation of the income distribution in the US in 2019 found below should help make this claim more intuitive.



The point is that something similar plausibly applies to future suffering: in terms of the distribution of future (expected) suffering, it seems reasonable to give disproportionate focus to the prevention of worst-case outcomes, as they contain more suffering (in expectation).

Futures in which the Abolitionist Project is completed, and in which our advocacy for the Abolitionist Project helps bring on its completion, say, a century sooner, are almost by definition not the kinds of future scenarios that contain the most suffering. That is, they are not worst-case futures in which things go very wrong and suffering gets multiplied in an out-of-control fashion.

Put more generally, it seems to me that advocating for the Abolitionist Project is not the best way to address worst-case outcomes, even if we assume that such advocacy has a positive effect in this regard. A more promising focus, it seems to me, is again to increase humanity's overall *willingness* and *capacity* to reduce suffering (the strategy that also seems most promising for advancing the Abolitionist Project itself). And this capacity should ideally be oriented toward the avoidance of very bad outcomes — outcomes that to me seem most likely to stem from bad sociopolitical dynamics.

IV.

Relatedly, a final critical point is that there may be some downsides to framing our goal in terms of *abolishing* suffering, rather than in terms of *minimizing suffering in expectation*. One reason is that the former framing may invoke our proportion bias, or what is known in the literature as proportion dominance: our tendency to intuitively care more about helping 10 out of 10 individuals rather than helping 10 out of 100, even though the impact is in fact the same.

Minimizing suffering in expectation would entail abolishing suffering *if* that were indeed the way to minimize suffering in expectation, but the point is that it might not be. For instance, it could be that the way to reduce the most suffering in expectation is to instead mostly focus on *reducing the probability* and *mitigating the expected badness* of worst-case outcomes. And framing our aim in terms of abolishing suffering, rather than the more general and neutral terms of minimizing suffering in expectation, can hide this possibility somewhat. (I say a bit more about this in Section 13.3 in my book; see also this section.)

Moreover, talking about the complete abolition of suffering can leave the broader aim of reducing suffering particularly vulnerable to objections — e.g. the objection that *completely abolishing* suffering seems risky in a number of ways. In contrast, the aim of reducing intense suffering is much less likely to invite such objections, and is more obviously urgent and worthy of priority. This is another strategic reason to doubt that the abolitionist framing is optimal.

Lastly, it would be quite a coincidence if the actions that maximize the probability of the complete abolition of suffering were also exactly those actions that minimize extreme suffering in expectation; even as these goals are related, they are by no means the same. And hence to the extent that our main goal is to minimize extreme suffering, we should probably frame our objective in these terms rather than in abolitionist terms.

Reasons in favor of prioritizing the Abolitionist Project

To be clear, there are also things to be said in favor of an abolitionist framing. For instance, many people will probably find a focus on the mere *alleviation* and *reduction* of suffering to be too negative and insufficiently motivating, leading them to disengage and drop out. Such people may find it much more motivating if the aim of reducing suffering is coupled with an inspiring vision about the complete abolition of suffering and increasingly better states of superhappiness.

As a case in point, I think my own focus on suffering was in large part inspired by the Abolitionist Project and the *The Hedonistic Imperative*, which gradually, albeit very slowly, eased my optimistic mind into prioritizing suffering. Without this light and inspiring transitional bridge, I may have remained as opposed to suffering-focused views as I was eight years ago, before I encountered David's work.

Brian Tomasik writes something similar about the influence of these ideas: "David Pearce's *The Hedonistic Imperative* was very influential on my life. That book was one of the key factors that led to my focus on suffering as the most important altruistic priority."

Likewise, informing people about technologies that can effectively reduce or even abolish certain forms of suffering, such as novel gene therapies, may give people hope that we *can* do something to reduce suffering, and thus help motivate action to this end.

But I think the two reasons cited above count more as reasons to *include* an abolitionist perspective in our "communication portfolio", as opposed to making it our *main* focus — not least in light of the four considerations mentioned above that count against the abolitionist framing and focus.

A critical question

The following question may capture the main difference between David's view and my own.

In previous conversations, David and I have clarified that we both accept that the avoidance of worst-case outcomes is, plausibly, the main priority for reducing suffering in expectation.

This premise, together with our shared moral outlook, seems to recommend a strong focus on minimizing the risk of worst-case outcomes. The critical question is thus: *What reasons do we have to think that prioritizing and promoting the Abolitionist Project is the* ***single best way****, or even* ***among the best ways****, to address worst-case outcomes?*

As noted above, I think there are good reasons to doubt that advocating the Abolitionist Project is among the most promising strategies to this end (say, among the top 10 causes to pursue), even if we grant that it has positive effects overall, including on worst-case outcomes in particular.

Possible responses

Analogy to smallpox

A way to respond may be to invoke the example of smallpox: Eradicating smallpox was plausibly the best way to minimize the risk of "astronomical smallpox", as opposed to focusing on other, indirect measures. So why should the same not be true in the case of suffering?

I think this is an interesting line of argument, but I think the case of smallpox is disanalogous in at least a couple of ways. First, smallpox is in a sense a much simpler and more circumscribed phenomenon than is suffering. In part for this reason, the eradication of smallpox was much easier than the abolition of suffering would be. As an infectious disease, smallpox, unlike suffering, has not evolved to serve any functional role in animals. It could thus not only be eradicated more easily, but also without unintended effects on, say, the function of the human mind.

Second, if we were primarily concerned about not spreading smallpox to space, and minimizing "smallpox-risks" in general, I think it *is* indeed plausible that the short-term eradication of smallpox would not be the ideal thing to prioritize with marginal resources. (Again, it is important to here distinguish what humanity at large should ideally do versus what the, say, 1,000 most dedicated suffering reducers should do with most of their resources, on the margin, in our imperfect world.)

One reason such a short-term focus may be suboptimal is that the short-term eradication of smallpox is already — or would already be, if it still existed — prioritized by mainstream organizations and governments around the world, and hence additional marginal resources would likely have a rather limited counterfactual impact to this end. Work to minimize the risk of spreading life forms vulnerable to smallpox is far more neglected, and hence does seem a fairly reasonable priority from a "smallpox-risk minimizing" perspective.

Sources of unwillingness

Another response may be to argue that humanity's unwillingness to reduce suffering derives mostly from the sense that the problem of suffering is intractable, and hence the best way to increase our willingness to alleviate and prevent suffering is to set out technical blueprints for its prevention. In David's words, "we can have a serious ethical debate about the future of sentience only once we appreciate what is — and what isn't — technically feasible."

I think there is something to be said in favor of this argument, as noted above in the section on reasons to favor the Abolitionist Project. Yet unfortunately, my sense is that humanity's unwillingness to reduce suffering does not primarily stem from a sense that the problem is too vast and intractable. Sadly, it seems to me that most people give relatively little thought to the urgency of

(others') suffering, especially when it comes to the suffering of non-human beings. As David notes, factory farming can be said to be "the greatest source of severe and readily avoidable suffering in the world today". Ending this enormous source of suffering is clearly tractable at a collective level. Yet most people still actively contribute to it rather than work against it, despite its solution being technically straightforward.

What is the best way to motivate humanity to prevent suffering?

This is an empirical question. But I would be surprised if setting out abolitionist blueprints turned out to be the single best strategy. Other candidates that seem more promising to me include informing people about horrific examples of suffering, as well as presenting reasoned arguments in favor of prioritizing the prevention of suffering.

To clarify, I am not arguing for any efforts to conserve suffering. The issue here is rather about what we should prioritize with our limited resources. The following analogy may help clarify my view: When animal advocates argue in favor of prioritizing the suffering of farm animals or wild animals rather than, say, the suffering of companion animals, they are not thereby urging us to conserve let alone increase the suffering of companion animals. The argument is rather that our limited resources seem to reduce more suffering if we spend them on these other things, even as we grant that it is a very good thing to reduce the suffering of companion animals.

In terms of how we rank the cost-effectiveness of different causes and interventions (cf. this distribution), I would still consider abolitionist advocacy to be quite beneficial all things considered, and probably significantly better than the vast majority of activities that we could pursue. But I would not quite rank it at the tail-end of the cost-effectiveness distribution, for some of the reasons outlined above.

Why I don't prioritize consciousness research

For altruists trying to reduce suffering, there is much to be said in favor of gaining a better understanding of consciousness. Not only may it lead to therapies that can mitigate suffering in the near term, but it may also help us in our large-scale prioritization efforts. For instance, clarifying which beings can feel pain is important for determining which causes and interventions we should be working on to best reduce suffering.

These points notwithstanding, my own view is that advancing consciousness research is not among the best uses of marginal resources for those seeking to reduce suffering. My aim in this post is to briefly explain why I hold this view.

Reason I: Scientific progress seems less contingent than other important endeavors

Scientific discoveries generally seem quite convergent, so much so that the same discovery is often made independently at roughly the same time (cf. examples of "multiple discovery"). This is not surprising: if we are trying to uncover an underlying truth — as per the standard story of science — we should expect our truth-seeking efforts to eventually converge upon the best explanation, provided that our hypotheses can be tested.

This is not to say that there is no contingency whatsoever in science, which there surely is — after all, the same discovery can be formalized in quite different ways (famous examples include the competing calculus notations of Newton and Leibniz, as well as distinct yet roughly equivalent formalisms of quantum mechanics). But the level of contingency in science still seems considerably lower than the level of contingency found in other domains, such as when it comes to which values people hold or what political frameworks they embrace.

To be clear, it is not that values and political frameworks are purely contingent either, as there is no doubt some level of convergence in these respects as well. Yet the convergence still seems significantly *lower* (and the contingency higher). For example, compare two of the most important events in the early 20th century in these respective domains: the formulation of the general theory of relativity (1915) and the communist revolution in Russia (roughly 1917-1922). While the formulation of the theory of general relativity did involve some contingency, particularly in terms of who and when, it seems extremely likely that the same theory would eventually have been

formulated anyway (after all, many of Einstein's other discoveries were made independently, roughly at the same time).

In comparison, the outcome of the Russian Revolution appears to have been far more contingent, and it seems that greater foreign intervention (as well as other factors) could easily have altered the outcome of the Russian Civil War, and thereby changed the course of history quite substantially.

This greater contingency of values and political systems compared to that of scientific progress suggests that we can generally make a greater counterfactual difference by focusing on the former, other things being equal.

Reason II: Consciousness research seems less neglected than other important endeavors

Besides contingency, it seems that there is a strong neglectedness case in favor of prioritizing the promotion of better values and political frameworks over the advancement of consciousness research.

After all, there are already many academic research centers that focus on consciousness research. By contrast, there is not a single academic research center that focuses primarily on the impartial reduction of suffering (e.g. at the level of values and political frameworks). To be sure, there is a lot of academic work that is *relevant* to the reduction of suffering, yet only a tiny fraction of this work adopts a comprehensive perspective that includes the suffering of all sentient beings across all time; and virtually none of it seeks to clarify optimal priorities relative to that perspective. Such impartial work seems exceedingly rare.

This difference in neglectedness likewise suggests that it is more effective to promote values and political frameworks that aim to reduce the suffering of all sentient beings — as well as to improve our strategic insights into effective suffering reduction — than to push for a better scientific understanding of consciousness.

Objection: The best consciousness research is also neglected

One might object that certain promising approaches to consciousness research (that we could support) are also extremely neglected, even if the larger field of consciousness research is not. Yet granting that this is true, I still think work on values and political frameworks (of the kind alluded to above) will be more neglected overall, considering the greater convergence of science compared to values and politics.

That is, the point regarding scientific convergence suggests that uniquely promising approaches to understanding consciousness are likely to be discovered eventually. Or at least it suggests that these promising approaches will be significantly *less* neglected than will efforts to promote values and political systems centered on effective suffering reduction for all sentient beings.

Reason III: Prioritizing the fundamental bottleneck — the willingness problem

Perhaps the greatest bottleneck to effective suffering reduction is humanity's lack of willingness to this end. While most people may embrace ideals that give significant weight to the reduction of suffering in theory, the reality is that most of us tend to give relatively little priority to the reduction of suffering in terms of our revealed preferences and our willingness to pay for the avoidance of suffering (e.g. in our consumption choices).

In particular, there are various reasons to think that our (un)willingness to reduce suffering is a bigger bottleneck than is our (lack of) understanding of consciousness. For example, if we look at what are arguably the two biggest sources of suffering in the world today — factory farming and wild-animal suffering — it seems that the main bottleneck to human progress on both of these problems is a lack of willingness to reduce suffering, whereas a greater knowledge of consciousness does not appear to be a key bottleneck. After all, most people in the US already report that they believe many insects to be sentient, and a majority likewise agree that farmed animals have roughly the same ability to experience pain as humans. Beliefs about animal sentience per se thus do not appear to be a main bottleneck, as opposed to speciesist attitudes and institutions that disregard non-human suffering.

In general, it seems to me that the willingness problem is best tackled by direct attempts to address it, such as by promoting greater concern for suffering, by reducing the gap between our noble ideals and our often less than noble behavior, and by advancing institutions that reflect impartial concern for suffering to a greater extent. While a better understanding of consciousness may be helpful with respect to the willingness problem, it still seems unlikely to me that consciousness research is among the very best ways to address it.

