Philosophical Considerations Relevant to Valuing Continued Human Survival: Conceptual Analysis, Population Axiology, and Decision Theory

Andreas Mogensen (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2023

GPI Working Paper No . 9-2023

Philosophical Considerations Relevant to Valuing Continued Human Survival: Conceptual Analysis, Population Axiology, and Decision Theory

Andreas L. Mogensen[*]

September 2023

Contents

[*] For extensive comments on the material of this report, I especially want to thank Hilary Greaves and Teru Thomas. I also wish to thank Adam Bales, Jacob Barrett, Alexander Barry, Lara Buchak, Tomi Francis, Richard Pettigrew, Andreas Schmidt, Shlomi Segall, Rhys Southan, Christian Tarsney, David Thorstad, Hayden Wilkinson, and Timothy Luke Williamson. This report draws on a range of material that already exists in the form of forthcoming papers or working papers. An abridged version of section 3 of this report is forthcoming in Hilary Greaves, Jacob Barrett, and David Thorstad, eds. (2024) *Essays on Longtermism*. Oxford: Oxford University Press, under the title 'Would a world without us be worse? Clues from population axiology.' Section 3.2 and 3.3 reproduce material from a working paper entitled, 'Do not go gentle: why the Asymmetry does not support anti-natalism.' Section 3.4 draws on some ideas explored in 'The weight of suffering,' forthcoming in *The Journal of Philosophy*. Lastly, section 4.2 condenses material from my working paper, 'Respect for others' risk attitudes and the long-run future.'

1 Introduction

Many think that human extinction would be a catastrophic tragedy, and that we ought to do more to reduce extinction risk. There is less agreement on exactly why. If some catastrophe were to kill everyone, that would obviously be horrific. Still, many think the deaths of billions of people don't exhaust what would be so terrible about extinction. After all, we can be confident that billions of people are going to die – many horribly and before their time - if humanity does *not* go extinct. The key difference seems to be that they will be survived by others. What's the importance of that?

Some take the view that the special moral importance of preventing extinction is explained in terms of the value of increasing the number of flourishing lives that will ever be lived, since there could be so many people in the vast future available to us (see Kavka 1978; Sikora 1978; Parfit 1984; Bostrom 2003; Ord 2021: 43-49). Others emphasize the moral importance of conserving existing things of value and hold that humanity itself is an appropriate object of conservative valuing (see Cohen 2012; Frick 2017). Many other views are possible (see esp. Scheffler 2013, 2018).

However, not everyone is so sure that human extinction would be regrettable. In the final section of the last book published in his lifetime, Parfit (2011: 920–925) considers what can actually be said about the value of all future history. No doubt, people will continue to suffer and despair. They will also continue to experience love and joy. Will the good be sufficient to outweigh the bad? Will it all be worth it? Parfit's discussion is brief and inconclusive. He leans toward 'Yes,' writing that our "descendants might, I believe, make the future very good." (Parfit 2011: 923) But 'might' falls far short of 'will'.

Others are confidently pessimistic. Some take the view that human lives are not worth starting because of the suffering they contain. Benatar (2006) adopts an extreme version of this view, which I discuss in section 3.3. He claims that "it would be better, all things considered, if there were no more people (and indeed no more conscious life)." (Benatar 2006: 146) Scepticism about the disvalue of human extinction is especially likely to arise among those concerned about our effects on non-human animals and the natural world. In his classic paper defending the view that all living things have moral status, Taylor (1981:

209) argues, in passing, that human extinction would "most likely be greeted with a hearty 'Good riddance!' " when viewed from the perspective of the biotic community as a whole. May (2018) argues similarly that because there "is just too much torment wreaked upon too many animals and too certain a prospect that this is going to continue and probably increase," we should take seriously the idea that human extinction would be morally desirable. Our abysmal treatment of non-human animals may also be thought to bode ill for our potential treatment of other kinds of minds with whom we might conceivably share the future and view primarily as tools: namely, minds that might arise from inorganic computational substrates, given suitable developments in the field of artificial intelligence (Saad and Bradley forthcoming).

This paper takes up the question of whether and to what extent the continued existence of humanity is morally desirable. For the sake of brevity, I'll refer to this as *the value of the future*, leaving the assumption that we conditionalize on human survival implicit. On its face, the case for assigning importance to reducing the risk of human extinction hinges largely on how we answer this question. Even if we're confident that the survival of humanity is a good thing, the question of exactly how good may determine how much weight to put on reducing extinction risk, relative to other priorities.

Considered in its full generality, this is an impossibly grand question. My aim in this paper is to outline and explore some key philosophical issues relevant to determining the value of the future, drawn from the fields of population ethics (section 3) and decision theory (section 4). I have more to say on the former than on the latter. Before that, I also do my part to clarify what we're even asking here (section 2).

All this is just a very small part of the puzzle. There are myriad empirical questions with which I do not engage at all. There are also many important philosophical questions that I leave on the table, including some in decision theory, such as ambiguity aversion. The selection of topics only partially reflects my judgments about relative importance. It also reflects the gaps in my own expertise, as well as my own guesses about the extent to which I have something to contribute on a given topic. I hope this report inspires others to contribute their own treatments of the many important topics I was unable to cover.

2 Conceptual Analysis - Clarifying the Question

What do we mean by the value of the future? I start, in section 2.1, by noting some basic assumptions about the theory of value. I then consider up to what time we should assume humanity doesn't go extinct for the purposes of this exercise (2.2), what we mean by 'humanity' (2.3), the place of non-human individuals in all this (2.4), and what it means to *not* go extinct (2.5). Finally, I discuss the role of uncertainty, both empirical and normative, in section 2.6.

2.1 Axiology

Our question is, roughly, whether and, if so, to what extent it would be morally better for humanity not to go extinct. Answering this question therefore requires us to have some sense of what sort of things are good and bad in and of themselves, and what sort of outcomes are better than others. Philosophers call a theory of this kind an *axiology*.

The good is here assumed to be agent-neutral, reflecting no particular individual's perspective on the world. I assume that our axiology has some bearing on what we ought to do or are morally required to do. I don't assume that nothing else so bears. In other words, I don't assume *consequentialism* (Mòzǐ 5th-3rd century BCE [2020]; Newcome 1728 [1732]; Bentham 1780 [1823]; Mill 1863; Moore 1903; Sidgwick 1906 [1981]). Nonetheless, I set aside all normative and deontic questions that are independent of or downstream from the axiological issues I'll cover.

We can think of our axiology as telling us, in the first instance, whether a given possible world is at least as good as another. Exactly what possible worlds are is a matter of philosophical controversy (Menzel 2021). Very roughly, we can think of a possible world as a complete history of everything. Note, however, that our question isn't whether it has *all* been worth it – past, present, and future (compare Parfit 2011: 920–921). For example, our aim isn't to consider whether a world with the actual world's past and a certain predicted future is better than a world in which nothing of value or disvalue ever exists. Thus, I won't address the plausibility of Schopenhauer's colourful claim that "it would have been much better if the sun had been able to call up the phenomenon of life as little on the earth as on the moon" (Schopenhauer 1850 [1970]: 11). Instead, we're to compare possible worlds that share the actual world's past - including terrestrial abiogenesis - but differ in respect of whether human beings go extinct at or before some future time t.

2.2 Up To What Time?

What time should we choose to be t?

Before we answer, here's one important thing to bear in mind. In considering the value of the world given that human beings don't go extinct at or before t, we don't assume that they then go extinct immediately thereafter. Rather, we assume that humanity survives until t and then ask how well the future can be expected to go given that assumption. Up to t, we take human survival for granted. Beyond t, our answer will be determined by our beliefs about how likely humanity is to survive given that it has survived until t. A near-future t is therefore compatible with a very long expected total duration.

There are a number of reasons we might choose t to be a date in the not-too-distant future. We have more control over humanity's survival in the near term than in the long term. We may also think the current era is historically unique, considered from the perspective of extinction risk (Sagan 1994: 172–179; Parfit 2011: 923; Ord 2021). The development of nuclear weapons, as well as other technological powers still on the horizon, gives us historically unprecedented powers to destroy ourselves. We don't yet have the civilizational wisdom, institutional safeguards, or distributed space settlements needed to survive long-term with those powers at our ready. Ord (2021: 31) argues on this basis that we face "an *unsustainable* level of risk. Thus, one way or another, this period is unlikely to last more than a small number of centuries."

A natural question, then, is what happens if we make it through this period. Ord (2021: 240) argues that beyond our current bottleneck lies "something extraordinarily vast and valuable – something in light of which all of history thus far will seem the merest prelude; a taste; a seed." But like Parfit's remarks quoted in my introduction, Ord's are about what we *could* achieve. His discussion is about "potential, not prophecy." (Ord 2021: 217) Pessimists might agree on our potential, but prophesy it will be dashed.

2.3 The Human Condition

We're asking about the value of continued human survival. So what does it mean to be *human*, in the sense that interests us? When we look back into evolutionary time, we find organisms - Neanderthals, Denisovans, Australopithecines - that we might not feel sure whether to count as human. When we look to the far future, things only get worse.

Imagine a future in which no human organisms remain, but there are billions upon billions of emulated minds running on computer hardware, derived originally in the (by then) distant past through scanning members of the species *Homo sapiens.* Suppose each of these minds retains a psychology and a virtual body much like ours. Aren't these people therefore human, in some meaningful sense (compare Kurzweil 2005: 374, Ord 2021: 236–239)? This outcome will certainly strike many of us as very different in value from the sort of thing we typically imagine as human extinction. This may suggest that we shouldn't be focusing on the extinction of some particular biological taxon, like the species *Homo sapiens* or the genus *Homo*. But then what exactly is it whose survival is at issue?

How we answer this question will depend ultimately on our ethical commitments and the way they shape our interest in the question of the value of continued human survival. For example, suppose we emphasize the moral importance of conserving existing things of value as the key factor explaining the moral importance of preventing human extinction. Then we'll presumably find it natural to adopt a definition that focuses on mental and physical similarity to currently existing people, and to count various post-human scenarios in which our descendants become radically alien as outside the circle (compare Cohen 2012: 149–152).

Others might think that our concern for the survival of humanity ought ultimately to be about the quality of the values that govern the future of our planet and the cosmos beyond, regardless of their similarity to our own. The values that we have now are no doubt badly mistaken in various ways (Williams 2015). Nonetheless, working from this foundation might offer the best chance of a morally good future. The alternatives are to cede posterity to the cold indifference of Nature or the values of some successor species that emerges to gain technological mastery over the living world, perhaps even at our expense. Neither might inspires much confidence. Viewed in this perspective, we nonetheless shouldn't care about the overall similarity of the bodies and minds of future people to our own, provided that they become alien for the right reasons. We might thus prefer to define which future individuals are to count as human in terms of normative constraints on the ratifiability of the transitions by which you get from us to them, focusing especially on changes in values.

The extent to which we need to grapple with these questions also depends on how we set t. The nearer to the present time that t is chosen to be, the less likely we are to face hard cases and the more likely is it that for any more concrete interpretation of 'humanity,' the extinction of humanity and the extinction of *Homo sapiens* coincide.

2.4 Non-Humans

I've framed our question in a way that might seem anthropocentric. The question is about the value of the future, given human survival up to time t. This obviously puts a lot of emphasis on *us*. Nonetheless, the question is emphatically *not* about the value of the future considered only in terms of how things stand with respect to future humans. Any and all morally statused beings who may exist contribute to the value of the future.

Roughly speaking, something counts as having *moral status* (or *moral standing*) insofar as they and/or their interests matter morally in their own right.[1] Many philosophers accept the view that *sentience*, conceived as a capacity for positively and/or negatively valenced conscious experience, is a necessary and sufficient condition for moral status (Singer 1993; DeGrazia 1996; Sebo 2018; Shepherd 2018). Some hold that consciousness on its own may suffice, even in the absence of a capacity for valenced feelings (Chalmers 2022: 339–345). Still others maintain that it is possible for an individual to have intentional states in the absence of phenomenal consciousness, and that preferences in the absence

[1] Note that on some views, there may exist things that matter morally in their own right by virtue of being intrinsically valuable, but which don't have interests in any meaningful sense. For example, this might be true of ecosystems, as suggested by Cahen (1988), as well as perhaps species and other objects of characteristic concern among environmental holists. On other views, individuals may not matter in their own right, but their interests do. For example, this may be true on some interpretations of a total utilitarian ethics (see Singer 1993: 121).

of phenomenology may suffice for moral status (Carruthers 2019: 171–174; Kagan 2019: 16–23). A minority view holds that organisms without mental lives also generally have moral standing (Goodpaster 1978; Taylor 1981; Varner 1998).

A further question concerns whether there is a hierarchy of moral statuses. Roughly speaking, this is the question of whether harms and benefits that are otherwise similar are to be weighted differently for different morally statused agents. For example, should the pain of a mouse be counted the same as that of a man if the pains are of the same intensity and duration and are otherwise similar in morally relevant respects? According to one popular view, the answer is 'Yes': pain is pain, and there are no differences of degree in moral status (Singer 1993; DeGrazia 1996; Rachels 2004). Many others defend a view on which moral status is graded and some individuals have a higher moral status than others (Nozick 1974: 35–42; Jaworska and Tannenbaum 2014; Kagan 2019).

For simplicity, my discussion is framed in a way that assumes no differences in moral status, at least as applied to the good-making properties of outcomes as instantiated by morally statused individuals. Since my discussion is focused on axiology, I arguably have no need to take a stand on whether there might be differences in moral status constituted by differences in the deontological constraints that protect individuals against being harmed for the sake of the greater good, as suggested by Nozick (1974: 39). It should nonetheless be easy enough to see how my conclusions would need to be adjusted given various hierarchical views.

Although I aim to be fully inclusive of all morally statused beings, it bears emphasis that the focus is nonetheless specifically on human extinction. Thus, our question does not concern, say, the extinction of all sentient life, except insofar as we should expect all sentient life to die out alongside humanity. There are clear cases of risks to continued human survival that should not be expected to have significant spill-over effects of this kind. In particular, human extinction resulting from a bioengineered pandemic is unlikely in general to pose an extinction risk to distantly related animals, given that genetic distance inhibits cross-species transmission of pathogens (Faria et al. 2013). According to Ord (2021: 167), the risk of a bioengineered pandemic makes up fully one fifth of the total risk of e over the 21st century.

Where non-human individuals feature in the discussion that follows, I focus almost exclusively on non-human animals of the kind that currently make up the vast majority of morally-statused individuals. Non-human animals have always made up the vast majority of morally statused individuals, of course, and we might reasonably project that they will continue to do so. Nonetheless, some may believe that non-human animals - and indeed human animals - make a negligible contribution to the expected value of continued human survival, which depends instead in large part on how things go in respect of possible future digital minds, which may not especially resemble the kinds of minds with which we are familiar. Minds that arise through software engineering rather than biological evolution and run on digital hardware may be thought to have the potential to make extremely positive or extremely negative contributions to the value of all future history by virtue of their ability to be inexpensively and reliably copied and to inhabit regions of the space of possible minds that are closed to biological sentience (Shulman and Bostrom 2021). Others may be inclined to dismiss that point of view as too speculative or outlandish.

My own focus is less on empirical prediction, and more on ensuring that we have the right evaluative framework. In that sense, I need take no stand in this debate. My reason for focusing on non-human animals is as follows. Precisely because possible abiological minds of the future might inhabit very alien regions of the space of possible minds, it is very difficult to reason in advance about what they will be like, and easy to get lost in conceptual fog. My hope is that by reasoning about the possibilities for good and bad lives available to the many and varied biological minds that now exist and that we understand reasonably well, we are doing the best we can do presently, from the perspective of normative theory, in gaining a more concrete understanding of how creatures with different kinds of mental lives might contribute to our evaluation of the prospect of continued human survival.