Reason IV: A better understanding of consciousness might enable deliberate harm

A final reason to prioritize other pursuits over consciousness research is that a better understanding of consciousness comes with significant risks. That is, while a better understanding of consciousness would allow benevolent agents to reduce suffering, it may likewise allow malevolent agents to increase suffering.

This risk is yet another reason why it seems safer and more beneficial to focus directly on the willingness problem and the related problem of keeping malevolent agents out of power — problems that we have by no means found solutions to, and which we are not guaranteed to find solutions to in the future. Indeed, given how serious these problems are, and how little control we have with regard to risks of malevolent individuals in power — especially in autocratic states — it is worth being cautious about developing tools and insights that can potentially increase humanity's ability to cause harm.

Objection: Consciousness research is the best way to address these problems

One might argue that consciousness research is ultimately the best way to address both the willingness problem and the risk of malevolent agents in power, or that it is the best way to solve at least one of those problems. Yet this seems doubtful to me, and like somewhat of a suspicious convergence. Given the vast range of possible interventions we could pursue to address these problems, we should be a priori skeptical of any intervention that we may propose as the best one, particularly when the path to impact is highly indirect.

Objection: We should be optimistic about solving these problems

Another argument in favor of consciousness research might be that we have reason to be optimistic about solving both the willingness problem and the malevolence problem, since the nature of selection pressure is about to change. Thanks to modern technological tools, benevolent agents will soon be able to design the world with greater foresight. We will deliberately choose genes and institutions to ensure that benevolence becomes realized to an ever greater extent, and in effect practically solve both the willingness problem and the malevolence problem.

But this argument seems to overlook two things. First, there is no guarantee that most humans will make actively benevolent choices, even if their choices will not be outright malevolent either. Most people may continue to optimize for things other than impartial benevolence, such as personal status and prestige, and they may continue to show relatively little concern for non-human beings.

Second, and perhaps more worryingly, modern technologies that enable intelligent foresight and deliberation for benevolent agents could be just as empowering for malevolent agents. The arms race between cooperators and exploiters is an ancient one, and I think we have strong reasons to doubt that this arms race will disappear in the next few decades or centuries. On the contrary, I believe we have good grounds to expect this arms race to get intensified, which to my mind is all the more reason to focus directly on reducing the risks posed by malevolent agents, and to promote

norms and institutions that favor cooperation. And again, I am skeptical that consciousness research is among the best ways to achieve these aims, even if it might be beneficial overall.[46]

46 For their comments, I thank Tobias Baumann, Winston Oswald-Drummond, and Jacob Shwartz-Lucas.

The dismal dismissal of suffering-focused views

Ethical views that give a foremost priority to the reduction of suffering are often dismissed out of hand. More than that, it is quite common to see such views discussed in highly uncharitable ways, and to even see them described with pejorative terms.

My aim in this post is to call attention to this phenomenon, as I believe it can distort public discourse and individual thinking about the issue. That is, if certain influential people consistently dismiss certain views without proper argumentation, and in some cases even use disparaging terms to describe such views, then this is likely to bias people's evaluations of these views. After all, most people will likely feel some social pressure not to endorse views that their intellectual peers call "crazy" or "monstrously toxic". (See also what Simon Knutsson writes about social mechanisms that may suppress talk about, and endorsements of, suffering-focused views.)

Many of the examples I present below are not necessarily that significant on their own, but I think the general pattern that I describe is quite problematic. Some of the examples involve derogatory descriptions, while others involve strawman arguments and uncharitable rejections of suffering-focused views that fail to engage with the most basic arguments in favor of such views.

My overall recommendation is simply to meet suffering-focused views with charitable arguments rather than with strawman argumentation or insults — i.e. to live up to the standards that are commonly accepted in other realms of intellectual discourse.

"Crazy" and "transparently silly" views

In his essay "Why I'm Not a Negative Utilitarian" (2013), Toby Ord writes that "you would have to be crazy" to choose a world with beings who experience unproblematic states over a world with beings who experience pure happiness (strict negative utilitarianism would be indifferent between the two, and according to some versions of negative utilitarianism, unproblematic mental states and pure happiness are the same thing, cf. Sherman, 2017; Knutsson, 2022).

Ord also writes that the view that happiness does not contribute to a person's wellbeing independently of its effects on reducing problematic states is a "crazy view", without engaging with any of the arguments that have been made in favor of the class of views that he is thereby dismissing — i.e. views according to which wellbeing consists in the absence of problematic states or frustrated desires (see e.g. Schopenhauer, 1819; 1851; Fehige, 1998; O'Keefe, 2009, ch. 12).

These may not seem like particularly problematic claims, yet I believe that Ord would consider it poor form if similar claims were made about his preferred view — for example, if someone claimed that "you would have to be crazy to choose to create arbitrarily large amounts of extreme suffering in order to create a 'sufficient' amount of pleasure" (cf. the Very Repugnant Conclusion; Creating Hell to Please the Blissful; and Intense Bliss with Hellish Cessation).

Similarly, Rob Bensinger writes that negative utilitarianism is "transparently false/silly". Bensinger provides a brief justification for his claim that I myself and others find unconvincing, and it is in any case not a justification that warrants calling negative utilitarianism "transparently false/silly".

Lazari-Radek and Singer's cursory rejection

In their book *The Point of View of the Universe*, Lazari-Radek and Singer seek to defend the classical utilitarian view of Henry Sigdwick. It would be natural, in this context, to provide an elaborate discussion of the moral symmetry between happiness and suffering that is entailed by classical utilitarianism — after all, such a moral symmetry has been rejected by various philosophers in a variety of ways, and it is arguably one of the most controversial features of classical utilitarianism (cf. Mayerfeld, 1996, p. 335).

Yet Lazari-Radek and Singer barely broach the issue at all. The only thing that comes close is a single page worth of commentary on the views of David Benatar, which unfortunately amounts to a misrepresentation of Benatar's views. Lazari-Radek and Singer claim that Benatar argues that "to have a desire for something is to be in a negative state" (p. 362). To my knowledge, this is not a claim that Benatar defends, and the claim is at any rate not critical to the main procreative asymmetry that he argues for (Benatar, 2006, ch. 2).

Lazari-Radek and Singer briefly rebut the claim about desires that they (I suspect wrongly) attribute to Benatar, by which they fail to address Benatar's core views in any meaningful way. They then proceed to write the following, which as far as I can tell is the closest they get to a defense of a moral symmetry between happiness and suffering in their entire book: "for people who are able to satisfy the basic necessities of life and who are not suffering from depression or chronic pain, life can reasonably be judged positively" (pp. 362-363).

This is, of course, not much of a defense of a moral symmetry. First of all, no arguments are provided in defense of the claim that such lives "can reasonably be judged positively" (a claim that one can reasonably dispute). Second, even if we grant that certain lives "can be judged positively" (in terms of the intrinsic value of their contents), it still does not follow that such lives that are "judged positively" can also morally outweigh the most horrific lives. This is an all-important issue for the classical utilitarian to address, and yet Lazari-Radek and Singer proceed as though their

claim that "life can reasonably be judged positively" also applies to the world as a whole, even when we factor in all of its most horrific lives. Put briefly, Lazari-Radek and Singer's cursory rejection of asymmetric and suffering-focused views is highly unsatisfactory.

(In a vein similar to the dismissive remarks covered in the previous section, Lazari-Radek and Singer also later write that "any sane person will agree" that a scenario in which 100 percent of humanity dies is worse than a scenario in which 99 percent of humanity dies, cf. p. 375. Regardless of the plausibility of that claim — which one might agree with even from a purely suffering-focused perspective — it is bad form to imply that people are not sane if they disagree with it, not least since the latter scenario could well involve far more suffering overall. Likewise, in a response to a question on Reddit, Singer dismisses negative utilitarianism as "hopeless" without providing any reasons as to why.)

"Arguably too nihilistic and divorced from humane values to be worth taking seriously"

The website utilitarianism.net is co-authored by William MacAskill, Richard Yetter Chappell, and Darius Meissner. The aim of the website is to provide "a textbook introduction to utilitarianism at the undergraduate level", and it is endorsed by Peter Singer (among others), who blurbs it as "the place to go for clear, full and fair accounts of what utilitarianism is, the arguments for it, the main objections to it, special issues like population ethics, and what living as a utilitarian involves."

Yet the discussion found on the website is sorely lacking when it comes to fundamental questions and objections concerning the relative importance of suffering versus happiness. In particular, like Lazari-Radek and Singer's *Point of View of the Universe*, the website contains no discussion of the moral symmetry between suffering and happiness that is entailed by classical utilitarianism, despite it being among the most disputed features of that view (see e.g. Popper, 1945; Mayerfeld, 1996; 1999; Wolf, 1996; 1997; 2004; O'Keefe, 2009; Knutsson, 2016; Mathison, 2018; Vinding, 2020).

Similarly, the discussion of population ethics found on the website is extremely one-sided and uncharitable in its discussion of suffering-focused and asymmetric views in population ethics, especially for a text that is supposed to serve as an introductory textbook.

For instance, they write the following in a critique of the Asymmetry in population ethics (the Asymmetry is roughly the idea that it is bad to bring miserable lives into the world but not good to bring happy lives into the world):

> But this brings us to a deeper problem with the procreative asymmetry, which is that it has trouble accounting for the idea that *we should be positively glad that the world (with all its worthwhile lives) exists*.

There is much to take issue with in this sentence. First, it presents the idea that "we should be positively glad that the world exists" as though it is an obvious and supremely plausible idea; yet it is by no means obvious, and it has been questioned by many philosophers. A truly "full and fair" introductory textbook would have included references to such counter-perspectives. Indeed, the authors of utilitarianism.net call it a "perverse conclusion" that an empty world would be better than a populated one, without mentioning any of the sources that have defended that "perverse conclusion", and without engaging with the arguments that have been made in its favor (e.g. Schopenhauer, 1819; 1851; Benatar, 1997; 2006; Fehige, 1998; Breyer, 2015; Gloor, 2017; St. Jules, 2019; Frick, 2020; Ajantaival, 2021/2022). Again, this falls short of what one would expect from a "full and fair" introductory textbook.

Second, the quote above may be critiqued for bringing in confounding intuitions, such as intuitions about the value of the world as a whole, which is in many ways a different issue from the question of whether it can be good to add new beings to the world for the sake of these beings themselves.

Third, the notion of "worthwhile lives" is not necessarily inconsistent with a procreative asymmetry, since lives may be deemed worthwhile in the sense that their continuation is preferable even if their creation is not (cf. Benatar, 1997; 2006; Fehige, 1998; St. Jules, 2019; Frick, 2020). Additionally, one can think that a life is worthwhile — both in terms of its continuation *and* creation — because it has beneficial effects for others, even if it can never be better for the created individual themself that they come into existence.

The authors go on to write:

> when thinking about what makes some possible universe *good*, the most obvious answer is that it contains a predominance of awesome, flourishing lives. How could that *not* be better than a barren rock? Any view that denies this verdict is arguably too nihilistic and divorced from humane values to be worth taking seriously.

This quote effectively dismisses all of the views cited above — the views of Schopenhauer, Fehige, Benatar, and Frick, as well as the Nirodha View in the Pali Buddhist tradition — in one fell swoop by claiming that they are "arguably too nihilistic and divorced from humane values to be worth taking seriously". That is, to put it briefly, a lazy treatment that again falls short of the minimal standards of a fair introductory textbook.

After all, classical utilitarians would probably also object if a textbook introduction were to effectively dismiss classical utilitarianism (and similar views) with the one-line claim that "views that allow the creation of lives full of extreme suffering in order to create pleasure for others are arguably too divorced from humane values to be worth taking seriously." Yet the dismissal is just as unhelpful and uncharitable when made in the other direction.

Finally, the authors also omit any mention of the Very Repugnant Conclusion, although one of the co-authors, William MacAskill, has stated that he considers it the strongest objection against his favored version of utilitarianism. It is arguably bad form to omit any discussion — or even a mention — of what one considers the strongest objection against one's favored view, especially if one is trying to write a fair and balanced introductory textbook that features that view prominently.

"Anti-natalism is neurotic self-hatred"

Psychologist Geoffrey Miller has given several talks about effective altruism, including one at EA Global, and he has also taught a full university course on the psychology of effective altruism. At the time of writing, Miller has more than 120,000 followers on Twitter, which makes him one of the most widely followed people associated with effective altruism, with more followers than Peter Singer.

Having such a large audience arguably raises one's responsibility to communicate in an intellectually honest and charitable manner. Yet Miller has repeatedly misrepresented the views of David Benatar and written highly uncharitable statements about antinatalism and negative utilitarianism, without seriously engaging with the arguments made in favor of these views.

For example, Miller has written on Twitter that "anti-natalism is neurotic self-hatred", and he has on several occasions falsely implied that David Benatar is a negative utilitarian, such as when he writes that "[Benatar's] negative utilitarianism assumes that only suffering counts, & pleasure can never offset it"; or when he writes that "Benatar's view boils down to the claim that all the joy, beauty, & love in the world can't offset even a drop of suffering in any organism anywhere. It's a monstrously toxic & nihilistic philosophy."

Yet the views that Miller attributes to Benatar are not views that Benatar in fact defends, and anyone familiar with Benatar's position knows that he does not think that "only suffering counts" (cf. his rejection of the Epicurean view of death, Benatar, 2006, ch. 7).