2.5 The Opposite of Extinction

Someone might argue that, for any t, if humanity is rendered extinct at or before t, then it would have been little better, and probably worse, had humanity not gone extinct then, since the closest possible counterfactual scenarios are ones in which the same extinction cause operates so as to only slightly delay the time of extinction. Therefore, human extinction at any given point in time can never be a catastrophic tragedy.

The problem posed here is analogous to a well-known problem in accounting for the harm of death (McMahan 2002: 107–117). The most widely accepted view maintains that death at t is a harm to the deceased to the extent that her life would have been better had she not died at t (Nagel 1970; McMahan 1988; Feldman 1991; Broome 1999: 170–173, 2004). But consider a young person who is killed when she absentmindedly steps off the curb into the path of a bus (McMahan 2002: 111). In thinking about how this person's life would have gone were she not killed that day, we could try to hold constant the crash that killed her and imagine that she did not die on that day but sustained severely debilitating injuries, to which she succumbed a few weeks later. This would suggest, counterintuitively, that her death in the collision that day was not especially tragic.

McMahan (2002: 107–117) proposes that in assessing the harm of death, we are to try to consider a counterfactual scenario where the event that caused the death at t is completely absent, rather than falling just short of being fatal. At the same time, we try to hold as much of the victim's actual life up to t constant. Plausibly, we should adopt the same kind of approach when thinking about human extinction. When considering the possibility that extinction at or before some time t may occur due to some cause or set of causes, the counterfactual scenario in which extinction is avoided is constructed by attempting to mentally remove the relevant causal factors as completely as possible, while holding constant as much of our actual history.

Note that we thereby close much of the gap that might otherwise seem to exist between thinking about *extinction risk* and *existential risk*. As defined by Bostrom (2013: 15), an existential risk is one that "threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development." Many of the catastrophe scenarios that satisfy the second but not the first disjunct in Bostrom's definition arguably end up being assumed away if we screen off not only outright extinction but also possible cases of civilizational collapse that might occur if the various extinction risks that now concern us materialize but prove unfatal.

In trying to assess how well someone's life would have gone had they not been killed at t, we sometimes face the problem that there is no way of holding the person's past substantially fixed while completely removing the cause of death. Consider someone who dies from a long-standing illness that significantly decreased their quality of life for many years prior. Since it's hard to know how to construct the relevant counterfactual scenario in this sort of case, it's hard to know how to formulate an unambiguous assessment of their death considered as a dated event (McMahan 2002: 113). For a possible parallel, consider risks from climate change. Extinction risks associated with extreme climate scenarios are poorly understood at present (Kemp et al. 2022). What is clear is that in bracketing the possibility of extinction as a result of climate change by assuming continued human survival up to some time t, it would be a mistake to focus only on scenarios in which there is no warming above pre-industrial levels at all or no more warming beyond the level of $1\,^{\circ}C$ above pre-industrial levels reached in 2017. This would clearly involve too great a revision in our actual history. Total committed emissions from existing and planned fossil fuel infrastructure already put the Paris Climate Agreement target of $1.5\,^{\circ}C$ warming out of reach (Lynas 2020: 270–273). It is much harder to say exactly what kind of scenario would represent a reasonable compromise between mentally removing the relevant extinction threat to the greatest extent possible while holding fixed our history and the extent of warming to which we are already committed.

2.6 Uncertainty

The survival of humanity is a kind of gamble. It could go well, and it could go very badly. Therefore, we need some way of representing our uncertainty, and some way to value uncertain prospects. The simplest case is that in which we have rational beliefs about the future that are capable of being represented by a probability function, $\Pr(\,\cdot\,)$, and our theory of value can be represented by a cardinal value function, $V(\,\cdot\,)$, allowing us to compare the expectation of this value function relative to the supposition that humanity goes extinct at or before t to the expectation given that humanity survives beyond t.

We need not assume at this stage that the right way to value uncertain prospects is in terms of their expected value, even granting that an ordering in terms of expected values is possible. We need only take on board the point that answering our question requires some way of valuing uncertain prospects.

All this should hopefully be uncontroversial. A more controversial assumption that I'll make is that we should also want our assessment of the value of the future to take account of our uncertainty about the correct value function - and even our uncertainty about the right way to value uncertain prospects. In doing so, we can draw on the recent literature on moral uncertainty (Lockhart 2000; Ross 2006; Sepielli 2009; Greaves and Ord 2017; MacAskill et al. 2020; Riedener 2020) and decision-theoretic uncertainty (MacAskill 2016; MacAskill et al. 2021). Some argue that uncertainty of this kind doesn't bear on what we ought to choose in any meaningful sense (Weatherson 2014; Harman 2015; Hedden 2016). If that's your view, feel free to skip the final part of the paper.

3 Population Axiology - Evaluating Extinction

We want to know whether and to what extent the continued existence of humanity is morally desirable. In addressing this question, we inevitably have to engage with questions of *population axiology* - questions about how to value outcomes that differ in the size and composition of the population (Greaves 2017). This is notoriously hard (Parfit 1984: 419–442; Ng 1989; Arrhenius 2000). As a result, it's difficult to be confident in any one theory, and natural to look across a range of theories. That's the goal of this section.

In section 3.1, I briefly consider a view with an especially infamous relationship to the desirability of extinction: *negative utilitarianism*. Section 3.2 considers a different view that bears important similarities to negative utilitarianism: *the procreation asymmetry*. I explain why the procreation asymmetry does not support the desirability of extinction in the way it is often assumed to do. In section 3.3, I explain how the same reasoning points to a flaw in Benatar's argument for the desirability of extinction. The next three sections cover *total utilitarianism* (3.4), *critical level utilitarianism* (3.5), and *average utilitarianism* and *variable value theories* (3.6). Section 3.7 is on *prioritarianism* and *egalitarianism*, where I primarily consider an argument that the value of equality counts for the desirability of human extinction. Section 3.8 is on *perfectionism*. Lastly, in section 3.9, I consider *conservatism about value* and explain why conservatism provides less support for the desirability of continued human survival than has been claimed.

3.1 Negative Utilitarianism

No theory of value bears a more infamous and apparently obvious relationship to the desirability of human extinction than *negative utilitarianism*. Popper (1961 [2011]: 602) claims that "pain cannot be outweighed by pleasure" and that our goal should be "the least amount of avoidable suffering for all". These remarks may be interpreted as follows.[2] We assume that welfare has both a positive aspect and a negative aspect. Call the former *well-being* and the latter *ill-being*. Negative utilitarianism says that one outcome, x, is worse than another outcome, y, if the sum of ill-being is greater in x.[3] [4]

The best-known objection to negative utilitarianism is that it entails the moral desirability of human extinction - or, at least, that it makes the desirability of human extinction easier to derive than it intuitively ought to be. Smart (1958) asks us to imagine a weapon capable of instantly and painlessly killing every human being. He argues that since everyone who could be killed painlessly in this way at any given time would experience some suffering if they continue to live, negative utilitarianism counts the painless death of every human being now living as better than their continued existence.

Stated in these terms, the argument isn't strictly valid. It neglects non-human welfare subjects (Knutsson 2019). Human extinction would minimize the sum of ill-being experienced by human beings, but can't be assumed to minimize the sum of ill-being considered in full generality without further argument. What Smart should have said, arguably, is that negative utilitarianism entails the desirability of the extinction of all welfare sub-

[2] Note that Popper (1961 [2011]: 501) would subsequently forswear this interpretation. Thanks to Jacob Barrett for bringing this passage to my attention.

[3] Consistent with Popper's remarks, we may also count x as worse than y if the sum of ill-being is equal in x and y, but the sum of well-being is greater in y. Thanks to Hilary Greaves for this observation.

[4] Note that negative utilitarianism may also be understood as a theory of right action on which the right act is that which minimizes the sum of ill-being. For present purposes, I focus simply on the negative utilitarian theory of value.

jects.[5] It doesn't obviously entail the desirability of human extinction except as part of such an event.

In any case, negative utilitarianism is not a view taken seriously by many contemporary philosophers. Nonetheless, there is something intuitively compelling in Popper's central idea: that suffering demands a moral response of a kind that mere absence of pleasure does not. As we'll see, many views that are taken seriously by contemporary moral philosophers incorporate some version of this idea.

3.2 The Procreation Asymmetry

A well-known view in population ethics with similarities to negative utilitarianism is the *procreation asymmetry*. In order to formulate the procreation asymmetry, we need the assumption that the set of possible lifetime welfare levels has a privileged zero point, above which lives are worth living, and below which they are not worth living.[6] The procreative asymmetry can then be understood as the view that whereas adding lives that are not worth living to the population makes the outcome worse, *ceteris paribus*, adding lives that are worth living to the population does not make the outcome better (or worse), *ceteris paribus* (McMahan 1981, Roberts 2011). Just as negative utilitarianism attaches no value to well-being, but only disvalue to ill-being, the procreation asymmetry attaches no value to additional lives in which well-being predominates, but only disvalue to additional lives in which ill-being predominates.

The procreation asymmetry is held to be intuitive, but has also been claimed to favour extinction (Sikora 1978; Rachels 1998; Holtug 2004). Holtug (2004) asks us to imagine that we can either choose to carry on the human race or let it go extinct by having no children. The procreation asymmetry supposedly entails that it would be better for those in this position to allow the human race to go extinct, "because, among the billions of people

[5] In fairness, he comes very close. While Smart (1958: 542) does pose his objection to negative utilitarianism in terms of "a weapon capable of … destroying the human race", he also describes whoever wields his imagined weapon as a "world-exploder" and suggests that, given negative utilitarianism, such a person is "the saviour of mankind, and for that matter of the animals too".

[6] See Arrhenius (2014: 24–34) for a survey of attempts to rigorously define the zero level.

they could cause to exist, there would surely be a few … who would be miserable; and while their misery would count against their being created, the happiness of the rest would count for nothing.” (Holtug 2004: 139)

As with the argument that negative utilitarianism entails the desirability of human extinction, this argument might be criticized for neglecting non-human welfare subjects.[7] Even apart from this, it is unsound. The desirability of extinction follows from the procreation asymmetry only if we assume that a bad thing plus a neutral thing adds up to a bad thing. While intuitive, it's well-known that those who defend the procreation asymmetry have powerful independent reasons to reject this assumption and posit that the neutrality of additional good lives is ‘greedy,’ being “able to swallow up bad things and neutralize them.” (Broome 2005: 409)

Here's why. The procreation asymmetry entails *the principle of neutrality*, which says that one population that differs from another only in that it involves any number of additional lives that are all worth living is not better or worse than the status quo population. This could be taken to mean that the smaller population is equal in value to the larger. This leads to very bad results, as demonstrated by Broome (2005).

Consider the following possible outcomes, consisting of one or two people - Afryea and possibly also Beom-seok - where a numerical value in a cell denotes a person's lifetime welfare level in the corresponding outcome, whereas ‘Ω’ denotes their non-existence:

	Afryea	Beom-seok
a	5	Ω
b	5	5
c	5	6

Table 1

[7] Note, however, that Holtug (2004: 139) asks us to imagine the choice described in the previous paragraph as one confronted “sometime in the future, [by] the last few inhabitants of earth”. Thus, it may be that in this scenario we are to imagine that there exist only those human beings who must choose whether to carry on their species.

If we interpret the principle of neutrality to mean that the smaller population is exactly as good as the larger when these populations differ only in that the larger involves some number of additional lives that are worth living, then b is exactly as good as a and a is exactly as good as c. Since 'exactly as good as' is transitive, b is exactly as good as c. Since c and b have the same population and c is at least as good for everyone and strictly better for someone, this is very hard to believe.

A more plausible interpretation of the principle of neutrality treats it as the view that one population that differs from another only in that it involves any number of additional good lives is neither better than, worse than, nor exactly as good as the smaller population. Since 'is neither better than, worse than, nor exactly as good as' isn't a transitive relation, the argument of the previous paragraph is defused.

However, Broome notes that when the principle of neutrality is interpreted in this way, we end up forced to conclude that one outcome that differs from another in two respects - one bad, the other neutral - need not be worse, all things considered. Consider the following possible outcomes for Afryea and Beom-seok:

	Afryea	Beom-seok
d	5	Ω
e	5	1
f	4	4

Table 2

Plausibly, f is better than e. This follows from *non-antiegalitarianism*, according to which, if the same people exist in x and y and y is perfectly equal with higher total (and so higher average) welfare than x, then y is better than x, *ceteris paribus* (Ng 1989). The principle of neutrality entails that e is not worse than d. It follows that f is not worse than d. This is surprising: f differs from d in just two respects, one of which is bad (the loss to Afryea of one unit of welfare) and one of which is assumed to be neutral (the addition of Beom-seok with a life worth living).

A priori, one might have thought that a bad thing plus a neutral thing must add up to a bad thing. Frick (2017) calls this the *decomposition principle.* Anyone who endorses the procreation asymmetry, and *ipso facto* the principle of neutrality, has strong independent reason to reject the decomposition principle. Broome thinks the decomposition principle is compelling, and rejects the principle of neutrality.[8] Many of his respondents think the decomposition principle may reasonably be jettisoned instead (Rabinowicz 2009; Frick 2017; Nebel 2019).[9] For present purposes, we can argue as follows. The argument that the procreation asymmetry entails the desirability of extinction fails, since it implicitly relies on the decomposition principle, whereas those who endorse the procreation asymmetry have strong independent reason to reject the decomposition principle, since its falsity is supported by the conjunction of the procreation asymmetry and non-anti-egalitarianism.

A more direct argument for the conclusion that the procreation asymmetry doesn't entail that it's bad to add good and bad lives in any mixture can be made by a slight modification of the argument by which Broome derives the conclusion that neutrality is 'greedy' (Francis 2019). Consider

[8] Broome also argues against the principle of neutrality on the grounds that if the value of additional lives worth living is 'greedy,' then we are not able to capture the intuitions taken to support it, since we are not typically able to ignore potential changes in population size when deciding how to act, given the ability of changes in the number of lives worth living to cancel out other morally significant changes we might bring about.

[9] In recent work, Frick (2022) adopts a different view, relying on the assumption that the betterness ordering over outcomes may be relative to choice sets in order to reject the conclusion that f is not worse than d. On his revised view, d is better than f when the choice set is $\{d, f\}$, f is better than e if the choice set is $\{e, f\}$, and e is not worse than d if the choice set is $\{d, e\}$. However, e is worse than d when the choice set is $\{d, e, f\}$ because f's availability makes e unjust to Beom-seok when e is chosen from the choice set set $\{d, e, f\}$, but not when chosen from the choice set $\{d, e\}$, and "the deontic fact that it is unjust or morally wrong to bring about some outcome *bears on* the axiological question how well the world goes if that outcomes is brought about." (Frick 2022: 249) Note, however, that there is no assumption that $d - f$ are outcomes for anyone to choose, as opposed to outcomes that might arise naturally. Thus, it is not clear how the appeal to choice set dependent betterness can be used to block the argument that f is not worse than d, since there are no choice sets in play here. Moreover, it is hard to think of a reason why f should be considered no worse than d if no one has a say about it, but worse than d if the outocme is up to us. Frick's proposed explanation for why f is worse than d is that everyone who exists in f is worse off than everyone who exists in d (Frick 2022: 236), which is equally true whether or not f and d are available to choose or arise independent of anyone's say-so.

	Afryea	Beom-seok	Csaba
g	5	Ω	Ω
h	5	1	1
i	5	-1	5

Table 3

It's plausible – and, at the very least, consistent with the the procreation asymmetry – to believe that i is better than h: the same people exist, and the loss suffered by Beom-seok seems to be outweighed by the greater gain experienced by Csaba. Since the principle of neutrality entails that g isn't better than h, it follows that g isn't better than i, although i differs from g only in terms of the addition of a life with negative overall welfare and a life with positive overall welfare.

I've argued that the procreation asymmetry doesn't automatically support the desirability of human extinction in the way often assumed, even when considered as part of the extinction of all sentient life. The addition of some number of lives that are not worth living alongside many that are worth living need not yield an outcome worse than that in which none of those lives are ever begun. But does the procreation asymmetry also allow us to say the addition of those many lives can make the outcome better? Is it compatible in principle with the verdict that human extinction might make the world a worse place overall?

Clearly, if we accept the procreation asymmetry, we cannot claim that the continued coming-into-being of new humans is desirable in light of the value of increasing the number of flourishing lives that will ever be lived. Nonetheless, that does not preclude us from identifying other considerations in light of which continued human survival may be viewed as desirable. A natural assumption is that if we can identify such values, then it is consistent with the procreation asymmetry to regard the continued existence of the human species as for the best, provided that we do not expect too many future lives to be so bad as not to be worth living (compare Frick 2017, discussed also in section 3.9).

However, it's not so easy to reach the conclusion that the survival of humanity is all-things-considered desirable in this way, even if we adopt an optimistic forecast of future welfare levels. So far, I've emphasized the idea that lives worth living exhibit 'greedy neutrality' in virtue of their ability to swallow up bad things and neutralize them. But 'greediness' goes both way. Adherents of the principle of neutrality appear committed to the idea that adding a life worth living to the population can also swallow up good things and neutralize them. A minor variation on the case in Table 2 illustrates this possibility:

	Afryea	Beom-seok
j	5	Ω
k	6	1
l	5	5

Table 4

By the principle of neutrality, l is not better than j. Non-anti-egalitarianism entails that l is better than k. It follows that k cannot be better than j, although k differs from j in just two respects, one of which is good (the gain to Afryea of one unit of welfare) and one of which is assumed to be neutral (the addition of Beom-seok with a life worth living).

It follows that if we accept the procreation asymmetry, we may well have trouble arguing for the desirability of the survival of humanity even if we can point to values that speak in favour of a continued human presence and even if all future lives are well worth living. The neutrality associated with the addition of worthwhile lives to the population could conceivably swallow up and neutralize those other values, forcing us toward the conclusion that extinction would not be worse than continued human survival.

3.3 An Aside on Benatar

Reasoning similar to that on which I relied in the previous subsection to contest the claim that the procreation asymmetry makes extinction desirable can also be used to identify a flaw in Benatar's philosophical argument that coming into existence is always a net harm, which forms the centrepiece of his widely discussed case for anti-natalism (Benatar 2006). This should be unsurprising, since Benatar's argument rests on an intrapersonal analogue of the procreation asymmetry and is motivated in part by its supposed ability to explain judgments congruent with the procreation asymmetry (Benatar 2006: 31–36).

I first explain Benatar's views about the harms and benefits of coming into existence. Suppose that m and n are two different outcomes, and that the table below shows Csaba's well-being in these outcomes at two different times, t_1 and t_2, with ω denoting Csaba's non-existence at a given time:

	t_1	t_2
m	ω	ω
n	5	-1

Table 5

Thus, in n, Csaba, lives for two periods. The first is purely blissful and involves no pain or frustration of any kind, whereas his only experience in the second is one of mild suffering. In m, he never lives at all, foregoing the bliss he would otherwise have experienced in the first period and the minor suffering he would have endured in the second. According to Benatar, n will be worse than m for Csaba in respect of the fact that he experiences suffering in the second period, but it won't be better for him in respect of the fact that he experiences happiness in the first period. The absence of happiness in the first period in m is neutral, Benatar claims, as opposed to a respect in which m would be worse than n. Benatar's claim is that a foregone pain is good even in worlds where the subject of the pain doesn't exist, but a foregone pleasure isn't bad when the subject of the pleasure doesn't exist.

From these principles, Benatar concludes that n is worse for Csaba, all things considered. Since anyone's life will involve some suffering, a similar argument establishes that it's better for anyone never to have been born, even if their life also involves great happiness. Therefore, "it would be better, all things considered, if there were no more people (and indeed no more conscious life)." (Benatar 2006: 164)

For present purposes, I set aside any assessment of the plausibility of Benatar's claims that n is worse than m in some respect and not better in any (see Harman 2009; Bradley 2010; DeGrazia 2010; McMahan 2013). Instead, I'll just focus on Benatar's inference from these premises to the conclusion that *n* is worse all things considered.

Benatar relies implicitly on the decomposition principle in making this inference. Thus, he tells us that “it is always preferable not to come into existence” since “coming into existence has disadvantages relative to never coming into existence whereas the positive features of existing are not advantages over never existing.” (Benatar 2006: 48) In other words, Benatar assumes that a bad thing plus a neutral thing must add up to a bad thing. This is just the decomposition principle, which we know we need not accept.

What’s more, it’s possible to show that Benatar himself has strong reason to deny the decomposition principle, at least given extremely plausible assumptions about acceptable trade-offs within a life. Consider

	t_1	t_2
o	ω	ω
p	1	1
q	5	-1

Table 6

In *p*, Csaba’s life involves no suffering at all at any time and a mild buzz of pleasure at each period of his existence. Benatar says that a life like *p* would be no worse for a person than never having been born: “About such an existence I say that it is neither a harm nor a benefit and we should be indifferent between such an existence and never existing.” (Benatar 2006: 29) However, it seems very plausible that *q* is better for Csaba than *p*. The harms suffered in the second period seem to be more than compensated for by the greater benefits in the first period. In that case, *q* can’t be worse for him than *o*, contrary to what Benatar claims. To avoid this conclusion, Benatar would seem to have to deny that it can ever be better for someone to undergo pain at one time for the sake of greater happiness at another.

In summary, just as the procreation asymmetry does not support the desirability of human extinction unless we assume the decomposition principle, whereas we have good reason to reject the decomposition principle insofar as we accept the procreation asymmetry, so Benatar's claims about the disadvantages and advantages that coming into existence has and does not have relative to non-existence do not support Benatar's anti-natalist conclusions. They would do so only if we assume the decomposition principle, which Benatar's assumptions give us good reason to reject. Even setting aside philosophical criticisms of those assumptions, we should conclude that they do not support the claim that it would be better if there were no more human beings or no more sentient life at all.

3.4 Total Utilitarianism

The next three sections consider a range of well-known utilitarian population axiologies that all reject the principle of neutrality, beginning with total utilitarianism. For two outcomes, x and y, with $N(x)$ and $N(y)$ individuals, respectively, if $u_i(x)$ denotes the lifetime welfare of individual i in x ($i = 1, \ldots, N(x)$) and $u_j(y)$ denotes the lifetime welfare of individual j in y ($j = 1, \ldots, N(y)$), then, on total utilitarianism, x is at least as good as y if and only if

$$\sum_{i=1}^{N(x)} u_i(x) \geq \sum_{j=1}^{N(y)} u_j(y) \tag{1}$$

In other words, we compare the sum of each individual's welfare within the two outcomes.

Total utilitarianism is often taken to justify assigning special moral importance to reducing the risk of human extinction (e.g., Bostrom 2003; Beckstead 2013; Greaves and Ord 2017). It tells a very straightforward story about why it should be good for there to be more flourishing lives - and better and better the more such lives will be lived - since additional lives that are worth living always make the outcome better and always do so to the same extent, no matter how big the population is already imagined as being. Assuming the number of our potential descendants to be vast and their lives reasonably good on average, if everyone now living were to die, the loss of value would be extraordinary - and the loss of all those billions of people now living would be only a very small part of it. Most of the value lost would be due to the fact that vast multitudes of people who would otherwise have enjoyed worthwhile lives will never be born (compare Parfit 1984: 453–454).

Much of the literature on population ethics focuses on counter-intuitive implications of the total utilitarian population axiology, especially the fact that it appears to entail the infamous *repugnant conclusion* (Parfit 1984). The repugnant conclusion states that for any population, A, of lives with very high welfare levels, there is a population, Z, of lives that are barely worth living such that Z is better than A. Like Parfit, many find this conclusion very hard to accept. But are trade-offs between lives of the kind that feature in the A and Z populations mentioned in the statement of the repugnant conclusion especially relevant in thinking about the value of continued human survival and the way total utilitarianism bears on its assessment?

They might well be. The argument sketched above for treating total utilitarianism as favouring continued human survival lays itself open to a charge of anthropocentrism. It focuses only on the good that would be lost via human extinction due to the fact that very many potential human lives would fail to contribute their happiness to the sum total. It ignores the possibility that all those human lives might displace far more non-human lives. As noted previously, scepticism about the value of continued human survival is especially likely to arise from concern for the negative impact of human beings on non-human animals and the natural world. Terrestrial vertebrate species have seen a mean decline of 28% in numbers over the last forty years, and, on the same time-scale, marine vertebrates have declined in abundance by 22% on average (Dirzo et al. 2014; McCauley et al. 2015). Alongside our treatment of factory farmed animals, these are among the kind of considerations highlighted by May (2018) in making the case that we should take seriously the desirability of human extinction.

Is there a case to be made for the undesirability of the continued existence of the human species in light of the fact that human beings so thoroughly decimate wild animal populations? Such a case will be hard to make if the members of the wild animal species that suffer population declines due to human activity do not have lives worth living, such that we should wish for their sake that they had never been born. (More on this later.) Assume, therefore, that their lives *are* worth living. Someone might then claim that the choice between a large human population and the much larger population of wild animals that may be projected to exist if we as a species were to go quietly extinct involves one important aspect of the choice between an A population and a Z population of the kind that feature in the repugnant conclusion, in that it partly involves a choice between a large population of individuals with high lifetime welfare levels and a much larger population of individuals whose lives are very near the neutral level.

Many human lives are very good. Globally, around 32% of human beings report that they are very happy, rising as high as 67% in Mexico and 64% in Uzbekistan (Inglehart et al. 2014).[10] Human lives may be thought of as ennobled by higher goods inaccessible to most or all non-human animals, such as autonomy, artistic creativity, and scientific understanding. The lives of non-human animals are typically much shorter than those of human beings and are lived in very hard conditions, characterized by scarcity of food and water and vulnerability to predation as normal parts of life (Tomasik 2015). Callicott (1980: 336) goes so far as to conjecture "that the pain and suffering of research and agribusiness animals is not greater than that enduring by free-living wildlife as a consequence of predation, disease, starvation, and cold". An individual whose lifespan, cognitive abilities, and affect balance are like those of a typical non-human animal living in the wild might be thought to have a life scarcely better, if at all, than a life like that originally imagined by Parfit (1986: 148) for the Z population of the repugnant conclusion, a population in which "people ... never suffer; but all they have is muzak and potatoes."[11]

The overall picture sketched here is obviously contestable. Self-reported happiness among human beings may be thought exaggerated (Benatar 2006: 64–69; Haybron 2008: 199–224). The claims made above about the welfare levels of wild non-human animals might seem overly pessimistic. I also haven't taken account of the possibility that shrinking wild animal populations may reduce the expected human population over all time due to the threat of ecosystem collapse (Dirzo et al. 2022). It might also be claimed that the total population of wild animals over all time is greater in expectation given continued human survival because only the technological capabilities of our descendants can safeguard the

[10] Overall, 83.6% of respondents in Inglehart et al. (2014) report that they are either very happy or quite happy.

[11] Notably, Parfit (2016: 118) later asks us to imagine the Z population mentioned in the statement of the repugnant conclusion as composed of animals that have only "enough slight pleasures like those of cows munching grass or lizards basking in the sun."

biosphere against total destruction due to the brightening of the Sun in roughly a billion years' time (Ord 2021: 218–223). Last but not least, I've made no serious effort to compare how much animal welfare really is lost as a result of defaunation and whether its sum total really is greater than the welfare maintained by an industrialized human civilization of billions. Nonetheless, it seems far from absurd to think that questions about the value of continued human survival that arise in response to the current defaunation crisis invoke comparisons of the kind that are at play in the repugnant conclusion.

On the other hand, Thomas (2018) and Nebel (2022) have recently brought renewed attention to the idea that total utilitarianism, suitably interpreted, need not entail the repugnant conclusion (compare Portmore 1999; Kitcher 2000; Carlson 2007). Total utilitarianism ranks outcomes by comparing $\sum_{i=1}^{N(x)} u_i(x)$ and $\sum_{j=1}^{N(y)} u_j(y)$. We have assumed implicitly that $u_i(x)$ and $u_j(y)$ are scalar quantities. Suppose instead that they are vectors of real numbers of the form (a, b). We let a be a measure of the dimension of an individual's welfare corresponding to something like Mill's 'higher pleasures' and b be a measure of the dimension of that individual's welfare corresponding to the 'lower pleasures'. Vectors can be added in the standard piecewise fashion: $(a,\ b) + (c,\ d) = (a + c,\ b + d)$. They can also be ordered by the standard lexicographic ordering: $(a,\ b) \geq (c\ , d)$ just in case $a > c$ or $a = c$ and $b \geq d$.[12] Thus, vector-valued representations of individual welfare levels are compatible with total utilitarianism.

Let's call a view of this kind *lexical total utilitarianism.* Assume that each dimension of welfare has a zero level. Say that an individual's life is *neutral* if $a = 0$ and $b = 0$, *barely worth living* if $a = 0$ and $b > 0$, and *good* if $a > 0$. So understood, lexical total utilitarianism avoids the repugnant conclusion, in that a population with good lives corresponding to the welfare level represented as $(100, 0)$ is better than any population with lives barely worth living corresponding to a welfare level represented as $(0, \epsilon)$ for any $\epsilon > 0$.

Lexical total utilitarianism also turns out to violate a principle that bears in a different way on the potential desirability of human extinction, given the harms we inflict on non-

[12] Nebel revises the standard lexicographic ordering in order to avoid certain objections. On the revised ordering, there exist $\Delta, \delta > 0$, such that for quantities of welfare $(a,\ b), (c,\ d)$, we have $(a,\ b) \geq (c,\ d)$ just in case (i) $a - c > \Delta$, (ii) $a \geq c$ and $b \geq d$, or (iii) $a \geq c$ and $(a - c)/(d - b) > \Delta/\delta$.

human animals. The principle I have in mind says, roughly, that the value of good lives can always be outweighed by the disvalue of sufficiently many lives with negative lifetime welfare levels. More exactly, for any background population and for any population of additional individuals, X, all of whom have positive welfare levels, it holds that, given any negative welfare level, there is some number, n, such that the addition of at least n individuals at that negative level alongside X makes their joint addition to the background population morally undesirable. Call this the *negative balancing principle*.

Lexical total utilitarianism entails the falsity of the negative balancing principle, presuming that we allow the vector elements to span negative as well as positive real values. Consider a null population, $\emptyset$, in which nothing of value exists. Letting the value of $\emptyset$ be represented by the vector $(0,0)$, lexical total utilitarianism entails that for some[13] $\pi > 0$, any ρ, and any $\sigma < 0$, a population of m people at (π, ρ) and n people at $(0, \sigma)$ is better than $\emptyset$ for any n.

The falsity of the negative balancing principle could bear on the desirability of the continued existence of human beings in the following way. Suppose we reject the negative balancing principle and hold that there are some lives so good that their addition to the population can justify the addition of any number of lives that are only barely not worth living. It does not seem altogether unreasonable to consider that some human lives may stand in this relationship to a large fraction of the intensively farmed animals whose poor welfare leads authors like May (2018) to look kindly on the possibility of human extinction. Compare, in particular, the best human lives against lives whose welfare level is like that of the average American broiler chicken (see Norwood and Lusk 2011: 128–31).