Miller also betrays a failure to understand Benatar's view when he writes:

> The asymmetry thesis is empirically false for humans. Almost all people report net positive subjective well-being in hundreds of studies around the world. Benatar is

basically patronizing everyone, saying ‘All you guys are wrong; you’re actually miserable’.

First, Benatar discusses various reasons as to why self-assessments of one’s quality of life may be unreliable (Benatar, 2006, pp. 64-69; see also Vinding, 2018). This is not fundamentally different from, say, evolutionary psychologists who argue that people’s self-reported motives may be wrong. Second, and more importantly, the main asymmetry that Benatar defends is not an empirical one, but rather an evaluative asymmetry between the presence and absence of goods versus the presence and absence of bads (Benatar, 2006, ch. 2). This evaluative asymmetry is not addressed by Miller’s claim above.

One might object that Miller’s statements have all been made on Twitter, and that tweets should generally be held to a lower standard than other forms of writing. Yet even if we grant that tweets should be held to a lower standard, we should still be clear that Miller blatantly misrepresents Benatar’s views, which is bad form on any platform and by any standard.

Moreover, one could argue that tweets should in some sense be held to a *higher* standard, since tweets are likely to be seen by more people compared to many other forms of writing (such as the average journal article), and perhaps also by readers who are less inclined to verify scholarly claims made by a university professor (compared to readers of other media).

More examples

Additional examples of uncharitable dismissals of suffering-focused views include statements from:

- Writer and EA Global speaker Riva-Melissa Tez, who wrote that “anti-natalism and negative utilitarianism is true ‘hate speech’”.
- YouTuber Robert Miles (>100k subscribers), who wrote: “Looks like it’s time for another round of ‘Principled Negative Utilitarianism or Undiagnosed Major Depressive Disorder?’” (See also here.)
- Daniel Faggella, who wrote: “If I didn’t know so many negative utilitarians who I liked as people, I’d call it a position of literal cowardice – even vice.” (The original post was even stronger in its tone: “If I didn’t know and respect so many negative utilitarians, I would openly call it a vice, and a position of childish, seething cowardice.”)
    - I find the remark about cowardice to be quite strange, as it seems to me that it takes a lot of courage to face up to the horror of suffering, and to set out to alleviate suffering with determination. And socially, too, it can take a lot of courage to

embrace strongly suffering-focused views in a social environment that often ridicules such views, and which often insinuates that there is something wrong with the adherents of these views.

- R. N. Smart, who wrote that negative utilitarianism allows "certain absurd and even wicked moral judgments", without providing any arguments as to whether competing moral views imply less "absurd or wicked" moral judgments, and without mentioning that classical utilitarianism — which Smart seems to express greater approval toward — has similar and arguably worse theoretical implications (cf. Knutsson, 2021; Ajantaival, 2022).

The following anecdotal example illustrates how uncharitable remarks can influence people's motivations and make people feel unwelcome in certain communities: An acquaintance of mine who took part in an EA intro fellowship heard a fellow participant dismiss antinatalism quite uncharitably, saying something along the lines of "antinatalism is like high school atheism, but edgier". My acquaintance thought that antinatalism is a plausible view, and the remark left them feeling unwelcome and discouraged from engaging further with effective altruism.

Conclusion

To be clear, my point is by no means that people should refrain from criticizing suffering-focused views, even in strong terms. My recommendation is simply that critics should strive to be even-handed, and to not misrepresent or unfairly malign views with which they disagree.

If we are trying to think straight about ethics, we should be keen not to let uncharitable claims and social pressures distort our thinking, especially since these factors tend to influence our views in hidden ways. After all, few people consciously think — let alone say — that social pressure exerts a strong influence on their views. Yet it is likely a potent factor all the same.

Beware frictions from altruistic value differences

I believe value differences pose some underappreciated challenges in large-scale altruistic efforts. My aim in this post is to outline what I see as the main such challenges, and to present a few psychological reasons as to why we should expect these challenges to be significant and difficult to overcome.[47]

To clarify, my aim in this post is not to make a case against value differences per se, much less a case against vigorous debate over values (I believe that such debate is healthy and desirable). Instead, my aim is to highlight some of the challenges and pitfalls that are *associated with* value differences, in the hope that we can better mitigate these pitfalls. After all, value differences are sure to persist among people who are trying to help others, and hence a critical issue is how well — or how poorly — we are going to handle these differences.

Examples of challenges posed by value differences among altruists

A key challenge posed by value differences, in my view, is that they can make us prone to tribal or otherwise antagonistic dynamics that are suboptimal by the lights of our own moral values. Such values-related frictions may in turn lead to the following pitfalls and failure modes:

- Failing to achieve moral aims that are already widely shared, such as avoiding worst-case outcomes (cf. "Common ground for longtermists").
- Failing to make mutually beneficial moral trades and compromises when possible (in ways that do not introduce problematic behavior such as dishonesty or censorship).
- Failing to update on arguments, whether they be empirical or values-related, because the arguments are made by those who, to our minds, seem like they belong to the "other side".[48]
- Some people committing harmful acts out of spite or primitive tribal instincts. (The sections below give some sense as to why this might happen.)[49]

47 By "value differences", I mean differences in underlying axiological and moral views relating to altruism. I don't have in mind anything that involves, say, hateful values or overt failures of moral character. Such moral failures are obviously worth being acutely aware of, too, but mostly for other reasons than the ones I explore here.

48 By analogy to how discriminatory hiring practices can cause economic inefficiencies, it seems plausible that values- and coalition-driven antagonisms can likewise cause "epistemic inefficiencies" (cf. Simler's "Crony beliefs").

49 That is, not only can values-driven antagonisms prevent us from capitalizing on potential gains, but they may in the worst case lead some people to actively sabotage and undermine just about everyone's moral aims, including the reflective moral aims of the emotion-driven actors themselves.

Of course, some of the failure modes listed above can have other causes beyond values- and coalition-related frictions. Yet poorly handled such frictions are probably still a key risk factor for these failure modes.

Reasons to expect values-related frictions to be significant

The following are some reasons to expect values-related frictions to be both common and quite difficult to handle by default.

Harmful actions based on different moral beliefs may be judged more harshly than intentional harm

One set of findings that seem relevant come from a 2016 anthropological study that examined the moral judgments of people across ten different cultures, eight of which were traditional small-scale societies (Barrett et al., 2016).

The study specifically asked people how they would evaluate a harmful act in light of a range of potentially extenuating circumstances, such as different moral beliefs, a mistake of fact, or self-defense. (The particular moral belief used in the study was that "striking a weak person to toughen him up is praiseworthy".)

While there was significant variation in people's moral judgments across cultures, there was nevertheless unanimous agreement that committing a harmful act based on different moral beliefs was *not* an extenuating circumstance. Indeed, on average across cultures, committing a harmful act based on different moral beliefs was considered worse than was committing the harmful act intentionally (see Barrett et al., 2016, fig. 5).

It is unclear whether this pattern in moral judgment necessarily applies to all or even most kinds of acts inspired by different moral beliefs. Yet these results still tentatively suggest that we may be inclined to see value differences as a uniquely aggravating factor in our moral judgments of people's actions — as something that tends to inspire harsher judgments rather than understanding.

Hot cognition about values-related beliefs, alliances, and opponents

Another relevant finding is that our minds appear to reflexively process moral and political groups and issues in ways that are strongly emotionally charged — an instance of "hot cognition". Specifically, we appear to affectively process our own groups and beliefs in a favorable light while similarly processing the "outgroup" and their beliefs in an unfavorable light. And what is striking about this affectively charged processing is that it appears to be swift and automatic, occurring prior

to conscious thought, which suggests that we are mostly unaware that it happens (Lodge & Taber, 2005; see also Kunda, 1990; Haidt, 2001).

These findings give us reason to expect that our reflexive processing of those who hold different altruistic values will tend to be affectively charged in ways that we are not aware of, and in ways that are not so easily changed (cf. Lodge & Taber, 2005, p. 476).

Coalitional instincts

A related reason to expect values-driven tensions to be significant and difficult to avoid is that the human mind plausibly has strong coalitional instincts, i.e. instincts for carving the world into, and smoothly navigating among, competing coalitions (Tooby & Cosmides, 2010).[50]

As John Tooby notes, these instincts may dispose us to blindly flatter and protect our own groups while misrepresenting and attacking other groups and coalitions. He likewise suggests that our coalitional instincts may push our public discourse less toward substance and more toward displaying loyalty to our own groups (see also Hannon, 2021).

In general, it seems that "team victory" is a strong yet often hidden motive in human behavior. And these coalitional instincts and "team victory" motives arguably further highlight the psychological challenges posed by value differences, not least since value differences often serve as the defining features of contrasting coalitions.[51]

Concrete suggestions for mitigating the risks of values-related frictions

Below are a few suggestions for how one might address the challenges and risks associated with values-related frictions. More suggestions are welcome.[52]

Acknowledging good-faith intentions and attempts to help others

It seems helpful to remind ourselves that altruists who have different values from ourselves are generally acting in good faith, and are trying to help others based on what they sincerely believe to be the best or most plausible views.

---

50 This point is closely related to the previous point, in that our hot cognition often reflects or manifests our coalitional instincts. For what it's worth, I believe that the concepts of coalitional instincts and (coalition-driven) hot cognition are two of the most powerful concepts for understanding human behavior in the realms of politics and morality.

51 Of course, values are by no means the only such coalition-defining feature. Other examples may include shared geographical location, long-term familiarity (e.g. with certain individuals or groups), and empirical beliefs. Indeed, it is my impression that empirical beliefs can be about as intense a source of coalitional identity and frictions as can value differences, even when we primarily hold the beliefs in question for epistemic rather than signaling reasons.

52 To be clear, I am not denying that there are also significant benefits to adversarial debate and discussion. But it still seems reasonable to make an effort to maximize the benefits while minimizing the risks.

Keeping in mind shared goals and potential gains from compromise

Another helpful strategy may be to keep in mind the shared goals and the important points of agreement that we have with our fellow altruists — e.g. a strong emphasis on impartiality, a strong focus on sentient welfare, a wide agreement on the importance of avoiding the very worst future outcomes, etc.

Likewise, it might be helpful to think of the positive-sum gains that people with different values may achieve by cooperating. After all, contrary to what our intuitions might suggest, it is quite conceivable that some of our greatest counterfactual gains can be found in the realm of cooperation with agents who hold different values from ourselves — e.g. by steering clear of "fights" and by instead collaborating to expand our Pareto frontier (cf. Hanson on "Expand vs Fight"). It would be tragic to lose out on such gains due to unwittingly navigating more by our coalitional instincts and identities than by impartial impact.

Making an effort to become aware of, and to actively reduce, the tendency to engage in reflexive ingroup liking and promotion

It is to be expected that we are prone to ingroup liking and ingroup promotion to a somewhat excessive degree (relative to what our impartial values would recommend). In that case, it may be helpful to become more aware of these reflexive tendencies, and to try to reduce them through deliberate "system-2" reasoning that is cautiously skeptical of our most immediate coalitional drives and intuitions, in effect adding a cooling element to our hot cognition.

Validating the difficulty of the situation

Finally, it may be helpful to take a step back and to validate how eminently understandable it is that strong reactions can emerge in the context of altruistic value differences.

After all, beyond the psychological reasons reviewed above, it is worth remembering that there is often a lot of identity on the line when value differences come up among altruists. Indeed, it is not only identity that is on the line, but also individual and collective priorities, plans, visions, and so on.

These are all quite foundational elements that touch virtually every level of our cognitive and emotional processing. And when all these elements effectively become condensed into a single conversation with a person who appears to have significant disagreements with us on just about all of these consequential issues, *and* our minds are under the influence of a fair dose of coalition-driven hot cognition, then no wonder that things start to feel a little tense and challenging.

Validating the full magnitude of this challenge might help lower the temperature, and in turn open the door to more fruitful engagements and collaborations going forward.

Research vs. non-research work to improve the world: In defense of more research and reflection

When trying to improve the world, we can either pursue direct interventions, such as directly helping beings in need and doing activism on their behalf, or we can pursue research on *how* we can best improve the world, as well as on *what* improving the world even means in the first place.

Of course, the distinction between direct work and research is not a sharp one. We can, after all, learn a lot about the "how" question by pursuing direct interventions, testing out what works and what does not. Conversely, research publications can effectively function as activism, and may thereby help bring about certain outcomes quite directly, even when such publications do not deliberately try to do either.

But despite these complications, we can still meaningfully distinguish more or less research-oriented efforts to improve the world. My aim here is to defend more research-oriented efforts, and to highlight certain factors that may lead us to underinvest in research and reflection. (Note that I here use the term "research" to cover more than just original research, as it also covers efforts to learn about existing research.)

Some examples

Perhaps the best way to give a sense of what I am talking about is by providing a few examples.

I. Cause Prioritization

Say our aim is to reduce suffering. Which concrete aims should we then pursue? Maybe our first inclination is to work to reduce human poverty. But when confronted with the horrors of factory farming, and the much larger number of non-human animals compared to humans, we may conclude that factory farming seems the more pressing issue. However, having turned our gaze to non-human animals, we may soon realize that the scale of factory farming is small compared to the scale of wild-animal suffering, which might in turn be small compared to the potentially astronomical scale of future moral catastrophes.

With so many possible causes one could pursue, it is likely suboptimal to settle on the first one that comes to mind, or to settle on any one of them without having made a significant effort considering where one can make the greatest difference.

II. Effective Interventions

Next, say we have settled on a specific cause, such as ending factory farming. Given this aim, there is a vast range of direct interventions one could pursue, including various forms of activism, lobbying to influence legislation, or working to develop novel foods that can outcompete animal products. Yet it is likely suboptimal to pursue any of these particular interventions without first trying to figure out which of them have the best expected impact. After all, different interventions may differ greatly in terms of their cost-effectiveness, which suggests that it is reasonable to make significant investments into figuring out which interventions are best, rather than to rush into action mode (although the drive to do the latter is understandable and intuitive, given the urgency of the problem).

III. Core Values

Most fundamentally, there is the question of what matters and what is most worth prioritizing at the level of core values. Our values ultimately determine our priorities, which renders clarification of our values a uniquely important and foundational step in any systematic endeavor to improve the world.

For example, is our aim to maximize a net sum of "happiness minus suffering", or is our aim chiefly to minimize extreme suffering? While there is significant common ground between these respective aims, there are also significant divergences between them, which can matter greatly for our priorities. The first view implies that it would be a net benefit to create a future that contains vast amounts of extreme suffering as long as that future contains a lot of happiness, while the other view would recommend the path of least extreme suffering.

In the absence of serious reflection on our values, there is a high risk that our efforts to improve the world will not only be suboptimal, but even positively harmful relative to the aims that we would endorse most strongly upon reflection. Yet efforts to clarify values are nonetheless extremely neglected — and often completely absent — in endeavors to improve the world.

The steelman case for "doing"

Before making a case for a greater focus on research, it is worth outlining some of the strongest reasons in favor of direct action (e.g. directly helping other beings and doing activism on their behalf).

We can learn a lot by acting

- The pursuit of direct interventions is a great way to learn important lessons that may be difficult to learn by doing pure research or reflection.
- In particular, direct action may give us practical insights that are often more in touch with reality than are the purely theoretical notions that we might come up with in intellectual isolation. And practical insights and skills often cannot be compensated for by purely intellectual insights.
- Direct action often has clearer feedback loops, and may therefore provide a good opportunity to both develop and display useful skills.

Direct action can motivate people to keep working to improve the world

- Research and reflection can be difficult, and it is often hard to tell whether one has made significant progress. In contrast, direct action may offer a clearer indication that one is really doing something to improve the world, and it can be easier to see when one is making progress (e.g. whether people altered their behavior in response to a given intervention, or whether a certain piece of legislation changed or not).

There are obvious problems in the world that are clearly worth addressing

- For example, we do not need to do more research to know that factory farming is bad, and it seems reasonable to think that evidence-based interventions that significantly reduce the number of beings who suffer on factory farms will be net beneficial.
- Likewise, it is probably beneficial to build a healthy movement of people who aim to help others in effective ways, and who reflect on and discuss what "helping others" ideally entails.

Certain biases plausibly prevent us from pursuing direct action

- It seems likely that we have a passivity bias of sorts. After all, it is often convenient to stay in one's intellectual armchair rather than to get one's hands dirty with direct work that may fall outside of one's comfort zone, such as doing street advocacy or running a political campaign.
- There might also be an omission bias at work, whereby we judge an omission to do direct work that prevents harm less harshly than an equivalent commission of harm.

The case for (more) research

I endorse all the arguments outlined above in favor of "doing". In particular, I think they are good arguments in favor of maintaining a strong element of direct action in our efforts to improve the world. Yet they are less compelling when it comes to establishing the stronger claim that we should focus *more* on direct action (on the current margin), or that direct action should represent the majority of our altruistic efforts at this point in time. I do not think any of those claims follow from the arguments above.

In general, it seems to me that altruistic endeavors tend to focus far too strongly on direct action while focusing far too little on research. This is hardly a controversial claim, at least not among aspiring effective altruists, who often point out that research on cause prioritization and on the cost-effectiveness of different interventions is important and neglected. Yet it seems to me that even effective altruists tend to underinvest in research, and to jump the gun when it comes to cause selection and direct action, and especially when it comes to the values that they choose to steer by.

A helpful starting point might be to sketch out some responses to the arguments outlined in the previous section, to note why those arguments need not undermine a case for more research.

We can learn a lot by acting — but we are arguably most limited by research insights

The fact that we can learn a lot by acting, and that practical insights and skills often cannot be substituted by pure conceptual knowledge, does not rule out that our potential for beneficial impact might generally be most bottlenecked by conceptual insights.

In particular, clarifying our core values and exploring the best causes and interventions arguably represent the most foundational steps in our endeavors to improve the world, suggesting that they should — at least at the earliest stages of our altruistic endeavors — be given primary importance relative to direct action (even as direct action and the development of practical skills also deserve significant priority, perhaps even more than 20 percent of the collective resources we spend at this point in time).

The case for prioritizing direct action would be more compelling if we had a lot of research that delivered clear recommendations for direct action. But I think there is generally a glaring shortage of such research. Moreover, research on cause prioritization often reveals plausible ways in which direct altruistic actions that seem good at first sight may actually be harmful. Such potential downsides of seemingly good actions constitute a strong and neglected reason to prioritize research more — not to get perpetually stuck in research, but to at least map out the main considerations for and against various actions.

To be more specific, it seems to me that the expected value of our actions can change a lot depending on how deep our network of crucial considerations goes, so much so that adding an extra layer of crucial considerations can flip the expected value of our actions. Inconvenient as it may be, this means that our views on what constitutes the best direct actions have a high risk of being unreliable as long as we have not explored crucial considerations in depth. (Such a risk always exists, of course, yet it seems that it can at least be markedly reduced, and that our estimates can become significantly better informed even with relatively modest research efforts.)

At the level of an individual altruist's career, it seems warranted to spend at least one year reading about and reflecting on fundamental values, one year learning about the most important cause areas, and one year learning about optimal interventions within those cause areas (ideally in that order, although one may fruitfully explore them in parallel to some extent; and such a full year's worth of full-time exploration could, of course, be conducted over several years). In an altruistic career spanning 40 years, this would still amount to less than ten percent of one's work time focused on such basic exploration, and less than three percent focused on exploring values in particular.

A similar argument can be made at a collective level: if we are aiming to have a beneficial influence on the long-term future — say, the next million years — it seems warranted to spend at least a few years focused primarily on what a beneficial influence would entail (i.e. clarifying our views on normative ethics), as well as researching *how* we can best influence the long-term future before we proceed to spend most of our resources on direct action. And it may be even better to try to encourage more people to pursue such research, ideally creating an entire research project in which a large number of people collaborate to address these questions.

Thus, even if it is ideal to mostly focus on direct action over the entire span of humanity's future, it seems plausible that we should focus most strongly on advancing research at this point, where relatively little research has been done, and where the explore-exploit tradeoff is likely to favor exploration quite strongly.

Objections: What about "long reflection" and the division of labor?

An objection to this line of reasoning is that heavy investment into reflection is premature, and that our main priority at this point should instead be to secure a condition of "long reflection" — a long period of time in which humanity focuses on reflection rather than action.

Yet this argument is problematic for a number of reasons. First, there are strong reasons to doubt that a condition of long reflection is feasible or even desirable, given that it would seem to require strong limits to voluntary actions that diverge from the ideal of reflection.

To think that we can choose to create a condition of long reflection may be an instance of the illusion of control. Human civilization is likely to develop according to its immediate interests, and seems unlikely to ever be steered via a common process of reflection. And even if we were to secure a condition of long reflection, there is no guarantee that humanity would ultimately be able to reach a sufficient level of agreement regarding the right path forward — after all, it is conceivable that a long reflection could go awfully wrong, and that bad values could win out due to poor execution or malevolent agents hijacking the process.

The limited feasibility of a long reflection suggests that there is no substitute for reflecting now. Failing to clarify and act on our values from this point onward carries a serious risk of pursuing a suboptimal path that we may not be able to reverse later. The resources we spend pursuing a long reflection (which is unlikely to ever occur) are resources not spent on addressing issues that might be more important and more time-sensitive, such as steering away from worst-case outcomes.

Another objection might be that there is a division of labor case favoring that only some people focus on research, while others, perhaps even most, should focus comparatively little on research. Yet while it seems trivially true that some people should focus more on research than others, this is not necessarily much of a reason against devoting more of our collective attention toward research (on the current margin), nor a reason against each altruist making a significant effort to read up on existing research.

After all, even if only a limited number of altruists should focus primarily on research, it still seems necessary that those who aim to put cutting-edge research into practice also spend time reading that research, which requires a considerable time investment. Indeed, even when one chooses to mostly defer to the judgments of other people, one will still need to make an effort to evaluate which people are most worth deferring to on different issues, followed by an effort to adequately understand what those people's views and findings entail.

This point also applies to research on values in particular. That is, even if one prioritizes direct action over research on fundamental values, it still seems necessary to spend a significant amount of time reading up on other people's work on fundamental values if one is to make a qualified judgment regarding which values one will attempt to steer by.

The division of altruistic labor is thus consistent with the recommendation that every dedicated altruist should spend at least a full year reading about and reflecting on fundamental values (just as the division of "ordinary" labor is consistent with everyone spending a certain amount of time on basic education). And one can further argue that the division of altruistic labor, and specialized

work on fundamental values in particular, is only fully utilized if most people spend a decent amount of time reading up on and making use of the insights provided by others.

Direct action can motivate people — but so can (the importance of) research

While research work is often challenging and difficult to be motivated to pursue, it is probably a mistake to view our motivation to do research as something that is fixed. There are likely many ways to increase our motivation to pursue research, not least by strongly internalizing the (highly counterintuitive) importance of research.

Moreover, the motivating force provided by direct action might be largely maintained as long as one includes a strong component of direct action in one's altruistic work (by devoting, say, 25 percent of one's resources toward direct action).

In any case, reduced individual motivation to pursue research seems unlikely to be a strong reason against devoting a greater priority to research at the level of collective resources and priorities (even if it might play a significant role in many individual cases). This is partly because the average motivation to pursue these respective endeavors seems unlikely to differ greatly — after all, many people will be more motivated to pursue research over direct action — and partly because urgent necessities are worth prioritizing and paying for even if they happen to be less than highly motivating.

By analogy, the cleaning of public toilets is also worth prioritizing and paying for, even if it may not be the most motivating pursuit for those who do it, and the same point arguably applies even more strongly in the case of the most important tasks necessary for achieving altruistic aims such as reducing extreme suffering. Moreover, the fact that altruistic research may be unusually taxing on our motivation (e.g. due to a feeling of "analysis paralysis") is a reason to think that such taxing research is generally neglected and hence worth pursuing on the margin.

Finally, to the extent one finds direct action more motivating than research, this might constitute a bias in one's prioritization efforts, even if it represents a relevant data point about one's personal fit and comparative advantage. And the same point applies in the opposite direction: to the extent that one finds research more motivating, this might make one more biased against the importance of direct action. While personal motivation is an important factor to consider, it is still worth being mindful of the tendency to overprioritize that which we consider fun and inspiring at the expense of that which is most important in impartial terms.

There are obvious problems in the world that are clearly worth addressing — but research is needed to best prioritize and address them

Knowing that there are serious problems in the world, as well as interventions that reduce those problems, does not in itself inform us about which problems are *most* pressing or which interventions are *most* effective at addressing them. Both of these aspects — roughly, cause prioritization and estimating the effectiveness of interventions — seem best advanced by research.

A similar point applies to our core values: we cannot meaningfully pursue cause prioritization and evaluations of interventions without first having a reasonably clear view of what matters, and what would constitute a better or worse world. And clarifying our values is arguably also best done through further research rather than through direct action, even as the latter may be helpful as well.

Certain biases plausibly prevent us from pursuing direct action — but there are also biases pushing us toward too much or premature action

The putative "passivity bias" outlined above has a counterpart in the "action bias", also known as "bias for action" — a tendency toward action even when action makes no difference or is positively harmful. A potential reason behind the action bias relates to signaling: actively doing something provides a clear signal that we are at least making an effort, and hence that we care (even if the effect might ultimately be harmful). By comparison, doing nothing might be interpreted as a sign that we do not care.

There might also be individual psychological benefits explaining the action bias, such as the satisfaction of feeling that one is "really doing something", as well as a greater feeling of being in control. In contrast, pursuing research on difficult questions can feel unsatisfying, since progress may be relatively slow, and one may not intuitively feel like one is "really doing something", even if learning additional research insights is in fact the best thing one can do.

Political philosopher Michael Huemer similarly argues that there is a harmful tendency toward too much action in politics. Since most people are uninformed about politics, Huemer argues that most people ought to be passive in politics, as there is otherwise a high risk that they will make things worse through ignorant choices.

Whatever one thinks of the merits of Huemer's argument in the political context, I think one should not be too quick to dismiss a similar argument when it comes to improving the long-term future — especially considering that action bias seems to be greater when we face increased uncertainty. At the very least, it seems worth endorsing a modified version of the argument that says that we should not be eager to act before we have considered our options carefully.

Furthermore, the fact that we evolved in a condition that was highly action-oriented rather than reflection-oriented, and in which action generally had far more value for our genetic fitness than did systematic research (indeed, the latter was hardly even possible), likewise suggests that we may be inclined to underemphasize research relative to how important it is for optimal impact from an impartial perspective.