The vast majority of terrestrial vertebrates slaughtered for food are chickens. In 2007, nearly 9 billion broilers were produced in the US. Many believe that broiler chickens have lives that are not worth living. Some claim they have "the most wretched lives of all farm animals." (Cooney 2014: 7) Herzog (2010: 156) describes the conditions in which broilers are raised as "Dante-esque: the chicks will never see sun nor sky. Because they are so top-heavy, broiler chickens spend most of their day lying down, often in litter contam-

[13] Given the lexicographic ordering, this is true for any $\pi > 0$. Relative to the ordering described in the preceding footnote, it is true for any π, m such that $\pi > \Delta/m$.

inated with excrement. As a result, many will develop breast blisters, hock burns, and sores on their feet." However, Norwood and Lusk (2011: 131) reject the view that broilers have lives not worth living. While conceding that "broilers lead short, unexciting lives," they conclude that "after reviewing all the obstacles to welfare and the nature of the birds, in our assessment, broiler farms do not cause large-scale suffering." They note that pre-slaughter mortality rates are low among broilers, and lower than most people expect, at just 5%. Furthermore, the birds have ample food and water, as well as litter for scratching.

Someone might not unreasonably conclude that if broiler chickens on average have lives that are not worth living, their lives might nonetheless be only *barely* not worth living. Since the idea of a life that is 'only barely not worth living' is so unclear,[14] there is little point in disputing this exact form of words. The key point is that if we reject the negative balancing principle, it seems open to us to claim that the best human lives are such that their value cannot be outweighed by any number of lives at a negative welfare level corresponding to a not-too-absurd estimate of the average lifetime welfare of the American broiler chicken. This opens the door to a novel way of resisting some arguments for the claim that human extinction may be morally desirable.

In summary, I have noted that whereas total utilitarianism is often treated as able to straightforwardly explain the badness of human extinction by virtue of the fact that it treats more and more good lives as ever better to the same extent, it also has distinctive and potentially surprising implications for how to make tradeoffs between the welfare of human beings and the welfare of the wild animals who would exist but for the continuation of human industrial civilization, or between the welfare of human beings and the sufferings of intensively farmed animals who exist due to human industrial civilization, depending on whether we combine total utilitarianism with a scalar interpretation of individual welfare values that entails the repugnant conclusion, or appeal instead to a form of lexical total utilitarianism designed to avoid that worrisome result.

[14] As, of course, is the idea of a life that is 'barely worth living'; see Cowie (2017).

3.5 Critical-Level Utilitarianism

In this section, I consider a generalization of total utilitarianism, known as *critical-level utilitarianism.* According to critical-level utilitarianism, x is at least as good as y if and only if

$$\sum_{i=1}^{N(x)} (u_i(x) - c) \geq \sum_{j=1}^{N(y)} (u_j(y) - c) \tag{2}$$

In other words, we compare the sum of each person's welfare net of c. Here, c is the so-called *critical-level*: the lifetime welfare level at which the addition of a person to the population brings about an outcome exactly as good as the outcome in which that additional person never exists. If c is equal to the zero level for lifetime welfare, critical-level utilitarianism reduces to total utilitarianism. We might instead prefer to set c to be positive (Broome 2004; Blackorby et al. 2005). Call this view *positive critical-level utilitarianism.*

Although positive critical-level utilitarianism rules out the principle of neutrality, it behaves in a sense like a weak form of the procreation asymmetry, at least in the following respect: it entails that for any two welfare levels equidistant from zero, adding one life at the positive level and one at the negative level is worse overall than adding neither. In that sense, good lives are not as good as bad lives are bad. However, it is also morally bad to add a person at the zero level, and at any positive welfare level less than c, unlike on the procreation asymmetry.

I will also say a little bit about why we might accept positive critical-level utilitarianism, as this will soon become relevant. On the one hand, positive critical-level utilitarianism can be motivated by a desire to avoid the repugnant conclusion in a way consistent with the assumption that welfare is a scalar quantity. Positive critical-level utilitarianism has also been motivated by appeal to the fact that it can give weight to the value of *unfragmented lives* (Blackorby et al. 2005: 151–152) or *longevity* (Broome 2004: 257–9), since positive critical-level utilitarianism counts, say, a long life whose overall lifetime welfare is 10 as better than two proportionally shorter lives, each with a lifetime welfare of 5.

As we'll now see, these motivations can come into conflict in a way that matters when assessing the bearing of positive critical-level utilitarianism on the value of the future.

The conflict can be brought out by noting that it is disturbingly plausible that the vast majority of non-human animals, even granting that they have positive welfare levels, nonetheless have welfare levels below the critical level, at least when the critical level is chosen to satisfy our intuitions about the repugnant conclusion (compare Williamson 2021). Some arguments that may be taken to support that conclusion were reviewed in the previous section. To these we may add the observation that many wild non-human animals adopt life-history strategies that emphasize high rates of reproduction and minimal parental investment, producing offspring in quantities several orders of magnitude greater than the replacement rate, almost none of which survive to reproduce. Some conclude, on this basis, that most wild animals have lives that are not worth living and that suffering predominates in nature (Ng 1995; Tomasik 2015; Horta 2010). These arguments are far from conclusive (Cuddington 2019; Groff and Ng 2019; Browning and Veit 2021). Nonetheless, even if these lives are above the zero level, if positive critical-level utilitarianism is true, they presumably have negative contributory value.

On the other hand, Blackorby et al. (2005) explicitly argue for a species-variable critical level, justifying this in terms of the idea that a positive critical level is intended to give weight to the value of unfragmented lives or longevity. Plausibly, these values are instantiated only to the extent that lives are unified over time via a sense of oneself as a creature with a past and a future. Blackorby et al. (2005: 236) write: “for species whose members have less integrated lives, a lower critical level is reasonable, with a level of zero for species whose members exist in the present only.” Call a view of this kind *species-variable critical-level utilitarianism.*

The problem is that species-variable critical-level utilitarianism entails the repugnant conclusion, at least if we assume that welfare is a scalar quantity. Suppose that we apply a positive critical level to human lives but not to the lives of ants. Then species-variable critical-level utilitarianism entails that for any large population of human beings all with lives at very positive welfare levels, it is possible to choose a larger population of ants whose lives are barely worth living and whose existence is better. If our motivation is to avoid the repugnant conclusion, we have good reason to reject species-variable critical-level utilitarianism. In that case, however, we should be reasonably confident that most non-human animals, even if they have positive lifetime welfare, have negative contributory value given positive critical-level utilitarianism, and that the world would be better without almost everything that moves in the waters, in the air, or upon the earth. Astonishingly, the tendency of human civilization to defaunate the planet would no longer be much of a strike against us, nor would the desirability of our continued existence receive support from claims that we are the biosphere's only hope of surviving on timescales beyond a billion years.

It is not just non-human animals, of course. Possibly, many existing people have lives only barely worth living (see Tännsjö 2002). Looking to the future, consider the scenario described by Hanson (2016), in which, the economy is made up principally of *emulated minds* ('ems') derived by scanning human brains. One striking feature of 'ems' is the ease with which they can be copied at scale. Hanson predicts that in an 'em' economy heavy use will be made of *spurs*, which are very short-lived copies existing for periods of minutes or hours, used to accomplish one-off tasks, such as search tasks that allow for fast parallel exploration of a search space. Hanson (2016: 195) expects that there will be "strong selection for ems who mostly accept sometimes being such spurs, and [we] expect most em work will be done by spurs". While the contours of personal identity become fuzzy in this scenario, it seems plausible that spurs, purely by virtue of being so extraordinarily short-lived, are individuals with lifetime welfare levels below the critical level when the critical level is set at a positive value capable of satisfying our intuitions about the repugnant conclusion.[15] Hanson's projected future therefore looks to be a moral catastrophe given positive critical-level utilitarianism.

In summary, positive critical-level utilitarianism makes it hard to avoid the conclusion that the contributory value of animal lives is largely negative and the future better and better insofar as there are fewer such lives in it. Exactly how the continued existence of human beings impacts on the size of the total population of non-human animals considered over all time therefore becomes important once again, but takes on a rather counterintuitive

[15] In other words, a world of spurs constitutes a *Short-Lived Z* outcome in the sense defined by Portmore (1999).

form. Positive critical-level utilitarianism may also have distinctive implications for the assessment of certain kinds of imagined futures involving digital minds, registering outcomes that might be evaluated very positively by total utilitarianism as tragic instead.

3.6 Average Utilitarianism and Variable Value Theories

I know of no one who treats the promotion of happiness as fundamentally more important than the alleviation of suffering. However, there are some fairly intuitive population axiologies that can behave in practice like critical-level utilitarianism with a negative critical level. Such views may in practice count the addition of good lives as making the outcome better than it would be made worse by the addition of a life not worth living equidistant from the zero level, *ceteris paribus*.

Specifically, we will look at *average utilitarianism* and theories that behave like average utilitarianism in the large population limit. On average utilitarianism, x is at least as good as y just in case

$$\sum_{i=1}^{N(x)} \frac{u_i(x)}{N(x)} \geq \sum_{j=1}^{N(y)} \frac{u_j(y)}{N(y)} \tag{3}$$

In other words, we compare the sum of welfare in each population, divided by the population size. According to *variable value theory* (Hurka 1982, Ng 1989), we instead have that x is at least as good as y just in case

$$g(N(x)) \sum_{i=1}^{N(x)} \frac{u_i(x)}{N(x)} \geq g(N(y)) \sum_{j=1}^{N(y)} \frac{u_j(y)}{N(y)} \tag{4}$$

where $g(\cdot)$ is a strictly concave, strictly increasing function with a horizontal asymptote. In other words, we compare the average welfare level, weighted by a function of the population size whose slope decreases as the population gets bigger and bigger and eventually ends up approximately flat. This function is chosen to ensure that variable value theory agrees approximately with average utilitarianism in evaluating population changes when the background population is large, and approximately with total utilitarianism when the background population is small. By 'the background population,' we have in mind all those individuals whose existence at a given welfare level is independent of the possible population changes under evaluation.

As Tarsney and Thomas (2020) show, given a suitably large background population, average utilitarianism and variable value theory behave in general approximately like critical-level utilitarianism in the evaluation of population changes, when c is set to be the average welfare of the background population. Therefore, if the average welfare of the background population is negative, average utilitarianism and variable value theory behave approximately like critical-level utilitarianism with a negative critical level.

The fact that average utilitarianism can value the addition of lives that are not worth living to the population if their welfare is less negative than that of the background population is well-known. In his *Hell Three* case, Parfit (1984: 422) asks us to imagine that "[m]ost of us have lives that are much worse than nothing. The exceptions are the sadistic tyrants who make us suffer. ... The tyrants claim truly that, if we have children, they will make these children suffer slightly less." Parfit notes that average utilitarianism entails that it would be better if we have children. Most population ethicists take this to be an extremely implausible implication of the view. When some colleagues and I examined non-philosophers' intuitions, we found that when people are asked to compare concrete examples of possible populations, they actually tended to judge it good to have more people with negative welfare when this meant a higher average welfare level, even when the additional people would have lives close to the absolute worst form of suffering imaginable (Caviola et al. 2022). When asked in the abstract, people did tend to agree that it is undesirable to bring into existence new people with lives that are not worth living even if their lives would be less intensely unhappy than those of the pre-existing population, though the mean rating was only 3.31 on a scale where 1 = *Very bad*, 4 = *Neither good nor bad*, and 7 = *Very good*.

Granting that average utilitarianism and variable value theory can in principle evaluate population changes in this way, how do things stand in actuality? As Tarsney and Thomas (2020: 22–3) note, there is a case to be made that there exists a large background population with negative average welfare. That is because the vast majority of the historical terrestrial population is composed of wild non-human animals. While the truth of the matter is hard to decide, there are some reasons to fear that wild non-human animals have negative lifetime welfare on average, as noted previously.

Of course, even if wild animals on Earth have negative average lifetime welfare, the issue remains open, since the terrestrial biosphere may be only a small fraction of the true background population. On the other hand, it may be thought that if there is life elsewhere, then, absent evidence to the contrary, the expectation of its average welfare level should not differ from the average for terrestrial life. However, the average welfare level for Earth-originating life over all time need not be the same as the average welfare level of the terrestrial background population. Keep in mind that the background population includes all those individuals whose existence at a given welfare level is independent of the possible population changes under evaluation. For our purposes, the future terrestrial population is not part of the background population, whereas developments paralleling those that may occur in our future form part of the background population for all suitably distant planets. Thus, setting our credences such that we do not expect the average welfare level elsewhere in the universe to differ from the terrestrial average need not mean that we do not expect the average to differ from the terrestrial background population.

The future, after all, is potentially enormous. The extinction of terrestrial multicellular life is projected at 0.8 – 1.2 billion years from today (Franck et al. 2006). Space settlement, astroengineering, and orbital changes could allow for the survival of Earth-originating welfare subjects way beyond that (Sandberg forthcoming). As members of the only terrestrial genus to have developed a capacity for cumulative technological culture, the future of Earth-originating life could be dominated by our descendants.

On the other hand, we may be especially wary of projecting terrestrial outcomes associated with the existence of organisms with a cumulative technological culture to exobiospheres, in light of the Fermi Paradox (Webb 2002). The observable universe is projected to contain roughly 10^{20} Earth-like planets, with 10^{9} in the Milky Way alone; the Earth is a late-comer among them, being among the last 20% of Earth-like planets to form (Behroozi and Peeples 2015).[16] Nonetheless, we do not find evidence of life elsewhere in the uni-

[16] 'Earth-like' here means having an orbital radius and energy flux similar to that of Earth, allowing for stable

verse. Arguably, this observation requires us to reject the Copernican assumption that the Earth is typical among Earth-like planets. One possibility is that it is atypical in respect of the emergence of organisms with a cumulative technological culture against the background of a biosphere populated by complex, intelligent animals, whose existence is itself not especially improbable, as argued by Powell (2020: 267–278).

In summary, a striking feature of average utilitarianism and variable value theories is that they make our assessment of population changes sensitive to how things stand with respect to the unaffected background population, which may be distant in space and time, requiring us to engage in exobiological speculations of the kind seen above. This is often taken to be a significant drawback of these theories, but as we'll see when we discuss decision theory (see especially sections 4.1 and 4.2), it is also very difficult to avoid in full generality (compare Goodsell 2021). In this section, I've otherwise focused primarily on the idea that if the average welfare level is negative for a suitably large background population, then average utilitarianism and variable value theories agree in practice with a form of critical-level utilitarianism that adopts a negative critical level. As I've noted, it is far from absurd to suppose that the antecedent of this conditional is satisfied. Thus, insofar as we have any sympathy for views of this kind, it need not be absurd to suppose that the continued survival of the human species could be judged as desirable - or at least not *un*desirable - even if the kind of lives in which it results are on average not worth living.

3.7 Prioritarianism and Egalitarianism

Each of the utilitarian axiologies discussed in sections 3.4-3.6 is indifferent to how a fixed sum of welfare is distributed when the population is fixed. In that sense, they are insensitive to the moral significance of distributive patterns.

An alternative that answers to this concern is *prioritarianism* (Parfit 1991; Holtug 2010; Adler 2012). Roughly, this is the view that improvements to a person's welfare are of greater moral value if the person's welfare level is of a lower absolute level. For concreteness, I shall interpret prioritarianism as *total prioritarianism*, the view that x is

surface reservoirs of liquid water.

at least as good as y just in case

$$\sum_{i=1}^{N(x)} f(u_i(x)) \geq \sum_{j=1}^{N(y)} f(u_j(y)) \tag{5}$$

where f is a strictly monotone increasing, strictly concave function with zero as a fixed point. Thus, we compare the sum of a function of each person's welfare, where the function doesn't increase linearly as a person's welfare improves, with its slope instead decreasing more and more as a given individual is imagined as better and better off.