This also seems true when it comes to our altruistic drives and behaviors in particular, where we have strong inclinations toward pursuing publicly visible actions that make us appear good and helpful (Hanson, 2015; Simler & Hanson, 2018, ch. 12). In contrast, we seem to have much less of an inclination toward reflecting on our values. Indeed, it seems plausible that we generally have an inclination *against* questioning our instinctive aims and drives — including our drive to signal altruistic intentions with highly visible actions — as well as an inclination against questioning the values held by our peers. After all, such questioning would likely have been evolutionarily costly in the past, and may still feel socially costly today.

Moreover, it is very unnatural for us to be as agnostic and open-minded as we should ideally be in the face of the massive uncertainty associated with endeavors that seek to have the best impact for all sentient beings (Vinding, 2020, sec. 9.1-9.2). This suggests that we may tend to be overconfident about — and too quick to conclude — that some particular direct action happens to be the optimal path for helping others.

Lastly, while some kind of omission bias plausibly causes us to discount the value of making an active effort to help others, it is not clear whether this bias counts more strongly against direct action than against research efforts aimed at helping others, since omission bias likely works against both types of action (relative to doing nothing). In fact, the omission bias might count more strongly against research, since a failure to do important research may feel like less of a harmful inaction than does a failure to pursue direct actions, whose connection to addressing urgent needs is usually much clearer.

The Big Neglected Question

There is one question that I consider particularly neglected among aspiring altruists — as though it occupies a uniquely impenetrable blindspot. I am tempted to call it “The Big Neglected Question”.

The question, in short, is whether anything can morally outweigh or compensate for extreme suffering. Our answer to this question has profound implications for our priorities. And yet astonishingly few people seem to seriously ponder it, even among dedicated altruists. In my view, reflecting on this question is among the first, most critical steps in any systematic endeavor to improve the world. (I suspect that a key reason this question tends to be shunned is that it seems too

dark, and because people may intuitively feel that it fundamentally questions all positive and meaning-giving aspects of life — although it arguably does not, as even a negative answer to the question above is compatible with personal fulfillment and positive roles and lives.)

More generally, as hinted earlier, it seems to me that reflection on fundamental values is extremely neglected among altruists. Ozzie Gooen argues that many large-scale altruistic projects are pursued without any serious exploration as to whether the projects in question are even a good way to achieve the ultimate (stated) aims of these projects, despite this seeming like a critical first question to ponder.

I would make a similar argument, only one level further down: just as it is worth exploring whether a given project is among the best ways to achieve a given aim before one pursues that project, so it is worth exploring which aims are most worth striving for in the first place. This, it seems to me, is even more neglected than is exploring whether our pet projects represent the best way to achieve our (provisional) aims. There is often a disproportionate amount of focus on impact, and comparatively little focus on what is the most plausible *aim* of the impact.

Conclusion

In closing, I should again stress that my argument is not that we should only do research and never act — that would clearly be a failure mode, and one that we must also be keen to steer clear of. But my point is that there are good reasons to think that it would be helpful to devote more attention to research in our efforts to improve the world, both on moral and empirical issues — especially at this early point in time.[53]

53 For helpful comments, I thank Teo Ajantaival, Tobias Baumann, and Winston Oswald-Drummond.

S-risk impact distribution is double-tailed

Co-authored with Tobias Baumann

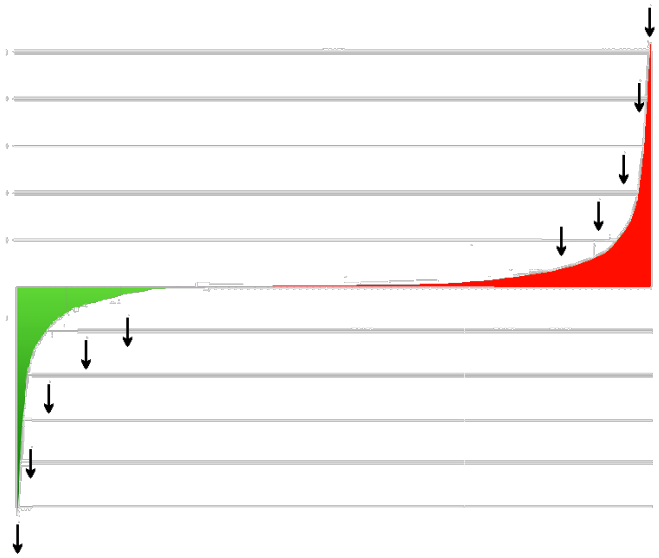
Summary

Discussions about s-risks often rest on a single-tailed picture, focused on how much suffering human civilization could risk *causing*. But when we consider the bigger picture, including s-risks from alien civilizations, we see that human civilization's expected impact on s-risks is in fact double-tailed. This likely has significant implications. For instance, it might mean that we should try to pursue interventions that are robust across *both* tails, and it tentatively suggests that, for a wide range of impartial value systems, it is safest to focus mostly on improving the *quality* of our future.

Introduction

What is the distribution of future expected suffering caused by human civilization?

If civilizations have the potential to *cause* large amounts of suffering, cf. the right tail on the figure below, we should also believe they have the potential to *prevent* large amounts of suffering in expectation.



The figure above shows percentiles along the x-axis and how much suffering is created or reduced by human civilization along the y-axis. The green tail is suffering reduced by human civilization, while the red tail is suffering caused (in expectation). A substantial fraction of the left tail will

amount to reductions of s-risks caused by extraterrestrial civilizations: preventing *their* red-tail scenarios.

The nature of the distribution

As the above figure suggests, the distribution is probably somewhat asymmetric, with more expected suffering caused than prevented. A key question is whether human civilization will be alone in our forward light cone — if so, then the green tail is much less pronounced (though it still does not disappear, since there may be other ways in which human civilization could reduce vast amounts of suffering, though these are more speculative, e.g. acausal trade, reducing universe generation, and unknown unknowns).

It does seem reasonably likely that we are currently alone in our corner of the universe, yet the question is not whether we are *currently* alone, but rather whether we will be alone in the future. And since we will not reach certainty about either of these questions any time soon, it seems that the green tail should be of significant magnitude in expectation. The question, then, is the precise *degree* of asymmetry.

A simple model

As a rough quantitative model, suppose there are N civilizations causing a total amount of S suffering. Without factoring in additional information, we can treat human civilization as a random sample from this set of N civilizations. In expectation, human civilization then causes suffering on a scale of S/N. However, this does not say much about the shape of the distribution of S. In particular, it may be skewed either way, or the variance may be very large compared to the expected value.

By analogy, consider human-caused animal suffering. In expectation, a random person may add to the amount of animal suffering (e.g. through meat consumption). But there are also many whose existence reduces animal suffering (e.g. through animal advocacy), and presumably some who greatly increase animal suffering (e.g. sadistic people who enjoy causing suffering on factory farms or slaughterhouses).

Of course, we do have additional information that may imply that human civilization is better or worse than a randomly sampled one. We could consider values, political dynamics, the frequency of severe conflict, or other factors that affect the likelihood of s-risks, and try to assess how humanity may be different from average. However, since we do not currently know much about what the "average civilization" looks like, it seems reasonable not to deviate much from the "agnostic prior".

Further research on this question may give us a better sense of where human civilization falls in this distribution.

Implications of a double-tailed distribution

If it is best to focus on tail risks, then it may be that the most effective strategy is to focus on *both* tails. That is, it may be optimal to focus on reducing expected suffering in the scenarios found toward the bookends of this distribution (though not necessarily only among the most extreme ends; it could well be optimal to focus on something like the 15 percent lowest and highest percentiles respectively, cf. the arrows on the figure above).

This is not particularly intuitive — after all, what would it mean to (also) focus on the left tail? This seems a question worthy of further research.

Robust interventions

A plausible implication may be that we should seek actions that are robustly good across both tails. For example, reducing extinction risks and increasing the probability of human-driven space colonization may be favorable relative to the left tail, for the purpose of preventing s-risks caused by extraterrestrial civilizations, yet generic extinction reduction also seems likely to increase human-driven s-risks (of course, some interventions may both reduce extinction risks *and* human-driven s-risks).

Likewise, there will probably be strategies that are beneficial if we only consider human-driven s-risks, yet which turn out to be harmful, or at least suboptimal, when we also take the left tail into account.

In contrast, things such as improving human values and cooperation seem beneficial relative to both tails: it reduces the probability of human-caused s-risks, and increases the probability that human-caused colonization is significantly better than ET colonization (conditional on human-driven colonization happening).

Quality may matter most

If interstellar colonization is feasible, the prevention of ET colonization may be more likely than one would naively think. As Lukas Finnveden writes:

> If you accept the self-indication assumption, you should be almost certain that we'll encounter other civilizations if we leave the galaxy. In this case, 95 % of the reachable universe will already be colonised when Earth-originating intelligence arrives, in

> expectation. Of the remaining 5 %, around 70 % would eventually be reached by other civilizations, while 30 % would have remained empty in our absence.

Similar conclusions are reached by Robin Hanson et al. in recent work on "grabby aliens", which suggests that *all* space is likely to be colonized anyway (assuming that "grabby aliens" will emerge).

Thus, on these (speculative) assumptions, *if* total space colonized is almost the same whether humans colonize or not, then, from an impartial moral perspective, the overall *quality* of the colonizing civilizations would seem the most important factor to consider, and plausibly also the safest thing to prioritize. This may hold for a variety of value systems, including classical utilitarianism. After all, if we do not know whether the expected quality of a colonization wave stemming from our own civilization is better or worse than the average quality of colonization waves stemming from other civilizations, it would seem imprudent to insistently push for colonization from our own civilization, and better to instead work to improve its trajectory.

Beware underestimating the probability of very bad outcomes: Historical examples against future optimism

It may be tempting to view history through a progressive lens that sees humanity as climbing toward ever greater moral progress and wisdom. As the famous quote popularized by Martin Luther King Jr. goes: “The arc of the moral universe is long, but it bends toward justice.”

Yet while we may *hope* that this is true, and do our best to increase the probability that it will be, we should also keep in mind that there are reasons to doubt this optimistic narrative. For some, the recent rise of right-wing populism is a salient reason to be less confident about humanity’s supposed path toward ever more compassionate and universal values. But it seems that we find even stronger reasons to be skeptical if we look further back in history. My aim in this post is to present a few historical examples that in my view speak against confident optimism regarding humanity’s future.

Germany in year 1900

In 1900, Germany was far from being a paragon of moral advancement. They were a colonial power, antisemitism was widespread, and bigoted anti-Polish Germanisation policies were in effect. Yet Germany anno 1900 was nevertheless far from being like Germany anno 1939-1945, in which it was the main aggressor in the deadliest war in history and the perpetrator of the largest genocide in history.

In other words, Germany had undergone an extreme case of moral regress along various dimensions by 1942 (the year the so-called Final Solution was formulated and approved by the Nazi leadership) compared to 1900. And this development was not easy to predict in advance. Indeed, for historian of antisemitism Shulamit Volkov, a key question regarding the Holocaust is: “Why was it so hard to see the approaching disaster?”

If one had told the average German citizen in 1900 about the atrocities that their country would perpetrate four decades later, would they have believed it? What probability would they have assigned to the possibility that their country would commit atrocities on such a massive scale? I suspect it would be very low. They might not have seen more reason to expect such moral regress than we do today when we think of our future.

A lesson that we can draw from Germany’s past moral deterioration is, to paraphrase Volkov’s question, that approaching disasters can be hard to see in advance. And this lesson suggests that we

should not be too confident as to whether we ourselves might currently be headed toward disasters that are difficult to see in advance.

Shantideva around year 700

Shantideva was a Buddhist monk who lived in ca. 685-763. He is best known as the author of *A Guide to the Bodhisattva's Way of Life*, which is a remarkable text for its time. The core message is one of profound compassion for all sentient beings, and Shantideva not only *describes* such universally compassionate ideals, but he also presents stirring encouragements and cogent reasoning in favor of acting on those ideals.

That such a universally compassionate text existed at such an early time is a deeply encouraging fact in one sense. Yet in another sense, it is deeply *dis*couraging. That is, when we think about all the suffering, wars, and atrocities that humanity has caused since Shantideva expounded these ideals — centuries upon centuries of brutal violence and torment imposed upon human and non-human beings — it seems that a certain pessimistic viewpoint gains support.

In particular, it seems that we should be pessimistic about notions along the lines of "compassionate ideals presented in a compelling way will eventually create a benevolent world". After all, even today, 1300 years later, where we generally pride ourselves of being far more civilized and morally developed than our ancestors, we are still painfully far from observing the most basic of compassionate ideals in relation to other sentient beings.

Of course, one might think that the problem is merely that people have yet to be exposed to compassionate ideals such as those of Shantideva — or those of Mahavira or Mozi, both of whom lived more than a thousand years before Shantideva. But even if we grant that this is the main problem, it still seems that historical cases like these give us some reason to doubt whether most people ever *will* be exposed to such compassionate ideals, or whether most people would accept such ideals upon being exposed to them, let alone be willing to act on them. The fact that these memes have not caught on to a greater degree than they have, despite existing in such developed forms a long time ago, is some evidence that they are not nearly as virulent as many of us would have hoped.

Speaking for myself at least, I can say that I used to think that people just needed to be exposed to certain compassionate ideals and compassion-based arguments, and then they would change their minds and behaviors due to the sheer compelling nature of these ideals and arguments. But my experience over the years, e.g. with animal advocacy, have made me far more pessimistic about the force of such arguments. And the limited influence of sophisticated expositions of these ideals and

arguments made many centuries ago is further evidence for that pessimism (relative to my previous expectations).