Note that, since f is strictly concave, it is worse to worsen lives that are already not worth living than it is good to improve lives that are already worth living by the same amount. Furthermore, for two welfare levels equidistant from zero, it is worse to add a person at the negative level than it is good to add the person at the positive level. In that sense, lives that are not worth living are bad to a greater degree than lives that are worth living are good (Holtug 2010: 255–256). Thus, prioritarianism supports a weak form of the procreation asymmetry and a more pessimistic attitude toward the value of the future than a view like total utilitarianism, which weights good and bad lives symmetrically. It bears some similarity to the view that "suffering is bad to a *greater degree* than happiness is good" (Mayerfeld 1996: 325).[17]

Prioritarianism's main rival among distribution-sensitive theories is *egalitarianism*. The view on which equality is a feature of outcomes that contributes to their ranking as morally better or worse is the view Parfit (1991) calls *telic egalitarianism*. On the standard telic egalitarian view, it is in itself bad if some people are worse off than others through no choice or fault of their own (Temkin 1993; Segall 2016). For the sake of brevity, I take the qualifier 'through no choice or fault of their own' as implicit in what follows. For con-

[17] See Mayerfeld (1999: 149–158) and Hurka (2010) for further discussion. Arguably, a view of this kind is best understood as a theory of individual welfare: i.e., as the view that pains and pleasures of equal intensity and duration together make a person's life go worse overall. So understood, this view is compatible with any of the axiological theories I discuss in section 3, since what divides them is not their theory of lifetime welfare but the function by which they map the lifetime welfare levels of different individuals to the moral value of outcomes.

creteness, I also assume that the currency of equality is lifetime welfare and that its scope is unrestricted in space and time. My focus will be on the plausibility of the following *egalitarian argument for anti-natalism*, which I take to be orthogonal to these assumptions.

Here is an informal statement of the argument. Inequality is in itself bad. If people continue to be born, there will be many more inequalities. Thus, things would be best with respect to equality if no more people came to exist (compare Temkin 1993: 216–218; Segall 2019: 421–422). Obviously, that is not to say that things would be best *all things considered*, since telic egalitarianism is compatible with and arguably demands a pluralist axiology on which the value of outcomes depends on more than just the level of equality. Nonetheless, the conclusion seems shocking. How plausible is this argument?

The argument is most naturally read as assuming that the disvalue of inequality in a population of n persons is the sum of the pairwise differences between people's welfare levels:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}|u_i - u_j| \tag{6}$$

This ensures that it will inevitably be worse from the perspective of equality if more people come into existence, since it is inevitable that more people will be worse off than others through no fault or choice of their own. Note, however, that the argument does not turn specifically on adopting (6) as our aggregate measure of inequality's badness. Other additive measures yield the same conclusion.

The argument can be challenged by appeal to aggregate measures of the disvalue of inequality that introduce certain kinds of averaging, such as the *Gini Coefficient*. The Gini Coefficient divides the sum of the pairwise differences between people's welfare levels by the square of the population size and by the average welfare level, $\bar{u}$:

$$\frac{1}{2n^2\bar{u}}\sum_{i=1}^{n}\sum_{j=1}^{n}|u_i - u_j| \tag{7}$$

If the aggregate disvalue of inequality is measured by (7) rather than (6), then it is possible to improve the outcome with respect to inequality as a result of an increase in the population. Consider an example discussed by Arrhenius (2013). By comparison with a two-person population in which one person is at 10 and another is at 20, a population in which one person is at 10 and 999 people are at 20 scores lower in respect of inequality according to (7), but is worse from the perspective of inequality according to (6).[18]

It is also possible to challenge the argument while retaining a simple additive measure of the badness of inequality across populations. The proper way of valuing equality, we might claim, should assign positive value to additional instances of pairwise equality between persons, and should not merely assign disvalue to instances of pairwise inequality. Thus, according to Arrhenius (2013: 85) "the more people who are unequal, the worse [is a population] in regards to egalitarian concerns, other things being equal; and the more people who are equal, the better [is a population] in regards to egalitarian concerns, other things being equal." Arrhenius calls this *positive egalitarianism*, in contrast with *negative egalitarianism*, which is the view presupposed by the egalitarian argument for anti-natalism. Arrhenius notes that positive egalitarianism may also allow us to count a two-person population in which one person is at 10 and another is at 20 as worse than a population in which one person is at 10 and 999 people are at 20, since, while the latter contains 999 relations of inequality by comparison with former's 1, it also contains 498,501 relations of equality, rather than none.[19]

[18] A different way in which to challenge egalitarian argument for anti-natalism's reliance on a simple additive measure of the badness of inequality is by claiming that the disvalue of inequality is conditional on existence. More exactly, we say that x is worse than y in respect of equality in virtue of the inequality between i and j in x only if i and j exist in y. This is in line with the view adopted by Parfit (1984: 425), who denies that we make the outcome worse with respect to equality by adding someone with a life worth living who is worse off than already existing people through no fault or choice of their own.

[19] We might also propose to resist the argument by counting additional instances of pairwise equality between persons as neutral and as exhibiting 'greedy' neutrality of the kind discussed in section 3 of this paper. However, arguments for attributing 'greediness' to the neutrality of additional good lives given the procreation asymmetry do not seem to carry over to the value of equality. The argument that the procreation asymmetry requires us to posit 'greedy' neutrality for good lives depends, ultimately, on the fact that, while the addition of a life worth living may be claimed to make the outcome neither better nor worse, any life can in principle always be better in respect of welfare, and its being so yields a morally better outcome, *ceteris paribus.* It is this structural fact about welfare and its value that requires us to reject the otherwise natural interpretation of the principle

We conclude that two key assumptions are required by the egalitarian argument for anti-natalism. Firstly, the argument presupposes an additive measure of the badness of inequality. Secondly, it presupposes negative egalitarianism. Since the relevant arguments are large in number and often complex, I do not have space to address the plausibility of these assumptions (see Temkin 1993: 191–231; Persson 2001, 2003; Rabinowicz 2003; Arrhenius 2013; Segall 2016, 2019; Gustafsson 2020; Arrhenius and Mosquera 2022). Nonetheless, in my view, both are credible, insofar as telic egalitarianism is.

A final way in which to challenge the egalitarian argument for anti-natalism parallels the objection raised against the argument that negative utilitarianism entails the desirability of human extinction in section 3.1. The objection then was that that argument ignored ill-being among non-human animals. While the view that non-human animals fall within the scope of distributive equality may strike us as hard to believe (McMahan 1996: 29–31; Kagan 2019: 58–78), it is accepted by Vallentyne (2005), and Kagan (2019) argues that it has to be accepted by egalitarians who endorse a non-hierarchical theory of moral standing of the kind on which I'm implicitly relying (see section 2.4). Someone who accepts a telic egalitarian view whose scope encompasses non-human animals could conceivably object that the argument completely ignores them.

This objection is clearly much weaker than the corresponding objection noted in section 3.1, simply because it's very hard to believe that inequalities involving non-human animals matter morally. It's weaker in another respect too. Assume that inequalities involving non-human animals do matter morally. Human extinction would not only entail the absence of further inequalities between human beings, but between human beings and non-human animals. If you accept that inequalities involving non-human animals matter morally, then you are under significant pressure to accept the conclusion that inequalities in welfare between human beings and non-human animals are much worse than those between human beings or between non-human animals, given that human

of neutrality as the claim that the addition of a life worth living leaves the outcome exactly equally as good as before (see Broome 2005: 405–407). Equality is importantly unlike welfare in this respect, in that some instance of pairwise equality between persons cannot be improved in respect of equality, in the way that any life worth living can be improved in respect of welfare.

beings have access to goods that may seem incomparably better than those available to non-human animals (Vallentyne 2005; Kagan 2019: 63–4). Unless you can find some way to resist this conclusion,[20] including non-human animals within the scope of distributive equality probably just adds to the egalitarian case for human extinction.

To round out this section, I want to briefly touch on one additional issue, which concerns the value of justice in general.

Some may balk at my use of the phrase 'in general'. The relationship between telic egalitarianism and justice is a matter of controversy. According to telic egalitarians like Temkin (1993), it is in itself bad if some people are undeservedly worse off than others, even if these inequalities arise completely independently of human agency. Temkin (1993: 22) does not shy away from describing such inequalities as involving "cosmic injustice". However, according to Rawls (1971: 102), such inequalities can be neither just nor unjust.

Regardless of where we stand on this issue, I think we should take seriously the idea that, insofar as it contributes to the ranking of outcomes as morally better or worse, justice behaves like equality in respect of the value of the future – or rather, as I have suggested it is plausible to think equality behaves. In other words, to the extent that it functions as a value contributing to the ranking of outcomes as morally better or worse, justice may be conceived as the privation of an evil, and thus as non-instrumentally valuable only insofar as it constitutively precludes additional instances of injustice, whose realization would make the outcome worse (compare Ewin 1970).

Crisp and Pummer (2020: 408) suggest something like this view when they claim that "promoting justice impartially can be thought of as a matter of reducing or preventing injustice", to be contrasted with beneficence, which "involves reducing harm and increasing benefit." Since there will inevitably be more injustices if people continue to be born, the goal of reducing or preventing injustice may be thought to be achieved most fully if no more people come into being. Once again, that is not to say that things would thereby be better *all things considered*, since there may be other values worth caring about besides justice. Nonetheless, it is notable that in contrast to the readiness with which telic

[20] Vallentyne (2005) explores several.

egalitarians concede the ability of considerations of aggregate welfare to outweigh their concern to ensure that people are not worse off than others through no choice of fault of their own (e.g., Temkin 1993: 282), it is not uncommon among political philosophers to find striking assertions of the primacy of justice among human values, as in Rawls's suggestion that the virtue of justice is "uncompromising" and that "an injustice is tolerable only when it is necessary to avoid an even greater injustice." (Rawls 1971: 4)[21]

This section began by briefly discussing prioritarianism, noting that it supports the view that lives that are not worth living are bad to a greater degree than lives worth living are good, and so supports a more pessimistic assessment of the overall value of the future than views that weigh good and bad lives symmetrically, such as total utilitarianism. I then discussed an argument that it would be better from the perspective of the value of equality if no more people were born, noting its key assumptions, and suggesting that similar assumptions may be thought of as applying to justice more generally, insofar as justice is something that contributes to the ranking of outcomes as morally better or worse. However, since almost nothing has been written on how to aggregate justice and injustice in order to determine when one world is better overall with respect to this value, we understand much less about those dimensions of the value of justice required to assess this line of argument than about the corresponding dimensions of the value of equality.

3.8 Perfectionism

The axiologies discussed so far are all compatible with welfarism, the thesis that value supervenes on well-being. A prominent non-welfarist axiology is *perfectionism* (Hurka 1993). Very roughly, perfectionism attributes intrinsic value to the achievement of excellence in pursuits such as the arts and sciences, politics, or sport. Hurka (1993) develops a version of what he calls 'narrow perfectionism.' In the tradition of Aristotle (350 BCE [1980]), Marx (1844 [2007]), Mill (1863), Nietzsche (1901 [2017]), and Nussbaum (2000), narrow perfectionism treats the good as the development to a high degree or the full realization of those capacities that are central to human nature.

[21] Thanks to Jacob Barrett for this observation.

Perfectionism, in Hurka's conception, is not a theory of individual welfare. The perfectionist good is a distinct dimension of the good for which a person might choose to sacrifice her well-being.[22] This is one natural way in which to read the choice faced by Achilles at Troy between a brief glorious life that ends violently or a long, contented life in Thessaly: "if I return home to the beloved land of my fathers / the excellence of my glory is gone, but there will be a long life / left for me, and my end in death will not come to me quickly." (*Iliad* 9.414-6, transl. Lattimore) The dilemmatic character of this choice is arguably best captured by interpreting it as a choice between two fundamentally distinct types of goods – excellence and well-being - rather than as concerning the most efficient means to realize the singular good of welfare.

One notable respect in which the narrow perfectionist good seems to differ from the good of well-being is by lacking an intrinsically bad, negative counterpart. With respect to a given person's welfare, outcomes can be classified as good, neutral, or bad. Pleasure has its counterpart in pain.[23] The narrow perfectionist good has been argued to have no intrinsically bad, negative counterpart, because there's no meaning to the idea of developing one's essentially human capacities to a negative degree (Hurka 1993: 100–101; Murphy 2001: 43–44; Sumner 2020: 429–431).

This might be thought to entail that narrow perfectionism can't support a negative verdict on the value of the future. But this doesn't follow. Note, for example, that while the value of a sum can't decrease as a result of the addition of non-negative terms, an average can decrease as a result of incorporating additional non-negative terms.[24] Notably, Hurka (1996: 69-83) argues that counterparts of aveage utilitarianism or variable value theory

[22] Hurka (1993) actually leans toward, but stops short of fully embracing, a monistic, 'pure perfectionism,' on which the achievement of excellence is the only good.

[23] As noted by Kagan (2014) and Sumner (2020) it is much less obvious how to identify the negative counterpart of well-being relative to non-hedonistic theories of welfare. See Bradford (2021) and Pallies (2022) for recent work on this issue.

[24] We can also imagine a perfectionist analogue of positive critical-level utilitarianism that would count lives sufficiently impoverished in perfectionist value as making the outcome worse. Thanks to Teru Thomas for this observation.

are more plausible as applied to perfectionist value, whereas a counterpart of total utilitarianism is much less believable. When it comes to welfare, surely more of what is good is always better. The same doesn't obviously hold for excellence. Intuitively, some artistic careers would have been better had they ended sooner, not because their late stages don't contain (what are by ordinary standards) genuine achievements, but because they fall conspicuously beneath the level of excellence established by the artist earlier in their career. For example, Francis Ford Coppola's career as a filmmaker would arguably have been a greater achievement overall if he had retired already in 1980 to pursue winemaking on the back of completing *The Godfather* in 1972, *The Godfather: Part II* and *The Conversation* in 1974, and then *Apocalypse Now* in 1979.

For reasons already noted, if averaging (or averaging within the large population limit) is plausible as a principle for aggregating perfectionist value across people, then perfectionism can contribute to a negative assessment of the value of the future even without the postulate of intrinsic perfectionist bads. All that's required is that our descendants fall below the standard for perfectionist achievement set by previous generations.

Futures in which perfectionist goods go into decline, sacrificed for the sake of comfort, ease, and safety, are a staple of cultural pessimist fears about modernity. Modernity supposedly makes it too easy to satisfy our wants and needs without effort or creativity, offers us an ever-expanding menu of mindless distractions, and encourages narrow individualism over the pursuit of shared, communal projects. The result is "the loss of a heroic dimension to life" (Taylor 1992: 4), a world of "secure and self-absorbed last men, devoid of thymotic striving for higher goals in our pursuit of private comforts" (Fukuyama 1992: 328) This fear is taken to extremes in fictional dystopias like Huxley's *Brave New World* and Pixar's *WALL-E*.

There can be purely philosophical reasons to predict an eventual decline in the level of perfectionist value realized in a suitably long-lived humanity. According to what Hurka calls the *single-peak perfection principle*, the first instance of a given achievement, such as climbing a certain mountain or proving a certain theorem, is most valuable, and subsequent repetitions of the same achievement diminish in value to zero (Hurka 1993: 79–82). Unless the space of possible achievements is suitably large, the single-peak perfection principle suggests that a decline in the value of humanity's achievements is inevitable in the long run.[25]

However, the narrow perfectionist view also faces distinctive challenges in its application to the very long run. It understands excellence as the full realization of "whatever properties are essential to humans and conditioned on their being living things." (Hurka 1993: 16) Setting aside the concern that such properties are biologically dubious (Kitcher 1999), it's not at all clear how to apply this framework over long-run timescales where a fixed human nature can't be assumed and humanity might conceivably transcend biology (Kurzweil 2005).

Notably, some of the declinist fears belonging to the genre of pessimism I've highlighted build in the worry that human beings or their descendants will 'degenerate' by becoming biologically adapted to stultifying aspects of modernity (Moynihan 2020: 312–322). H. G. Wells' time traveller explains the Eloi, the physically and intellectually diminutive people he encounters on the surface in the year 802,701, in these terms: "Humanity had been strong, energetic, and intelligent, and had used all its abundant vitality to alter the conditions under which it lived. And now came the reaction of the altered conditions." (Wells 1895: 74) Narrow perfectionism seems to have the implication that, considered as a result of biological evolution under natural selection, the loss of all higher faculties from the genus *Homo* wouldn't be regrettable, in that individuals so adapted don't fail in the realization of their nature in foregoing the use of the sophisticated cognitive capacities we think of as distinctive of our species. This feels hard to believe - and difficult to square with the spirit of a perfectionist ethics.