Of course, this is not to say that we can necessarily do better than to promote compassion-based ideals and arguments. It is merely to say that the best we can do might be a lot less significant — or be less likely to succeed — than what many of us had initially expected.

Lewis Gompertz and J. Howard Moore in the 19th century

Lewis Gompertz (ca. 1784-1861) and J. Howard Moore (1862-1916) both have a lot in common with Shantideva, as they likewise wrote about compassionate ethics relating to all sentient beings. (And all three of them touched on wild-animal suffering.) Yet Gompertz and Moore, along with other figures in the 19th century, wrote more explicitly about animal rights and moral vegetarianism than did Shantideva. Two observations seem noteworthy with regard to these writings.

One is that Gompertz and Moore both wrote about these topics before the rise of factory farming. That is, even though authors such as Gompertz and Moore made strong arguments against exploiting and killing other animals in the 19th century, humanity still went on to exploit and kill beings on a far greater scale than ever before in the 20th century, indeed on a scale that is still increasing today.

This may be a lesson for those who are working to reduce risks of astronomical suffering at present: even if you make convincing arguments against a moral atrocity that humanity is committing or otherwise heading toward, and even if you make these arguments at an early stage where the atrocity has yet to (fully) develop, this might still not be enough to prevent it from happening on a continuously expanding scale.

The second and closely related observation is that Gompertz and Moore both seem to have focused exclusively on animal exploitation as it existed in their own times. They did not appear to focus on preventing the problem from getting worse, even though one could argue, in hindsight, that such a strategy might have been more helpful overall.

Indeed, even though Moore's outlook was quite pessimistic, he still seems to have been rather optimistic about the future. For instance, in the preface to his book *The Universal Kinship* (1906), he wrote: "The time will come when the sentiments of these pages will not be hailed by two or three, and ridiculed or ignored by the rest; *they will represent Public Opinion and Law*."

Gompertz appeared similarly optimistic about the future, as he in his *Moral Inquiries* (1824, p. 48) wrote: "though I cannot conceive how any person can shut his eyes to the general state of misery throughout the universe, I still think that it is for a wise purpose; that the evils of life, which could

not properly be otherwise, will in the course of time be rectified ...” Neither Gompertz nor Moore seem to have predicted that animal exploitation would be getting far worse in many ways (e.g. the horrible conditions of factory farms) or that it would increase vastly in scale.

This second observation might likewise carry lessons for animal activists and suffering reducers today. If these leading figures of 19th-century animal activism tacitly underestimated the risk that things might get far worse in the future, and as a result paid insufficient attention to such risks, could it be the case that most activists today are similarly underestimating and underprioritizing future risks of things getting even worse still? This question is at least worth pondering.

On a general and concluding note, it seems important to be aware of our tendencies to entertain wishful thinking and to be under the spell of the illusion of control. Just because a group of people have embraced some broadly compassionate values, and in turn identified ongoing atrocities and future risks based on those values, it does not mean that those people will be able to steer humanity’s future such that we avoid these atrocities and risks. The sad reality is that universally compassionate values are far from being in charge.

Radical uncertainty about outcomes need not imply (similarly) radical uncertainty about strategies

Our uncertainty about how the future will unfold is vast, especially on long timescales. In light of this uncertainty, it may be natural to think that our uncertainty about strategies must be equally vast and intractable. My aim in this brief post is to argue that this is not the case.

Analogies to games, competitions, and projects

Perhaps the most intuitive way to see that vast outcome uncertainty need not imply vast strategic uncertainty is to consider games by analogy. Take chess as an example. It allows a staggering number of possible outcomes on the board, and chess players generally have great uncertainty about how a game of chess will unfold, even as they can make some informed predictions (similar to how we can make informed predictions in the real world).

Yet despite the great outcome uncertainty, there are still many strategies and rules of thumb that are robustly beneficial for increasing one's chances of winning a game of chess. A trivially obvious one is to not lose pieces without good reason, yet seasoned chess players will know a long list of more advanced strategies and heuristics that tend to be beneficial in many different scenarios. (For an example of such a list, see e.g. here.)

Of course, chess is by no means the only example. Across a wide range of board games and video games, the same basic pattern is found: despite vast uncertainty about specific outcomes, there are clear heuristics and strategies that are robustly beneficial.

Indeed, this holds true in virtually any sphere of competition. Politicians cannot predict exactly how an election campaign will unfold, yet they can usually still identify helpful campaign strategies; athletes cannot predict how a given match will develop, yet they can still be reasonably confident about what constitutes good moves and game plans; companies cannot predict market dynamics in detail, yet they can still identify many objectives that would help them beat the competition (e.g. hire the best people and ensure high customer satisfaction).

The point also applies beyond the realm of competition. For instance, when engineers set out to build a big project, there are usually many uncertainties as to how the construction process is going to unfold and what challenges might come up. Yet they are generally still able to identify strategies that can address unforeseen challenges and get the job done. The same goes for just about any

project, including cooperative projects between parties with different aims: detailed outcomes are exceedingly difficult to predict, yet it is generally (more) feasible to identify beneficial strategies.

Disanalogy in scope?

One might object that the examples above all involve rather narrow aims, and those aims differ greatly from impartial aims that relate to the interests of all sentient beings. This is a fair point, yet I do not think it undermines these analogies or the core point that they support.

Granted, when we move from narrower to broader aims and endeavors, our uncertainty about the relevant outcomes will tend to increase — e.g. when our aims involve far more beings and far greater spans of time. And when the outcome space and its associated uncertainty increases, we should also expect our strategic uncertainty to become greater. Yet it plausibly still holds true that we can identify at least some reasonably robust strategies, despite the increase in uncertainty that is associated with impartial aims. At the minimum, it seems plausible that our strategic uncertainty is still smaller than our outcome uncertainty.

After all, if such a pattern of lower strategic uncertainty holds true of a wide range of endeavors on a smaller scale, it seems reasonable to expect that it will apply on larger scales too. Besides, it appears that at least some of the examples mentioned in the previous section would still stand even if we greatly increased their scale. For example, in the case of many video games, it seems that we could increase the scale of the game by an arbitrary amount without meaningfully changing the most promising strategies — e.g. accumulate resources, gain more insights, strengthen your position. And similar strategies are plausibly quite robust relative to many goals in the real world as well, on virtually any scale.

Three robust strategies for reducing suffering

If we grant that we can identify some strategies that are robustly beneficial from an impartial perspective, this naturally raises the question as to what these strategies might be. The following are three examples of strategies for reducing suffering that seem especially robust and promising to me. (This is by no means an exhaustive list.)

- **Movement and capacity building:** Expand the movement of people who strive to reduce suffering, and build a healthy and sustainable culture around this movement. Capacity building also includes efforts to increase the insights and resources available to the movement.

- **Promote concern for suffering:** Increase the level of priority that people devote to the prevention of suffering, and increase the amount of resources that society devotes to its alleviation.
- **Promote cooperation:** Increase society's ability and willingness to engage in cooperative dialogues and positive-sum compromises that can help steer us away from bad outcomes.

The golden middle way: Avoiding overconfidence and passivity

To be clear, I do not mean to invite complacency about the risk that some apparently promising strategies could prove harmful. But I think it is worth keeping in mind that, just as there are costs associated with overconfidence, there are also costs associated with being too uncertain and too hesitant to act on the strategies that seem most promising. All in all, I think we have good reasons to pursue strategies such as those listed above, while still keeping in mind that we do face great strategic uncertainty.

Some pitfalls of utilitarianism

My aim in this post is to highlight and discuss what I consider to be some potential pitfalls of utilitarianism. These are not necessarily pitfalls that undermine utilitarianism at a theoretical level (although some of them might also pose a serious challenge at that level). As I see them, they are more pitfalls at the practical level, relating to how utilitarianism is sometimes talked about, thought about, and acted on in ways that may be suboptimal by the standards of utilitarianism itself.

I should note from the outset that this post is not inspired by recent events involving dishonest and ruinous behavior by utilitarian actors; I have been planning to write this post for a long time. But recent events arguably serve to highlight the importance of some of the points I raise below.

Restrictive formalisms and "formalism first"

A potential pitfall of utilitarianism, in terms of how it is commonly approached, is that it can make us quick to embrace certain formalisms and conclusions, as though we have to accept them on pain of mathematical inconsistency.

Consider the following example: Alice is a utilitarian who thinks that a certain mildly enjoyable experience, *x*, has positive value. On Alice's view, it is clear that no number of instances of *x* would be worse than a state of extreme suffering, since a state of extreme suffering and a mildly enjoyable experience are completely different categories of experience. Over time, Alice reads about different views of wellbeing and axiology, and she eventually changes her position such that she finds it more plausible that no experiential states are above a neutral state, and that no states have intrinsic positive value (i.e. she comes to embrace a minimalist axiology).

Alice thus no longer considers it plausible to assign positive value to experience *x*, and instead now assigns mildly negative value to the experience (e.g. because the experience is not entirely flawless; it contains some bothersome disturbances). Having changed her mind about the value of experience *x*, Alice now feels mathematically compelled to say that sufficiently many instances of that experience are worse than any experience of extreme suffering, even though she finds this implausible on its face — she still thinks state *x* and states of extreme suffering belong to wholly different categories of experience.

To be clear, the point I am trying to make here is not that the final conclusion that Alice draws is implausible. My point is rather that certain prevalent ways of formalizing value can make people feel needlessly compelled to draw particular conclusions, as though there are no coherent

alternatives, when in fact there are. More generally, there may be a tendency to "put formalism first", as it were, rather than to consider substantive plausibility first, and to then identify a coherent formalism that fits our views of substantive plausibility.

Note that the pitfall I am gesturing at here is not one that is strictly implied by utilitarianism, as one can be a utilitarian yet still reject standard formalizations of utilitarianism. But being bound to a restrictive formalization scheme nevertheless seems common, in my experience, among those who endorse or sympathize with utilitarianism.

Risky and harmful decision procedures

A standard distinction in consequentialist moral theory is that between 'consequentialist criteria of rightness' and 'consequentialist decision procedures'. One might endorse a consequentialist criterion of rightness — meaning that consequences determine whether a given action is right or wrong — without necessarily endorsing consequentialist decision procedures, i.e. decision procedures in which one decides how to act based on case-by-case calculations of the expected outcomes.

Yet while this distinction is often emphasized, it still seems that utilitarianism is prone to inspire suboptimal decision procedures, also by its own standards (as a criterion of rightness). The following are a few of the ways in which utilitarianism can inspire suboptimal decision procedures, attitudes, and actions by its own standards.

Allowing speculative expected value calculations to determine our actions

A particular pitfall is to let our actions be strongly determined by speculative expected value calculations. There are various reasons why this may be suboptimal by utilitarian standards, but an important one is simply that the probabilities that go into such calculations are likely to be inaccurate. If our probability estimates on a given matter are highly uncertain and likely to change a lot as we learn more, there is a large risk that it is suboptimal to make any strong bets on our current estimates.

The robustness of a given probability estimate is thus a key factor to consider when deciding whether to act on that estimate, yet it can be easy to neglect this factor in real-world decisions.

Underestimating the importance of emotions, virtues, and other traits of moral actors

A related pitfall is to underestimate the significance of emotions, attitudes, and virtues. Specifically, if we place a strong emphasis on the consequences of actions, we might in turn be inclined to underemphasize the traits and dispositions of the moral actors themselves. Yet the traits and

dispositions of moral actors are often critical to emphasize and to actively develop if we are to create better outcomes. Our cerebral faculties and our intuitive attitudinal faculties can both be seen as tools that enable us to navigate the world, and the latter are often more helpful for creating desired outcomes than the former (cf. Gigerenzer, 2001).

A specific context in which I and others have tried to argue for the importance of underlying attitudes and traits, in contrast to mere cerebral beliefs, is when it comes to animal ethics. In particular, engaging in practices that are transparently harmful and exploitative toward non-human beings is harmful not only in terms of how it directly contributes to those specific exploitative practices, but also in terms of how it shapes our emotions, attitudes, and traits — and thus ultimately our behavior.

More generally, to emphasize outcomes while placing relatively little emphasis on the traits of humans, as moral actors, seems to overlook the largely habitual and disposition-based nature of human behavior. After all, our emotions and attitudes not only play important roles in our individual motivations and actions, but also in the social incentives that influence the behavior of others (cf. Haidt, 2001).

In short, if one embraces a consequentialist criterion of rightness, it seems that there are good reasons to cultivate the temperament of a virtue ethicist and the felt attitudes of a non-consequentialist who finds certain actions unacceptable in practically all situations.

Uncertainty-induced moral permissiveness

Another pitfall is to practically surrender one's capacity for moral judgment due to uncertainty about long-term outcomes. In its most extreme manifestations, this might amount to declaring that we do not know whether people who committed large-scale atrocities in the past acted wrongly, since we do not know the ultimate consequences of those actions. But perhaps a more typical manifestation is to fail to judge, let alone oppose, ongoing harmful actions and intolerant values (e.g. clear cases of discrimination), again with reference to uncertainty about the long-term consequences of those actions and values.

This pitfall relates to the point about dispositions and attitudes made above, in that the disposition to be willing to judge and oppose harmful actions and views plausibly has better overall consequences than a disposition to be meek and unwilling to take a strong stance against such things.