Cases of this kind also indirectly serve to highlight a difficult problem that Hurka ex-

[25] In the famous passage at the very end of *Reasons and Persons*, where he discusses the importance of avoiding human extinction, Parfit argues that perfectionism favours continued human survival "because what matters most would be the *highest* achievements of these kinds, and these highest achievements would come in future centuries." (Parfit 1984: 454) Setting aside the question of whether in fact the likes of Shakespeare or Bach will ever be surpassed, Parfit here appears to assert a *maximax* criterion for comparing worlds according to perfectionist value, on which worlds are ranked purely according to the highest value achieved by the most perfect individual(s). A similar view may be attributed to Nietzsche (1901 [2017]). See Hurka (1993: 75–80) for criticism of the maximax criterion.

plicitly brackets: namely, how to account for members of other animal species within the scope of a perfectionist axiology. So far, I've implicitly assumed that under narrow perfectionism, perfectionist goods are to be aggregated, in the first instance, within a given biological species. For example, in considering whether the average level of perfectionist goods will decline over the long-term as a result of the supposedly stultifying effects of modernity on human beings, I implicitly assumed that we're concerned with the average level of perfectionist value achieved by humans, rather than the relevant average being taken with respect to all living things. This arguably makes sense within the framework of narrow perfectionism, since it conceives of the good in each of its concrete instances as species-relative. The question remains how to aggregate across species to arrive at a global evaluation of outcomes.

Let me wrap up. I suspect that perfectionist ideals play a greater role in our thinking about the value of the future than is typically acknowledged. According to Ord (2021: 217), "It is because [our] potential is so vast and glorious that the stakes of existential risk are so high." *Glory* is a distinctively perfectionist ideal. It is that for which Achilles chose to die violently in his youth beneath the walls of Troy rather than live a long, happy, and unremarkable life in his homeland. In considering how perfectionism bears on our assessment of continued human survival, I've stressed the fact that it's significantly more plausible that perfectionist goods ought to be aggregated by averaging, and how this may justify an Achillean preference for our species' history to be shorter than it might otherwise be, even if the future available to us would be comfortable and contented. On the other hand, there are significant, as-yet unresolved difficulties in applying the perfectionist axiology at the grand scales we're considering, due to the instability of human nature over long-run timescales and the need, ultimately, to aggregate across the many and varied forms of life and rational activity that exist today and tomorrow.

3.9 Conservatism

According to Cohen (2012: 149), (small 'c') conservatives "exhibit a bias in favour of retaining what is of value, even in the face of replacing it by something of greater value". Things which are of value are to be valued not merely as vehicles by which goodness or beauty enter into the world, but for themselves. Therefore, we have reason to oppose their destruction, even when necessary to bring something of greater value into existence.

Frick (2017) argues that our moral reasons for ensuring humanity's survival can be understood as conservative in this sense. We ought to ensure our species' survival not because it's better for there to be more beings with good lives, but because humanity is valuable in itself in virtue of its unique intellectual, affective, and ethical capacities, and the appropriate response to recognising the value of a thing involves caring about its continued existence. The conservative attitude explains straightforwardly why most people care about humanity extending on and on into the future, but are not concerned to maximize the size of the population that can be supported at any given point in time.

It might be thought that conservatism about value is not a hypothesis about value in the sense that average or total utilitarianism is a hypothesis about value and should therefore be set aside in an inquiry like ours. Cohen (2012: 155) even claims that conservatism is incompatible with maximizing consequentialism. His intuition, I take it, is that because conservatism says that we have reason to retain what is of value rather than replace it with something of greater value, it entails that we can have most reason to choose an outcome that is sub-maximal in the axiological ordering. This inference is invalid. The axiological framework assigns values to outcomes. Conservatism is a claim about the proper response to the value of individual things or individual people. It's an open question how to relate the ordering of outcomes in terms of moral value to assignments of moral value to particulars. In principle, there's no reason why we're barred from ordering possible worlds in such a way as to give weight to the longevity of valuable things.[26]

[26] A different way in which conservatism might resist being captured by any axiological hypothesis is by invoking agent-relativity. For example, the claim that it's right to have a bias in favour of retaining existing value-bearers could be understood as a claim that we have reason to conserve valuable things that exist already *now* at the time at which you and I exist. Thus, we needn't have a bias in favour of outcomes in which value-bearing entities that come into existence in future are retained rather than subsequently replaced, although future agents ought to have such a bias. Reasons to value the conservation of particular value-bearers are therefore agent-relative. By contrast, an axiological ordering is agent-neutral. But, as a matter of fact, the picture of conservatism that I've just sketched is one that Cohen (2012: 166) explicitly rejects in favour a view that treats the conservation of future and present existents symmetrically. This is not to deny that some aspects of conserva-

How might a conservative attitude toward the value of humanity shape our thinking about the value of the human future? First and foremost, let's consider the core the point at which Frick (2017) is ultimately driving, which relates to issues addressed already in section 3.2. If we accept the procreation asymmetry, we can't claim that the continued coming-into-being of new humans is desirable in light of the value of individual welfare considered as a good to be promoted. However, we may try to argue that the continued existence of humanity is to be desired for some other reason: for example, in virtue of the appropriateness of adopting a conservative attitude to the value of humanity.

As already noted, it's not so easy to reach the conclusion that the survival of humanity is all-things-considered desirable in this way, even if we adopt an optimistic forecast of future welfare levels. Adherents of the principle of neutrality are committed to the idea that adding lives worth living to the population can swallow up good things and neutralize them. It follows that if we accept the procreation asymmetry, we face obstacles in arguing for the desirability of the survival of humanity even if we can point to values besides the promotion of welfare that speak in favour of a continued human presence, such as the value conservatives attach to the continued existence of valuable things.

Here's a different issue for us to consider. Suppose that total utilitarianism entails that the continued existence of humanity is undesirable. There will be too much suffering. Any minimally plausible moral outlook must allow that if some suitably high proportion of future lives will be very horrible if we go on, then it would be better that humanity cease to exist. Might a conservative claim that the importance of retaining existing things of value, even when they can be replaced with something better, can nonetheless make the continued existence of humanity desirable in less extreme scenarios where total utilitarianism entails that the continued existence of humanity is morally undesirable?

Plausibly not. In Cohen's discussion, the emphasis is on conserving not something

tive valuing have an agent-relative character. Cohen distinguishes between two different types of conservative valuing, which he calls 'personal valuing' and 'particular valuing'. Personal valuing is agent-relative, in the sense that involves valuing the continued existence of something because of its relationship to you. Particular valuing, by contrast, is conservative valuing that is independent of any personal relationship to the valued object and responds to its value as such. The latter gets most of the space in Cohen's paper.

which is intrinsically valuable *in some respect*, but something which is intrinsically valuable *on the whole*. Thus, Cohen insists that unjust social arrangements aren't appropriate objects of the conservative attitude, even if they're valuable in some respect or other. Arguably, if the continued existence of humanity would yield so much suffering that total utilitarianism recommends its extinction, this is strong evidence that humanity lacks intrinsic value in the overall sense Cohen has in mind.

We may have distinctive, valuable qualities, but that's not all. We're also capable of profound evil and of callous indifference to profound evil. As a result, it's not so obvious that humanity is an appropriate object of conservative valuing. A philosopher who appears to take the opposite view is Nozick (1989). Nozick (1989: 238–239) argues that in the aftermath of the Holocaust, it "now would not be a *special* tragedy if humankind ended .... That species, the one that committed *that*, has lost its worthy status." The exact sense in which Nozick means to rule out that the end of humanity would be 'a *special* tragedy' isn't altogether clear, but one plausible reading is that Nozick means to deny that it is fitting to value humanity in the distinctive way that conservatives might value the Grand Canyon or the paintings of Fra Angelico, as objects of value worth preserving in their own right.

Conservatism itself might be thought to yield distinctive reasons for finding humanity to be unworthy of being so valued. In the literature on existential risks from artificial intelligence, it is sometimes noted that the gulf between our intelligence and that of future software agents could place us in a position of vulnerability with respect to machine intelligence like that in which the rest of the biosphere stands to us (Russell 2019: 132–136; Ord 2021: 142–143). What is striking is the implication that, in respect of other living things, we are exactly the kind of existential threat the authors warn us of and hope to avert. Provided that we accept a holist theory of environmental ethics that attributes intrinsic value to species, ecosystems, and other supraorganismic wholes (Leopold 1949; Rolston 1988; Callicott 1989), conservatism provides the most natural moral framework within which to justify an imperative to conserve existing species and ecosystems against these existential threats from human industrial civilization.

As a result of human activity, extinction rates for mammals, amphibians, birds, and reptiles over the last 500 years are at least as high as those responsible for the previous 'Big

Five' mass extinction events in the geological record and could yield comparable extinction magnitudes within as little as three hundred years (Barnosky et al. 2011). Now, mass extinctions do not only destroy. For example, the tropical rainforests whose losses we now lament arose as a result of the asteroid impact that triggered the most recent Cretaceous-Paleogene mass extinction (Carvalho et al. 2021). Thomas (2017) notes that "[a]fter every fall during the history of life there has been a subsequent rise in diversity" (40) and suggests that "we should consider whether we are on the brink of a sixth major genesis of new life." (117) Among other observations, he notes that the rate of phenotypic change in human harvested systems is more than thrice the normal rate (Darimont et al. 2009), and that we are currently witnessing unprecedented rates of plant speciation (Thomas 2015). But these are considerations of precisely the sort that provide cold comfort to conservatives, whose concern is for the preservation of particular valuable things as they already are. Theirs is the most natural moral framework within which to lament all that's already been lost and all that is still being lost in the face of human rapacity.

All in all, while conservatism provides an intuitive framework for understanding the sense in which many people are distinctively concerned for the human species to continue existing, it provides less support for the desirability of continued human survival than has been claimed, both because there are unacknowledged obstacles to combining conservatism with the principle of neutrality to derive verdicts about all-things-considered desirability, and because there are reasons to doubt that humanity is a proper object of conservative valuing, some of which are themselves distinctively conservative in nature.

3.10 Summary

We have considered a number of different value theories discussed among contemporary moral philosophers and how they might bear on the value to be assigned to the continued survival of humanity and to efforts to reduce the risk of human extinction. The results, I think, are often surprising, sometimes disturbing, and occasionally hopeful. The procreation asymmetry does not speak for going gentle into that good night. The putative desirability of human extinction considered in relation to the deleterious effects of human activities on non-human animals invokes many classic problems in population axiology related to the valuation of lives that are only weakly positive, as well as some new ones related to the valuation of lives that are only weakly negative. There is a plausible case to be made that telic egalitarians must inevitably look upon continued human survival as less desirable, that perfectionism counsels us to someday burn out rather than fade away, and, finally, that the conservative case for preserving humanity has been overstated.

4 Decision Theory - Evaluating Risks of Extinction

The survival of humanity is a kind of gamble. It could go well, and it could go terribly wrong. How do we value an uncertain prospect of this kind, even presuming that we know how to value each of its possible outcomes? This section explores the significance of different ways of answering this question, with particular emphasis on the question of how to balance potential upsides and downsides of gambles. I also address the significance of one important way in which we might fail to know the value of every possible outcome of a gamble: namely, by being ignorant of the correct axiology. Furthermore, I discuss how our assessment might reflect our uncertainty about the right way of evaluating uncertain prospects. Throughout, I assume a broadly Bayesian perspective on which a rational agent's uncertainty is to be represented by a unique probability measure on an algebra on the set of possible worlds, W.

In section 4.1, I consider *expected value* and its relation to *expected utility theory*, with particular emphasis on the possibility that the two are mutually incompatible because a rational agent cannot be represented by an unbounded utility function. As a result, a rational agent may be required to be risk seeking or risk averse in respect of moral value, depending on their views about the overall goodness of the world as a whole. In section 4.2, I discuss risk-weighted expected utility theory. I criticise the claim that we are required to use a risk avoidant risk function when making decisions on behalf of future people, as well as a model used to support the conclusion that a risk function of that kind should lead us to prefer extinction. Finally in section 4.3, I discuss normative uncertainty, both axiological and decision-theoretic. In particular, I highlight important asymmetries in the space of possible theories that may bias evaluations made under normative uncertainty in the direction of pessimism.

4.1 Expected Value and Expected Utility

We can think of an event like continued human survival up to time t as a set of possible worlds. For any event, E, assuming for simplicity that W is countable and that $V(\,\cdot\,)$ is a real-valued cardinal measure of the moral value of worlds, E's value considered as an event may be identified with its expected moral value,

$$\mathrm{EV}(E) := \sum_{w \in W} \Pr(w \mid E) \cdot V(w) \tag{8}$$

Thus, if the moral value of outcomes is identified with total welfare as on total utilitarianism (section 3.4), then uncertain prospects are valued at their expected total welfare.

If we value events in this way, we are said to be *risk neutral* with respect to total welfare. We are indifferent between any event, E, and any *mean-preserving spread* of E that pushes probability mass toward the extremes while leaving the expectation unchanged.[27] We might instead be *risk averse* with respect to moral value, strictly preferring E to any mean-preserving spread of E. This might not seem so implausible to us. When choosing in morally charged situations, shouldn't we be especially cautious? Isn't that kind of caution at least permissible?

On its face, how we answer this question bears on how we assess the prospect of continued human survival. Even if better in expectation, the continued existence of humanity might seem like a riskier prospect than a world without us. In a world without us in which non-human animals continue to exist, there is a limit to how badly wrong things might go. In one to two billion years' time, the increasing brightness of the sun will trigger a runaway greenhouse effect, and the seas will boil away. There will be no more suffering. By contrast,

[27] More exactly, E' is a mean-preserving spread of E just in case $\mathrm{EV}(E) = \mathrm{EV}(E')$ and there is a bijection, f, between E and E', such that if $w' = f(w)$ then $V(w') = V(w) + Z(w)$, where $Z(w)$ is a random variable whose expected value is zero conditional on any possible value of $V(w)$.

the continued existence of humanity would greatly increase the potential spatiotemporal extent - and perhaps also the potential severity - of ill-being among Earth-originating welfare subjects (Althaus and Gloor 2016).

It's tempting to infer that the view that an impartially beneficent agent should value gambles at their expected moral value is supported by the fact that orthodox decision theory requires rational agents to value uncertain prospects in accordance with their *expected utility* (Arnauld and Nicole 1662; Bernoulli 1738; Ramsey 1926; von Neumann and Morgenstern 1947; Savage 1972). Rougly speaking, 'utility,' so understood, is a measure of the agent's strength of preference, to be contrasted with its use by classical utilitarians – and by many moral philosophers today – as a synonym for 'welfare' (Broome 1991).