After all, while there is significant uncertainty about the long-term future, one can still make reasonable inferences about which broad directions we should ideally steer our civilization toward over the long term (e.g. toward showing concern for suffering in prudent yet morally serious ways).

Utilitarians have reason to help steer the future in those directions, and to develop traits and attitudes that are commensurate with such directional changes. (See also "Radical uncertainty about outcomes need not imply (similarly) radical uncertainty about strategies".)

Uncertainty-induced lack of moral drive

A related pitfall is uncertainty-induced lack of moral drive, whereby empirical uncertainty serves as a stumbling block to dedicated efforts to help others. This is probably also starkly suboptimal, for reasons similar to those outlined above: all things considered, it is likely ideal to develop a burning drive to help other sentient beings, despite uncertainty about long-term outcomes.

Perhaps the main difficulty in this respect is to know which particular project or aim is most important to work on. Yet a potential remedy to this problem (here conveyed in a short and crude fashion) might be to first make a dedicated effort toward the concrete goal of figuring out which projects or aims seem most worth pursuing — i.e. a broad and systematic search, informed by copious reading. And when one has eventually identified an aim or project that seems promising, it might be helpful to somewhat relax the "doubting modules" of our minds and to stick to that project for a while, pursuing the chosen aim with dedication (unless something clearly better comes up).

A more plausible approach

The previous sections have mostly pointed to suboptimal ways to approach utilitarian decision procedures. In this section, I want to briefly outline what I would consider a more defensible way to approach decision-making from a utilitarian perspective (whether one is a pure utilitarian or whether one merely includes a utilitarian component in one's moral view).

I think two key facts must inform any plausible approach to utilitarian decision procedures:

1. We have massive empirical uncertainty.
2. We humans have a strong proclivity to deceive ourselves in self-serving ways.

These two observations carry significant implications. In short, they suggest that we should generally approach moral decisions with considerable humility, and with a strong sense of skepticism toward conclusions that are conveniently self-serving or low on integrity.

Given our massive uncertainty and our endlessly rationalizing minds, the ideal approach to utilitarian decision procedures is probably one that has a rather large distance between the initial question of "how to act" and the final decision to pursue a given action — at least when one is trying to calculate one's way to an optimal decision (as opposed to when one is relying on commonly endorsed rules of thumb or intuitions). And this distance should probably be especially

large if the decision that at first seems most recommendable is one that other moral views, along with common-sense intuitions, would deem profoundly wrong.

In other words, it seems that utilitarian decision procedures are best approached by assigning a fairly high prior to the judgments of other ethical views and common-sense moral intuitions (in terms of how plausible those judgments are from a utilitarian perspective), at least when these other views and intuitions converge strongly on a given conclusion. And it seems warranted to then be quite cautious and slow to update away from that prior, in part because of our massive uncertainty and our self-deceived minds. This is not to say that one could not end up with significant divergences relative to other widely endorsed moral views, but merely that such strong divergences probably need to be supported by a level of evidence that exceeds a rather high bar.

Likewise, it seems worth approaching utilitarian decision procedures with a prior that strongly favors actions of high integrity, not least because we should expect our rationalizing minds to be heavily biased toward low integrity — especially when nobody is looking.

Put briefly, it seems that a more defensible approach to utilitarian decision procedures would be animated by significant humility and would embody a strong inclination toward key virtues of integrity, kindness, honesty, etc., partly due to our strong tendency to excuse and rationalize deficiencies in these regards.

The link between utilitarian judgments and Dark Triad traits: A cause for reflection

There are many studies that find a modest but significant association between proto-utilitarian judgments and the personality traits of psychopathy (impaired empathy) and Machiavellianism (manipulativeness and deceitfulness). (See Bartels & Pizarro, 2011; Koenigs et al., 2012; Gao & Tang, 2013; Djeriouat & Trémolière, 2014; Amiri & Behnezhad, 2017; Balash & Falkenbach, 2018; Karandikar et al., 2019; Halm & Möhring, 2019; Dinić et al., 2020; Bolelli, 2021; Luke & Gawronski, 2021; Schönegger, 2022.)

Specifically, the aspect of utilitarian judgment that seems most associated with psychopathy is the willingness to commit harm for the sake of the greater good, whereas endorsement of impartial beneficence — a core feature of utilitarianism and many other moral views — is associated with empathic concern, and is thus negatively associated with psychopathy (Kahane et al., 2018; Paruzel-Czachura & Farny, 2022). Another study likewise found that the connection between psychopathy and utilitarian moral judgments is in part explained by a reduced aversion to carrying out harmful acts (Patil, 2015).

Of course, whether a particular moral view, or a given feature of a moral view, is associated with certain undesirable personality traits by no means refutes that moral view. But the findings reviewed above might still be a cause for self-reflection among those of us who endorse or sympathize with some form of utilitarianism.

For example, maybe utilitarians are generally inclined to have fewer moral inhibitions compared to most people — e.g. because utilitarian reasoning might override intuitive judgments and norms, or because utilitarians are (perhaps) above average in trait Machiavellianism, in which case they might have fewer strongly felt moral inhibitions to overcome in the first place. And if utilitarians do tend to have fewer or weaker moral restraints of certain kinds, this could in turn dispose them to be less ethical in some respects, also by their own standards.

To be clear, this is all somewhat speculative. Yet, at the same time, these speculations are not wholly unmotivated. In terms of potential upshots, it seems that a utilitarian proneness to reduced moral restraint, if real, would give utilitarian actors additional reason to be skeptical of inclinations to disregard common moral inhibitions against harmful acts and low-integrity behavior. In short, it would give utilitarians even more reason to err on the side of integrity.[54]

54 For helpful comments, I am grateful to Tobias Baumann, Simon Knutsson, and Winston Oswald-Drummond.

Distrusting salience: Keeping unseen urgencies in mind

The psychological appeal of salient events and risks can be a major hurdle to optimal altruistic priorities and impact. My aim in this post is to outline a few reasons to approach our intuitive fascination with salient events and risks with a fair bit of skepticism, and to actively focus on that which is important yet unseen, hiding in the shadows of the salient.

General reasons for caution: Availability bias and related biases

The human mind is subject to various biases that involve an overemphasis on the salient, i.e. that which readily stands out and captures our attention.

In general terms, there is the availability bias, also known as the availability heuristic, namely the common tendency to base our beliefs and judgments on information that we can readily recall. For example, we tend to overestimate the frequency of events when examples of these events easily come to mind.

Closely related is what is known as the salience bias, which is the tendency to overestimate salient features and events when making decisions. For instance, when deciding to buy a given product, the salience bias may lead us to give undue importance to a particularly salient feature of that product — e.g. some fancy packaging — while neglecting less salient yet perhaps more relevant features.

A similar bias is the recency bias: our tendency to give disproportionate weight to recent events in our belief-formation and decision-making. This bias is in some sense predicted by the availability bias, since recent events tend to be more readily available to our memory. Indeed, the availability bias and the recency bias are sometimes considered equivalent, even though it seems more accurate to view the recency bias as a consequence or a subset of the availability bias; after all, readily remembered information does not always pertain to recent events.

Finally, there is the phenomenon of belief digitization, which is the tendency to give undue weight to (what we consider) the single most plausible hypothesis in our inferences and decisions, even when other hypotheses also deserve significant weight. For example, if we are considering hypotheses A, B, and C, and we assign them the probabilities 50 percent, 30 percent, and 20 percent, respectively, belief digitization will push us toward simply accepting A as though it were true. In other words, belief digitization pushes us toward altogether discarding B and C, even though B and C collectively have the same probability as A. (See also related studies on Salience Theory and on the overestimation of salient causes and hypotheses in predictive reasoning.)

All of the biases mentioned above can be considered different instances of a broader cluster of availability/salience biases, and they each give us reason to be cautious of the influence that salient information has on our beliefs and our priorities.

The news: A common driver of salience-related distortions

One way in which our attention can become preoccupied with salient (though not necessarily crucial) information is through the news. Much has been written against spending a lot of time on the news, and the reasons against it are probably even stronger for those who are trying to spend their time and resources in ways that help sentient beings most effectively.

For even if we grant that there is substantial value in following the news, it seems plausible that the opportunity costs are generally too high, in terms of what one could instead spend one's limited time learning about or advocating for. Moreover, there is a real risk that a preoccupation with the news has outright harmful effects overall, such as by gradually pulling one's focus away from the most important problems and toward less important and less neglected problems. After all, the prevailing news criteria or news values decidedly do not reflect the problems that are most important from an impartial perspective concerned with the suffering of all sentient beings.

I believe the same issue exists in academia: A certain issue becomes fashionable, there are calls for abstracts, and there is a strong pull to write and talk about that given issue. And while it may indeed be important to talk and write about those topics for the purpose of getting ahead — or not falling behind — in academia, it seems more doubtful whether such topical talk is at all well-adapted for the purpose of making a difference in the world. In other words, the "news values" of academia are not necessarily much better than the news values of mainstream journalism.

The narrow urgency delusion

A salience-related pitfall that we can easily succumb to when following the news is what we may call the "narrow urgency delusion". This is when the news covers some specific tragedy and we come to feel, at a visceral level, that this tragedy is the most urgent problem that is currently taking place. Such a perception is, in a very important sense, an illusion.

The reality is that tragedy on an unfathomable scale is always occurring, and the tragedies conveyed by the news are sadly but a tiny fraction of the horrors that are constantly taking place around us. Yet the tragedies that are always occurring, such as children who suffer and die from undernutrition and chickens who are boiled alive, are so common and so underreported that they all too readily fade from our moral perception. To our intuitions, these horrors seemingly register as mere baseline horror — as unsalient abstractions that carry little felt urgency — even though the horrors in

question are every bit as urgent as the narrow sliver of salient horrors conveyed in the news (Vinding, 2020, sec. 7.6).

We should thus be clear that the delusion involved in the narrow urgency delusion is not the "urgency" part — there is indeed unspeakable horror and urgency involved in the tragedies reported by the news. The delusion rather lies in the "narrow" part; we find ourselves in a condition that contains *extensive* horror and torment, *all* of which merits compassion and concern.

So it is not that the salient victims are less important than what we intuitively feel, but rather that the countless victims whom we effectively overlook are far more important than what we (do not) feel.

Massive problems that always face us: Ongoing moral disasters and future risks

The following are some of the urgent problems that always face us, yet which are often less salient to us than the individual tragedies that are reported in the news:

- Prevalent forms of human suffering (e.g. due to cancer, the second most common cause of human death, or due to political oppression — a recent report concluded that 70 percent of the world's population live in autocracies).
- The industrial farming and slaughter of non-human animals.
- The suffering of wild animals due to natural processes.
- Risks of astronomical future suffering (s-risks).

These common and ever-present problems are, by definition, not news, which hints at the inherent ineffectiveness of news when it comes to giving us a clear picture of the reality we inhabit and the problems that confront us.

As the final entry on the list above suggests, the problems that face us are not limited to ongoing moral disasters. We also face risks of future atrocities, potentially involving horrors on an unprecedented scale. Such risks will plausibly tend to feel even less salient and less urgent than do the ongoing moral disasters we are facing, even though our influence on these future risks — and future suffering in general — could well be more consequential given the vast scope of the long-term future.

So while salience-driven biases may blind us to ongoing large-scale atrocities, they probably blind us even more to future suffering and risks of future atrocities.

Salience-driven distortions in efforts to reduce s-risks

There are many salience-related hurdles that may prevent us from giving significant priority to the reduction of future suffering. Yet even if we do grant a strong priority to the reduction of future suffering, including s-risks in particular, there are reasons to think that salience-driven distortions still pose a serious challenge in our prioritization efforts.

Our general availability bias gives us some reason to believe that we will overemphasize salient ideas and hypotheses in efforts to reduce future suffering. Yet perhaps more compelling are the studies on how we tend to greatly overestimate salient hypotheses when we engage in predictive and multi-stage reasoning in particular. (Multi-stage reasoning is when we make inferences in successive steps, such that the output of one step provides the input for the next one.)

After all, when we are trying to predict the main sources of future suffering, including specific scenarios in which s-risks materialize, we are very much engaging in predictive and multi-stage reasoning. Therefore, we should arguably expect our reasoning about future causes of suffering to be too narrow by default, with a tendency to give too much weight to a relatively small set of salient risks at the expense of a broader class of less salient (yet still significant) risks that we are prone to dismiss in our multi-stage inferences and predictions.

This effect can be further reinforced through other mechanisms. For example, if we have described and explored — or even just imagined — a certain class of risks in greater detail than other risks, then this alone may lead us to regard those more elaborately described risks as being more likely than less elaborately explored scenarios. Moreover, if we find ourselves in a group of people who focus disproportionally on a certain class of future scenarios, this may further increase the salience and perceived likelihood of these scenarios, compared to alternative scenarios that may be more salient in other groups and communities.

Reducing salience-driven distortions

The pitfalls mentioned above seem to suggest some concrete ways in which we might reduce salience-driven distortions in efforts to reduce future suffering.

First, they recommend caution about the danger of neglecting less salient hypotheses when engaging in predictive and multi-stage reasoning. Specifically, when thinking about future risks, we should be careful not to simply focus on what appears to be the single greatest risk, and to effectively neglect all others. After all, even if the risk we regard as the single greatest risk indeed *is* the single greatest risk, that risk might still be fairly modest compared to the totality of future risks, and we might still do better by deliberately working to reduce a relatively broad class of risks.