The argument would go through straightforwardly if we had some way of establishing that an impartially beneficent agent's preferences over outcomes are to be represented by a utility function that is an increasing linear function of their moral value. One obstacle to arguing in this way is that, on their standard interpretation, the representation theorems that underwrite orthodox decision theory (von Neumann and Morgenstern 1947; Savage 1972) are not supposed to permit us to assign a cardinal utility representation to the agent's preferences over outcomes that is prior to her preferences over uncertain prospects (Dreier 1996), and so don't let us treat a person's utility function as linear in some good independent of the assumption that she's risk neutral with respect to that good.[28]

Worse yet, when the possible outcomes associated with uncertain prospects are permitted to be countably infinite, (generalisations of) the standard axioms of expected utility theory require that the utility function is *bounded*, since unbounded utility functions may involve the decision-maker preferring a gamble to each of its possible outcomes (see Fishburn 1970: 194, 206–207; Kreps 1988: 59-65; Hammond 1998: 43–48; McGee 1999; Russell

[28] We might think this points to a flaw in the standard way of thinking about representation theorems and their role in decision theory. It's common to object that we ought to be able to disentangle the agent's attitude to risk from the way she values outcomes (Watkins 1977; Hansson 1988; Buchak 2013). We'll look at a decision theory that allows us to do just that in the next section, though it's one that rejects expected utility maximization as a rational requirement. In the recent philosophical literature, Peterson (2004) and Easwaran (2014) show how to derive the norm of expected utility maximization starting from the agent's preferences over outcomes and deriving normative constraints on preferences over gambles as a result.

and Isaacs 2021).[29] This suggests a sense in which orthodox decision theory points toward the conclusion that a rational agent is required *not* to be risk neutral with respect to moral value. It seems very plausible that moral value is unbounded. Total welfare, for example, seems to be an unbounded quantity. There can always in principle be more and more people with lives at a given non-zero welfare level. Therefore, a rational agent cannot have a utility function that is linear in total welfare.

Many understandably find it incredible that we are required to have bounded utilities (McGee 1999; Arntzenius et al. 2004; Nover and Hájek 2004; Monton 2019). Nonetheless, given that bounded utility follows from generalizations of the standard axioms, which ought to seem compelling to anyone who accepts expected utility theory for finite uncertain prospects, it is worth briefly considering how a rational agent might asses the expected utility of continued human survival given a bounded utility function that is strictly monotone increasing in moral value.

A natural way to imagine such a utility function is as a *logistic function* on moral value (Kosonen 2022), of the form

$$u(x) = \frac{L}{1 + e^{-k(x-x_0)}} + C \tag{9}$$

L and C together determine the upper and lower bounds on the function. Their values are arbitrary for our purposes, given that the utility function is invariant under positive linear transformation. The parameter k determines the steepness of function, and x_0 determines the inflection point. In the abstract, the logistic function has the following shape:

[29] Note, however, that Russell and Isaacs (2021) also show that if utilities need not be real valued, then utilities need only satisfy a weaker property they call *limitedness*.

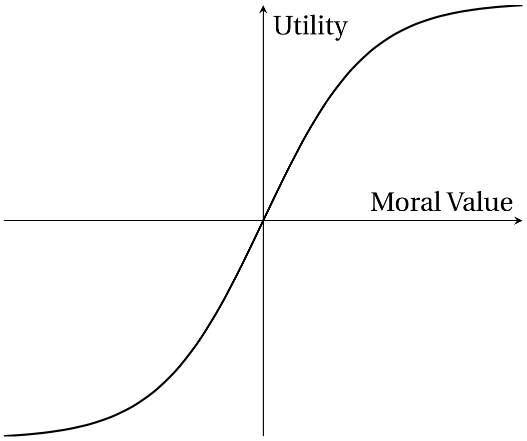

Figure 1

The agent is risk averse in respect of moral value for outcomes to the right of the inflection point, where the utility function is concave, and risk seeking in respect of moral value for outcomes to the left of the inflection point, where the utility function is convex. Insofar as any non-arbitrary choice can be made here, it is natural to identify the x-coordinate of the inflection point with the moral value of the null world, with outcomes to its left as those worse than a world without anything of value at all, and outcomes to its right as better.

Given this utility function, how the agent assesses the prospect of continued human survival will depend on how they assess the value of the world as a whole, taking into account the distribution of moral value across all spacetime coordinates, no matter how distant. For example, if they are suitably confident that there exists a suitably large background population with negative average lifetime welfare and they identify moral value with total welfare, they might be highly risk seeking in evaluating the prospect of continued human survival, since they expect the value of the world as a whole to be far to the left of the inflection point. Because their utility function is non-linear in total welfare, the agent is forced to take account of exobiological issues like those discussed in section 3.6 in evaluating the prospect of continued human survival, even though they accept a separable axiology.[30]

In principle, we could imagine that these exobiological issues are important for any moral decision the agent might make under conditions of uncertainty, because they determine the agent's risk attitude with respect to moral value. However, it seems more natural to imagine that the growth rate of the utility function is sufficiently small that utility can be safely approximated by a linear function in ordinary, small-stakes cases. It is only when considering really grand questions – such as, potentially, the continued survival of humanity – that the non-linearity of the agent's utility function might become significant. On the other hand, if the growth rate of the logistic function is low enough, even the possible changes in total welfare associated with extinction and continued survival might remain small enough that the non-linearity of the agent's utility function can be safely ignored for practical purposes.

In summary, whereas there is no doubt something intuitively attractive about the prescription to value an event at its expected moral value, it is harder than we may initially expect to derive support for this form of risk neutrality from orthodox decision theory's prescription to maximize expected utility. In fact, orthodox decision theory seems to allow us to construct a strong case against risk neutrality with respect to moral value, assuming moral value to be unbounded. A reasonably plausible utility function on moral value compatible with natural generalizations of the axioms of expected utility theory may require the agent to be risk-seeking or risk-averse depending on conjectures about the overall goodness of the world, at least when confronting suitably grand questions, such as (perhaps) the desirability of continued human survival.

4.2 Risk-Weighted Expected Value

In the recent philosophical literature, an especially influential alternative to expected utility theory is defended by Buchak (2013). Building on the rank-dependent utility model originally developed by Quiggin (1982), Buchak argues for the rationality of maximizing *risk-weighted expected utility*. We might similarly claim that a rational, impartially benef-

[30] For relevant discussion, see Beckstead and Thomas (2021), Goodsell (2021), Russell (2021).

icent agent values uncertain prospects at their *risk-weighted expect moral value* (REV) (Pettigrew 2022).

Let's assume for simplicity that we are back to dealing with finite events.[31] For any event, E, suppose we are able to list E's elements in ascending order from worst to best. In other words, we index the worlds in E as $w_1, w_2, \ldots, w_n$, where $V(w_i) \leq V(w_{i+1})$ for $i = 1, ..., n-1$. Then the risk-weighted expected value of E is

$$\text{REV}(E) := V(w_1) + \sum_{k=2}^{n} \left[(V(w_k) - V(w_{k-1})) \cdot r \left(\sum_{i=k}^{n} \text{Pr}(w_i \mid E) \right) \right] \tag{10}$$

In other words, we add to the moral value of the worst possible outcome in E, $V(w_1)$, the difference in value between the second-to-last-ranked world and the worst outcome, $(V(w_2) - V(w_1))$, weighted by applying a function, $r(\,\cdot\,)$, to the probability of attaining a world in E at least as good as w_2, and then we add the difference in value between the third-to-last-ranked world and the second-to-last-ranked world, $(V(w_3) - V(w_2))$, weighted by $r(\,\cdot\,)$ of the probability of attaining a world in E at least as good as w_3, and so on.

We call $r(\,\cdot\,)$ the *risk function* and stipulate that it is a strictly increasing function that maps the unit interval to itself and has 0 and 1 as fixed points. As its name suggests, the risk function is intended to encode the agent's attitude toward risk. If $r(\,\cdot\,)$ is linear, then risk-weighted expected value reduces to expected value. According to Buchak, we are not rationally required to have a linear risk function. When $r(\,\cdot\,)$ is strictly convex, the agent is said to be *risk avoidant* and weights the worst possible outcomes out of proportion to their probability. An agent who is risk avoidant in this sense can be risk averse in respect of goods on which her utility function is linear.

It may be thought that if a wide range of risk functions are rationally permissible, then even if an impartially beneficent agent is required to have preferences represented by a utility function that is linear in moral value (or a utility function that can be so approx-

[31] Risk-weighted expected utility theory potentially gives us some additional resources to tame the problems associated with countably infinite gambles noted in the previous section (Buchak 2013: 73–74), albeit not without violating the principles ultimately adopted by Buchak (2013: 82–113) in axiomatizing risk-weighted expected utility theory.

imated for the possible outcomes under consideration), different impartially beneficent agents with the same beliefs and values can permissibly differ significantly in their evaluation of continued human survival by using their own idiosyncratic risk function.

Buchak (2019) argues for a notably different conclusion. According to her *future risk-avoidance principle*, "If we are making a decision whose largest effects concern a large group of future individuals, then we should make a very risk avoidant choice" (Buchak 2019: 78). Therefore, when assessing the prospect of continued human survival for the sake of making decisions about global priorities, we are morally constrained to rely on a strictly convex risk function and can't use our own idiosyncratic risk function unless it has this property. In a recent paper, Pettigrew (2022) argues for the striking conclusion that (a suitably refined version of) the future risk-avoidance principle supports the verdict that agents exclusively concerned with the moral value of outcomes should prefer human extinction because of the significant weight they are required to place on avoiding catastrophic outcomes.

Buchak arrives at the future risk-avoidance principle as follows. We are morally required, she claims, to choose in accordance with a risk attitude that is sensitive to the risk attitudes of the agents potentially affected by our decision. We ought not simply follow our own idiosyncratic attitude toward risk and impose it on others. Moreover, when the risk attitudes of those we affect through our actions are unknown to us, we are morally required to default to the most risk avoidant risk attitude within reason, so that our choice cannot reasonably be rejected as excessively risky by those we might affect.[32] These com-

[32] Note that Buchak here distinguishes between those risk attitudes that are *rational*, in the sense of satisfying certain minimal coherence constraints, and those attitudes that are *reasonable*, in the sense of satisfying certain additional substantive normative constraints (compare Rawls 1993). In this sense, it may not be irrational to prefer the destruction of the whole world to the scratching of your finger, but it is unreasonable. The most risk avoidant risk attitude within reason is taken to be the extremal member of the set of reasonable risk attitudes, and not the set of rational risk attitudes. Clearly, it is difficult to say exactly what kind of risk avoidant attitude is within the outer bounds of reasonableness, but to give the reader some sense of this, Buchak (2019: 73) suggests that "it is not unreasonable to care about the bottom half of consequences five times as much as the top half, but that is close to the reasonable lower limit." Thus, consider a gamble over two differently valued outcomes x, and y, where each has a .5 probability and $u(x) > u(y)$. Then, according to Buchak, it is not unreasonable to have a risk function, r, such that $r(1) - r(.5) = 5 \cdot r(.5)$, or, equivalently, such that $r(.5) = 1/6$.

mitments are encapsulated in Buchak's *risk principle*: "When making a decision for an individual, choose under the assumption that he has the most risk avoidant attitude within reason unless we know that he has a different risk-attitude, in which case, choose using his risk attitude." (Buchak 2017: 632) Since the risk attitudes of future people are unknown to us, Buchak infers from the risk principle that acts that *exclusively* affect future individuals are to be governed by the most risk avoidant risk attitude within reason. Therefore, for any plausible principle for aggregating across groups of individuals with diverse risk attitudes, decisions that *primarily* affect future people should be governed by a highly risk avoidant risk attitude - one that doesn't stray too far from the most risk avoidant within reason.

The final step in Buchak's argument for the future risk-avoidance principle is questionable, or so I'll now argue. It is not true that any plausible principle of aggregation in this context delivers the future risk-avoidance principle. Indeed, it is hard to see how any could.

Suppose that we are attracted to an additive principle for interpersonal aggregation. When people all have or are imputed to have the same risk function, $r(\cdot)$, we then face the choice of whether to prefer the act that maximizes the risk-weighted expectation of the sum of each person's welfare relative to $r(\cdot)$ or the sum of the risk-weighted expectation relative to $r(\cdot)$ of each person's welfare. In fixed population cases, there is no difference between the sum of each person's expected welfare and the expectation of the sum of each person's welfare, so these coincide when $r(\cdot)$ is linear. The same is not true when a nonlinear risk function is applied (Blessenohl 2020).

To see this, consider the following case, modified from Nebel (2021: 103–104) and based on the famous *Allais Paradox* (Allais 1953).[33] One of 100 numbered tickets is drawn at random, yielding the following welfare outcomes for Afryea and Beom-seok, each of whom is known to have the risk function $r(\Pr(X)) = (\Pr(X))^2$.

Nonetheless, this is close to the outer bounds of reasonableness. Thanks to Lara Buchak for help in clarifying this example.

[33] My presentation also draws inspiration from personal communication with Kacper Kowalczyk.

	Ticket #1-89	Ticket #90-99	Ticket #100
A	Afryea: $2+\epsilon$ Beom-seok: ϵ	Afryea: $5+\epsilon$ Beom-seok: $2+\epsilon$	Afryea: ϵ Beom-seok: $2+\epsilon$
B	Afryea: 2 Beom-seok: 0	Afryea: 2 Beom-seok: 5	Afryea: 2 Beom-seok: 0

Table 7

For small enough $\epsilon > 0$, B maximizes the sum of each person's risk-weighted expected welfare. Indeed, both have higher risk-weighted expected welfare under B than A. In that sense, B is *ex ante Pareto superior* to A. However, A maximizes the risk-weighted expectation of the sum of each person's welfare. Indeed, choice of A is guaranteed to yield higher total welfare and, more generally, dominates choice of B given any axiology that satisfies the following *anonymized ex post Pareto* principle: if x and y have the same population and x can be obtained from y by permuting welfare levels and then making each person better off, then x is better than y. All of total utilitarianism, critical-level utilitarianism, average utilitarianism, variable value theory, and total prioritarianism satisfy anonymized ex post Pareto.

It is highly plausible, nonetheless, that the moral considerations that motivate the risk principle require us to strictly prefer B to A in the choice above. Buchak claims that "we cannot choose a more-than-minimally risky gamble for another person unless we have some reason to think that he would take that gamble himself" and should "take only the risks that no one could reasonably reject." (Buchak 2019: 74) Both Afryea and Beom-seok would prefer B to A if choosing rationally on their own behalf. Choice of B is, in that sense, uniquely justifiable to each in light of her risk attitude, and each could reasonably reject choice of A on the grounds that there is another option that they rationally prefer and that everyone else is rational in preferring also (compare Frick 2015: 186–191). To the extent that we find the risk principle compelling, we therefore ought to strictly prefer B to A. In other words, we ought to prefer the option that maximizes the sum of each person's risk-weighted expected welfare, not the risk-weighted expectation of the sum of each person's welfare.

However, this principle doesn't deal well with variable population choices of the kind we inevitably confront when thinking about the future. Suppose there is some person who may or may not exist if a certain gamble is chosen. For any outcome in which she does not exist, her welfare level in that outcome will be undefined (see Broome 1999: 16; Bykvist 2007; Rabinowicz and Arrhenius 2015). As a result, neither the expected value, nor the risk-weighted expected value of the gamble, is defined from her perspective. By extension, the sum of each person's risk-weighted expected welfare is undefined. Far from supporting the future risk-avoidance principle, a principle of maximizing the sum of each individual's risk-weighted expected welfare simply breaks down when thinking about the long-run future.

To be clear, the problem doesn't depend on aggregating by taking an unweighted sum. Any principle that relies on aggregating each person's risk-weighted expected welfare relative to a given option threatens to break down when some person does not exist in every outcome associated with that option, because that person's risk-weighted expected welfare is undefined relative to choice of that option. By contrast, in any possible outcome, the aggregate welfare of each person existing in that outcome is well-defined, and by extension, so is the risk-weighted expectation of aggregate welfare.

We should also note that since the moral considerations that motivate the risk principle require us to strictly prefer B to A in Table 7, it is plausible that the moral considerations that motivate the risk principle are not related to achieving morally good outcomes and instead reflect what we may think of as characteristically Kantian concerns, such as respect for the autonomous personhood of others as a constraint on permissible action (Kant 1785 [1998]; O'Neill 1989; Darwall 1996; Korsgaard 1996; Scanlon 1998). That's because A is guaranteed to be better than B given anonymized ex post Pareto, and it seems plausible that the correct axiology satisfies anonymized ex post Pareto. Therefore, even apart from the issue noted in the previous paragraph, a principle like the risk principle ought to be set aside in an inquiry like ours, which is purely axiological (see section 2.1). We are left with no apparent reason to reject the thought that when a wide range of risk functions are rationally permissible, different impartially beneficent agents with the same beliefs and values can differ significantly in their evaluation of continued human survival by using their own idiosyncratic risk function, at least insofar as they are concerned exclusively with the goodness of outcomes.