Second, the tendency to judge scenarios to be more likely when we have thought about them in detail would seem to recommend that we avoid exploring future risks in starkly unbalanced ways. For instance, if we have explored one class of risks in elaborate detail while largely neglecting another, it seems worth trying to outline concrete scenarios that exemplify the more neglected class of risks, so as to correct any potentially unjustified disregard of their importance and likelihood.

Third, the possibility that certain ideas can become highly salient in part for sociological reasons may recommend a strategy of exchanging ideas with, and actively seeking critiques from, people who do not fully share the outlook that has come to prevail in one's own group.

In general, it seems that we are likely to underestimate our empirical uncertainty (Vinding, 2020, sec. 9.1-9.2). The space of possible future outcomes is vast, and any specific risk that we may envision is but a tiny subset of the risks we are facing. Hence, our most salient ideas regarding future risks should ideally be held up against a big question mark that represents the many (currently) unsalient risks that confront us.

Put briefly, we need to cultivate a firm awareness of the limited reliability of salience, and a corresponding awareness of the immense importance of the unsalient. We need to make an active effort to keep unseen urgencies in mind.

Popular views of population ethics imply a priority on preventing worst-case outcomes

A wide variety of views can support a focus on preventing worst-case outcomes. More than that, it appears that the views of population ethics held by the general population also, on average, imply a priority on preventing futures with large numbers of miserable beings. My aim in this post is to elaborate on this point and to briefly explore its relevance.

Asymmetric scope sensitivity

A recent study set out to investigate people's intuitions on population ethics, exploring how people judge the value of different populations of happy and unhappy individuals (Caviola et al., 2022a). The study included a number of sub-studies, which generally found that people endorse a weak asymmetry in population ethics. That is, people tend to believe that miserable lives and suffering weigh somewhat stronger than do happy lives and happiness, even when the misery and happiness in question are claimed to be equally intense (Caviola et al., 2022a, p. 8).

One of the sub-studies sought to examine the participants' evaluations of populations of varying sizes (these participants were all from the US). Subjects were asked to consider different pairs of hypothetical civilizations, each of which would last for millions of years and differ only in terms of their size. Specifically, the participants were asked to rate how strongly they would prefer a larger over a smaller happy population, as well as how strongly they would prefer a smaller over a larger unhappy population. These ratings were made in pairwise population comparisons of 1,000 vs 10,000, 1 million vs. 10 million, and 1 billion vs 10 billion people (Caviola et al., 2022a, p. 11).

The study found that subjects generally preferred a larger happy population in the case of 1,000 vs 10,000 people, yet this preference declined as the populations in question became larger. In fact, the preference for larger happy populations declined so strongly that the participants on average preferred the smaller happy population in the 1 billion vs 10 billion comparison (Caviola et al., 2022a, p. 12). The study likewise reported that "the median response for the ideal population size was 1.5 million for the happy civilization" (Caviola et al., 2022a, p. 12).

In contrast, subjects generally preferred a smaller unhappy population, and this preference became increasingly strong when population pairs became larger. As the authors summarized these findings: "people show 'asymmetric scope sensitivity' with respect to happy and unhappy population sizes.

And this asymmetric scope sensitivity was more pronounced the larger the population sizes got." (Caviola et al., 2022a, p. 12).

It is worth noting that the study did not specify whether the populations under consideration would represent all humans in the world, though it was specified that each of the populations in the pairwise comparisons would have "multiple Earth-like planets available to them" (Caviola et al., 2022a, p. 11). Subjects were also informed that the populations in question would have "no issues with resource depletion, environmental degradation or overpopulation", yet it is nevertheless possible that worries about overpopulation influenced the responses of some of the participants (Caviola et al., 2022a, p. 11, p. 13).

Why this popular view would support a focus on preventing worst-case outcomes

It is fairly straightforward to see how people's average preferences regarding different populations of happy and unhappy people would support a focus on preventing worst-case outcomes.

If people's average preference for a larger happy population declines and eventually reverses, whereas the average preference for smaller unhappy populations gets stronger as the populations in question increase, then it seems to follow that people on average consider it more valuable to prevent the existence of a very large unhappy population than to ensure the existence of a very large happy population. In other words, people generally seem to think that it is more valuable to prevent worst-case outcomes than to create best-case outcomes. And this asymmetry may well become stronger when the prospect of space colonization and astronomical stakes enter the picture. That is, when the populations under consideration are not just 1 billion versus 10 billion, but many orders of magnitude larger — e.g. 1 trillion vs 10 trillion, 1 quadrillion vs 10 quadrillion, etc.

To be sure, the findings above do not directly show that people endorse a strong asymmetric scope sensitivity for astronomical-scale populations, since it does not include any such populations. (It would be great to have studies that directly explore this issue.) Yet given the trend toward increasingly strong preferences for smaller populations in comparisons of larger populations, it seems reasonable to expect that the asymmetry would also hold — and plausibly grow even stronger — on astronomical scales.

But again, even if people's average preference for smaller unhappy populations does not get stronger still for populations larger than 10 billion, it is worth reiterating that people already seem to consider it more valuable to prevent worst-case outcomes when considering Earth-scale populations.

Other surveys likewise indicate that the prevention of future suffering, including worst-case outcomes in particular, is considered a key priority among the general population. For example, one survey (n=14,866) asked people what a future civilization should strive for, and found that the most popular aim, ranked highest by roughly a third of the respondents, was "minimizing suffering" (Future of Life Institute, 2017). Similarly, a small survey (n=99) asked people whether they would accept one minute of extreme suffering in order to add a given number of happy years to their lives, and found that a plurality, around 45 percent, said that no number of happy years could lead them to accept the offer (Tomasik, 2015).

Lastly, a pilot study (n=172) conducted by Caviola et al. asked people at what ratio of happy to unhappy lives they would be willing to push a button to create a new world, and found that people on average thought that it required roughly a ratio of 100:1 happy people to unhappy people (Caviola et al., 2022b, p. 7; Contestabile, 2022, sec. 4).[55] In other words, people generally seem to endorse a rather strong asymmetry when it comes to the creation of new worlds, which may suggest that people would likewise endorse a strong asymmetry in the case of space colonization aimed at populating new worlds. (But again, it would be helpful to have studies that probe this question directly.)

Similar views in the academic literature

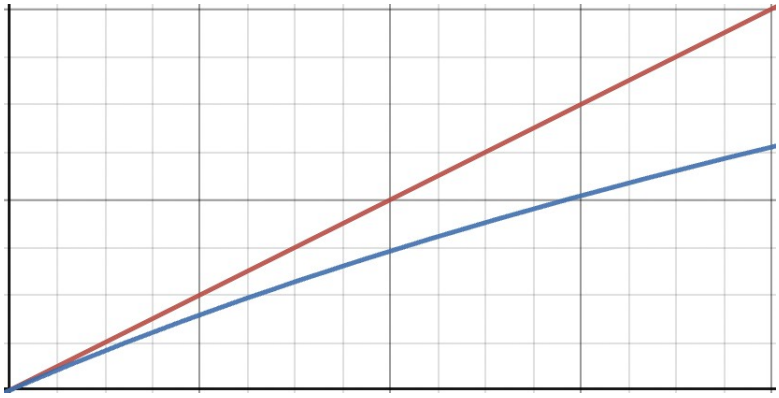
It is worth noting that people's average preferences regarding population sizes appear to resemble a class of views that have been explored, and to some degree endorsed, by academic philosophers. In particular, asymmetric scope sensitivity is implied by theoretical views that maintain that happy lives, or states of happiness in general, add diminishing marginal value to the world, whereas miserable lives, or states of suffering in general, add non-diminishing disvalue to the world (Hurka, 1983; 2010; Parfit, 1984, pp. 406-412; Knutsson, 2016).[56]

These asymmetric views likewise tend to support a focus on the prevention of worst-case outcomes, especially on astronomical scales, where the asymmetry can become extremely strong (Vinding, 2020, sec. 6.2). The figures below illustrate how a view of this kind might see the population-ethical

55 The absolute population sizes that were included in the questions in this study were rather small, ranging from 10 to 1,000, and hence people's asymmetric scope sensitivity likely did not apply strongly in these versions of the ratio question. One might thus expect that the ratio would be even higher if the question were phrased in terms of larger populations; this also seems worth exploring in future studies.

56 Specifically, Hurka argues that happy lives plausibly add diminishing marginal value to the world (Hurka, 1983), whereas he appears sympathetic to an asymmetric evaluation of the disvalue of suffering that would regard its marginal disvalue as non-diminishing, or at least as less strongly diminishing (cf. Hurka, 2010, p. 200). Similarly, Parfit explicitly wrote that "the badness of extra suffering never declines" (p. 406), while he appeared to consider it more plausible that pleasure has diminishing marginal value, even though he ultimately felt compelled to reject the latter view based on its implications (Parfit, 1984, pp. 406-412).

asymmetry on different scales, being fairly weak on smaller scales while being strong on larger scales.



Earth-scale asymmetry between the disvalue of unhappy lives (red) vs the value of happy lives (blue).

Astronomical-scale asymmetry between the disvalue of unhappy lives (red) vs the value of happy lives (blue)

Why this is relevant: Potential implications for democratic institutions

What is the relevance of people's average preferences regarding population ethics, and why, specifically, is it relevant what those views would imply for our priorities?

To my mind, the main relevance of these findings lies in their potential implications for the policies of representative political institutions. That is, if political efforts to improve the long-term future are to represent people's views and preferences in a genuinely democratic fashion, then these findings seem to carry significant implications for those political efforts and priorities.

In short, if the findings regarding asymmetric scope sensitivity do indeed reflect people's average preferences on population ethics, they seem to imply that representative political institutions would make it a priority to prevent worst-case outcomes that involve large miserable populations. At the very least, the findings suggest that the prevention of such outcomes would be a stronger priority for democratic institutions than would the creation of a very large happy future population.

Notice that the statements above are descriptive in nature, being phrased in terms of what democratic political institutions would do, if they were to be representative. I am not claiming that the priorities in question would be right by virtue of being democratic. But given that many of us happen to live in (more or less) representative democracies, it seems worth investigating which priorities those systems would imply by the standards of their own stated ideals.

To elaborate further, the popular support for asymmetric scope sensitivity suggests that a focus on preventing worst-case outcomes is by no means a fringe priority, but rather a priority that most people seem to weigh considerably higher than the creation of a very large happy population. Those aiming to prevent worst-case outcomes may thus have good reason to advance this goal in the political realm, with an awareness that the aim of preventing worst-case outcomes plausibly has greater democratic legitimacy and support than does the goal of creating a very large happy population (cf. the studies and surveys cited above). And this seems especially true if efforts to create a large happy population come at the opportunity cost of preventing worst-case outcomes, or even if they come at the opportunity cost of failing to prevent intense suffering for currently existing individuals.[57]

References

Caviola, L. et al., (2022a). Population ethical intuitions. *Cognition*, 218, 104941. Ungated

Caviola, L. et al. (2022b). Supplementary Materials for Population ethical intuitions. Ungated

Contestabile, B. (2022). Is There a Prevalence of Suffering? Ungated

Future of Life Institute. (2017). Superintelligence survey. Ungated

Hurka, T. (1983). Value and Population Size. *Ethics*, 93(3), pp. 496-507.

Hurka, T. (2010). Asymmetries In Value. *Nous*, 44(2), pp. 199-223. Ungated

Knutsson, S. (2016). What is the difference between weak negative and non-negative ethical views? Ungated

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Tomasik, B. (2015). A Small Mechanical Turk Survey on Ethics and Animal Welfare. Ungated

Vinding, M. (2020). *Suffering-Focused Ethics: Defense and Implications*. Ratio Ethica. Ungated

57 For helpful comments, I am grateful to David Althaus and Tobias Baumann.

Other Resources on Suffering-Focused Ethics

Books

The Hedonistic Imperative (1995) by David Pearce

Suffering and Moral Responsibility (1999) by Jamie Mayerfeld

The Battle for Compassion: Ethics in an Apathetic Universe (2011) by Jonathan Leighton

Can Biotechnology Abolish Suffering? (2017) by David Pearce

Suffering-Focused Ethics: Defense and Implications (2020) by Magnus Vinding

Reasoned Politics (2022) by Magnus Vinding

Avoiding the Worst: How to Prevent a Moral Catastrophe (2022) by Tobias Baumann

The Tango of Ethics: Intuition, Rationality and the Prevention of Suffering (2023) by Jonathan Leighton

Compassionate Purpose: Personal Inspiration for a Better World (forthcoming) by Magnus Vinding

Essays and articles

Essays on Reducing Suffering by Brian Tomasik

Essays and articles by Simon Knutsson

Essays and articles by Jonathan Leighton

Minimalist axiologies sequence by Teo Ajantaival

The Case for Suffering-Focused Ethics and Tranquilism by Lukas Gloor

A longtermist critique of "The expected value of extinction risk reduction is positive" by Anthony DiGiovanni

Resources for Sustainable Activism by various authors

Organizations

Animal Ethics

Center for Reducing Suffering (CRS)

Center on Long-Term Risk (CLR)

Organization for the Prevention of Intense Suffering (OPIS)

See also the following collection of Suffering-Focused Ethics Resources