Of course, those risk functions might themselves be risk avoidant. It is therefore also worth exploring reasons why the claim that even moderate risk avoidance should lead an impartially beneficent agent to prefer human extinction may not be sound, by examining the model used by Pettigrew (2022) to argue for that conclusion.

For simplicity, the model assumes four possible futures for humanity. In *lh*, we have a long future in which the average quality of life among human beings is very high. In *mh*, the future is mediocre in quality, being either long and mediocre in respect of the average quality of life or high in respect of the average quality of life but short in duration. In *ext*, humanity goes extinct within this century, with average welfare at a mediocre level. In *lm*, the future is long and miserable, and most people do not have lives worth living. Letting moral value be accounted for in units corresponding to a year of life at the very high welfare level achieved in *lh*, the values and probabilities of these outcomes are assigned by Pettigrew as follows:

	lm	lh	ext	mh
Probability	10^{-7}	10^{-5}	10^{-2}	$1 - 10^{-2} - 10^{-5} - 10^{-7}$
Moral Value	-10^{19}	10^{19}	10^{4}	10^{11}

Table 8

Within this model, the expected moral value of non-extinction exceeds the expected moral value of extinction, but for the risk function $r\,(\Pr(X)) = (\Pr(X))^2$, the risk-weighted expected moral value of extinction is greater. In fact, this holds true for the risk function $r\,(\Pr(X)) = (\Pr(X))^k$ for values of k greater than about 1.38. Thus, even agents who are moderately risk avoidant prefer extinction.

One respect in which Pettigrew's model could be criticized is on the same ground as Smart's objection to negative utilitarianism: namely, its exclusive focus on human welfare. In response, Pettigrew (2022: 26) argues that taking account of the welfare of non-human individuals "is unlikely to change the problem significantly. It only means that there are more minds to contain great pleasure in the long happy future (*lh*), but also more to contain great suffering in the long miserable one (*lm*)." But that is not all. If the moral value of outcomes depends only on human welfare, then the aftermath of an extinction event that wipes out humanity is a riskless prospect, a guarantee of an indefinitely neutral outcome. This no longer holds if we give moral weight to the well-being of non-human welfare subjects. The assumption that human extinction is riskless stacks the deck in favour of the conclusion that risk-avoidant agents should prefer humanity's end.

One important respect in which human extinction constitutes a risky prospect is that it may coincide with or allows for the emergence of a different bearer of advanced technological capabilities within our region of the universe. For example, this might occur if human extinction is triggered by advanced artificial intelligence that is misaligned with human values (Bostrom 2014; Russell 2019; Ord 2021: 138–152). According to Ord (2021: 167), the majority of existential risk in the 21st century derives from AI. Our continued existence may seem like the safer bet, since our values are much easier to predict. Better the Devil you know.

Note, moreover, that even if we knew exactly how humanity would end and were certain that our demise would permanently neutralize the Earth and its environs, there would still be an important sense in which extinction constitutes a risky prospect, assuming that our preferences over risky prospects track their risk-weighted expected moral value and that we assign value to the welfare of non-human individuals. That's because these preferences are in principle sensitive to background uncertainty about the value of outcomes independent of our choices.

Here's an illustration of this phenomenon. Suppose your goal is to maximize total welfare and you can choose between C, obtaining 2 units of total welfare for sure, or D, a fifty-fifty gamble over obtaining 1 or 4 units of total welfare. If you want to maximize expected total welfare, you should choose D. If you are risk-avoidant and have the risk function $r(\Pr(X)) = (\Pr(X))^2$, then it might seem you should prefer C.

Suppose, however, that all the possible events described above merely represent potential gains relative to the background level of total welfare, which is independent of your choice between C and D and of the outcome of the gamble offered under D. That is, you either add 2 units to the background level for sure, or you add either 1 or 4 units, each with 0.5 probability. Moreover, you don't know the background level of total welfare. Suppose you are unsure whether the background welfare level is 0, 5, or 10, and you assign probabilities of 0.2, 0.6, and 0.2 to each these possibilities, respectively. Now the risk-weighted expected value of D exceeds that of C, given the stated risk function. Taking background uncertainty into account, your preferences align with those of the expected value maximizer, whose preferences are unaffected by background uncertainty. This result is typical: when we account for background uncertainty, the preferences of risk-avoidant agents tend to converge with those of agents with linear risk functions (i.e., $r(\Pr(X)) = \Pr(X))$, by injecting an element of risk into options that might otherwise have seemed riskless (Buchak 2013: 226–229; Thoma and Weisberg 2017; Thoma 2019; Buchak 2022; compare Tarsney 2020).

As a result, when evaluating the prospect of continued human survival in terms of the risk-weighted expectation of total welfare relative to a convex risk function, we need to take into account not only our uncertainty about how things play out here on Earth, but also what the overall distribution of welfare in its entirety looks like across all space and time. Even if our axiology is separable and our utility function is linear in the good, the non-linearity of the risk function requires us to take account of exobiological issues like those discussed originally in section 3.6.

Given the potential enormity of the total sum of welfare in the Universe as a whole, taking account of background uncertainty can push our preferences in respect of what happens here in our cosmic neighbourhood in line with those we'd have given a linear risk-function, even if our risk-function is convex. Consider again the probabilities and values in Table 8, and suppose we replace each possible outcome with a gamble on the background level of total welfare, which is independent of what happens here on Earth. For simplicity, let's adopt the very crude assumption that there is a 0.6 probability that the background level is zero, because life evolves only on Earth, a 0.2 probability that it is -10^{25}, because life is common in the galaxy and suffering predominates among living things, and a 0.2 probability that it is 10^{25}, because life is common and happiness predominates among living things. Under these assumptions, non-extinction is preferred for risk avoidant risk functions of the form $r(\Pr(X)) = (\Pr(X))^k$ with integer values of k as high as $k = 21$. Thus, a preference for continued survival may be highly robust to even extreme degrees of risk avoidance when taking background uncertainty into account.

Risk-weighted expected utility theory allows us to represent an agent as risk averse in respect of some good even if they value each additional increment of that good the very same. Arguments have been advanced for thinking that we ought to be risk avoidant in this sense when making decisions that affect future generations, but I have given reasons for thinking that those arguments do not succeed and must anyway appeal to moral considerations of a kind set aside in this inquiry. In addition, the sensitivity to background uncertainty exhibited by risk-weighted expected utility theory and the tendency of risk-weighted expected utility theory to agree with expected utility theory given suitably great background uncertainty limits the extent to which the theory makes any distinctive contribution to our thinking about the value of the future.

4.3 Normative Uncertainty

No one can be sure of any of the axiological theories discussed in section 3. Rather than simply reporting the value of the future relative to whichever value theory we are most confident in, we may believe that our assessment should take account of the verdicts of the many different theories to which we assign some degree of confidence and somehow aggregate them (Lockhart 2000; Sepielli 2009; Ross 2006; MacAskill et al. 2020).

This is most straightforward when all theories can be represented by interval-scale measurable value functions that are unit comparable. Informally, this means the theories not only rank events as better or worse, but also encode information about how much better or worse different options are relative to one another, and that we can meaningfully compare the sizes of these value-differences across theories.

For axiological theories that together satisfy these assumptions, arguably the dominant view among philosophers who think that normative uncertainty bears in some sense on what we ought to do is that the evaluations of different possible worlds across different theories should be aggregated under conditions of axiological uncertainty so that worlds in W are ranked according to their expected value (Lockhart 2000; Sepielli 2009; MacAskill and Ord 2020; MacAskill et al. 2020; Riedener 2020). In other words, if $V_1(\,\cdot\,), \dots, V_n(\,\cdot\,)$ are value functions representing theories $T_1, \dots, T_n$ and $\Pr(\,\cdot\,)$ is a probability function representing the agent's degree of confidence in each theory, then w is to be preferred to w' under axiological uncertainty just in case

$$\sum_{i=1}^{n} \Pr(T_i) \cdot V_{T_i}(w) \geq \sum_{i=1}^{n} \Pr(T_i) \cdot V_{T_i}(w') \tag{11}$$

When theories are not all interval-scale measurable and unit comparable, the question of how to aggregate across them becomes significantly more complex and controversial. MacAskill et al. (2020) propose that we impute additional structure to the available theories, so that we can still use an expectational criterion for decision-making. For example, they propose to impute unit-comparability between cardinal value theories that are not otherwise comparable by normalizing their variance. Tarsney (2021) argues that we ought instead to ignore the richer structure inherent in some but not all theories, such as using a decision criterion that ignores cardinal value differences when aggregating across both ordinal and cardinal theories.

An agent who aggregates evaluations across theories by calculating expected values can be expected to behave for the most part roughly as if she were certain of a pluralist axiology whose broad outline can be predicted *a priori* due to asymmetries in the space of plausible axiological theories.[34] For example, I don't know of any plausible view on which inequality is intrinsically good. The only sensible views are that it is intrinsically bad or neutral (see section 3.7). Therefore, agents who reason under axiological uncertainty based on the expected values of outcomes will behave for the most part roughly as if they were certain of a pluralist theory that attaches some degree of intrinsic disvalue to inequality, even if they are reasonably confident that only the sum of individual welfare is of intrinsic moral importance (MacAskill et al. 2020: 185). If the argument of section 3.7

[34] Salient exceptions arise in relation to opportunities for moral learning, where the agent may find it desirable to pay in order to gain evidence about the true moral theory, whereas someone who was already certain of a given moral theory would not. Thanks to Teru Thomas for this observation.

is to be believed, they are thereby required to downgrade their evaluation of continued human survival.

Consider, similarly, the location of the critical level. Among theories that satisfy separability and reject the principle of neutrality, there seem to be only two plausible views about its location. Either it is the zero level for lifetime welfare, as on total utilitarianism or prioritarianism, or it is positive, as on positive critical-level utilitarianism. There is no plausible separable theory of which I know on which the critical level is negative. In the simplified case where $\Pr(\text{TU}) = p$ is your confidence in total utilitarianism and, for some version of positive critical-level utilitarianism with critical level c, $\Pr(\text{PCLU}) = 1 - p$ is your confidence in that theory, the agent acts for the most part roughly as if she is certain of positive critical-level utilitarianism with critical level $(1–p)c$ (MacAskill et al. 2020: 186–7). This is somewhat disturbing, because, as noted in section 3.5, there is a good case to be made that under the most plausible way of setting a positive critical level, most lives fall well below it.

By restricting ourselves to separable population axiologies, we set aside average utilitarianism. As noted in section 3.6, given a suitable background population, average utilitarianism behaves like critical-level utilitarianism with a negative critical level when it comes to the ordinal ranking of possible population changes. Suppose that average utilitarianism in fact treats the addition of a certain life at a welfare level of zero as an improvement, with positive critical-level utilitarianism delivering the opposite verdict. Then our relative confidence in average utilitarianism may in principle cancel out our confidence in positive critical-level utilitarianism when evaluating this population change.

Note, however, that even if we assume that average utilitarianism and positive critical-level utilitarianism are exactly equally plausible, we can't say whether their opposed evaluations cancel out unless we are able to determine whether the change is better according to average utilitarianism to the same extent that it is worse according to positive critical-level utilitarianism. Unfortunately, it's not at all obvious how to compare value differences across average utilitarianism and critical-level utilitarianism (see MacAskill 2014: 93–95; Hedden 2016). Nonetheless, Greaves and Ord (2017) argue that in the limit, the stakes according to critical-level utilitarianism swamp the stakes according to average utilitarianism. They claim that this *large-population swamping result* is almost entirely neutral with respect to how to make intertheoretic value comparisons.

The core idea can be explained as follows. Imagine adding a person at a fixed welfare level to a background population with a fixed average welfare level. For increasing sizes of the background population, the change in value decreases relative to average utilitarianism and approaches zero in the limit. It remains constant under critical-level utilitarianism. Therefore, the ratio of the amount at stake according to critical-level utilitarianism to the amount at stake according to average utilitarianism goes to infinity in the limit. This holds true regardless of how we make intertheoretic comparisons between average utilitarianism and critical-level utilitarianism, so long as intertheoretic comparisons are independent of the size of the background population.

The large-population swamping result has broader significance and applies to any choice about whether to aggregate putative goods or bads by summing or averaging. In section 3.7, it was noted that telic egalitarianism supports the desirability of human extinction if the disvalue of inequality is measured additively, rather than by a measure like the Gini coefficient. For my own part, I feel extremely uncertain which of these provides the right measure of the badness of inequality. However, even if we are most confident in the Gini measure, the additive measure will determine our assessment of the contribution of the disvalue of inequality to the value of the future under moral uncertainty once the background population is above a given size, provided that the way we compare how much is at stake across these different measures doesn't vary as a function of the background population.

Nonetheless, the fact that the large-population swamping result requires us to assume that the way we make intertheoretic comparisons is independent of the size of the background population is a more significant concession than may be apparent at first glance.

As noted previously, MacAskill et al. (2020) propose to impute unit-comparability between cardinal value theories that are not otherwise comparable by applying statistical normalization techniques. In particular, they advocate normalizing theories that are not otherwise unit-comparable at their variance. Very roughly, the argument for normalizing variance across the different theories is that this gives each theory in which the agent is equally confident an equal say in determining the most appropriate option (see MacAskill et al. 2020: 77–111). In order to normalize variance across theories, a choice has to be made about whether to do so by normalizing the theories in terms of the variance of the values they assign to the options currently available to the agent in a particular choice situation (the *narrow approach*), or in respect of some larger set of options – such as the set of all options to which the agent's ur-prior assigns non-zero probability (the *broad approach*). MacAskill et al. (2020: 101–105) argue for the narrow approach because it is more usable in practice as a guide to action and provides a straightforward way of taming radically incomplete theories, which otherwise threaten to derail the ability of expected moral value calculations to provide any guidance at all.

Note, then, that when using the narrow approach to variance normalization, it is impossible for average utilitarianism to be swamped by critical-level utilitarianism in evaluating the choice of whether to add a person at a given welfare level to a given background population with a fixed average welfare level. Quite generally, it is impossible for there to be more at stake for one theory than another in the choice between any pair of options in a binary option set. As a result, the large-population swamping result is less robust than it may first appear.

The discussion so far has focused on uncertainty over axiological theories. We might also be uncertain about how to value uncertain prospects. Some argue that in conditions of decision-theoretic uncertainty, there is a decision-relevant sense of 'ought' such that our ranking of options ought to be guided by an assessment derived by aggregating evaluations across the different decision theories among which we are uncertain, rather than the theory in which we are most confident or the theory that (unbeknownst to us) is true. Thus, in the context of uncertainty between evidential and causal decision theory, MacAskill (2016) suggests that one should maximize *meta expected value*: i.e., the probability weighted average of the choiceworthiness assigned to one's options by the different decision theories over which one is uncertain.

Earlier, I highlighted Buchak's claim, expressed in her *future risk-avoidance principle*, that when choosing in a way that impacts the welfare of future people, we ought to use a very risk avoidant risk function. We may well reject this view in favour of the view that we should maximize expected moral value. This corresponds to maximizing risk-weighted expected value relative to a linear risk function. Nonetheless, we shouldn't be certain that this view is right and Buchak's is wrong. On the other hand, I know of no one who seriously entertains the converse of Buchak's view. In other words, I know of no one who holds that when choosing in a way that impacts the welfare of future people, we ought to be risk seeking.[35] Suppose, then, that under conditions of decision theoretic uncertainty,[36] one should maximize meta expected value.[37] Then, arguably, one should assess options using a convex combination of a linear and a strictly convex risk function. This is itself a strictly convex function, and so maximizing meta expected value here will be equivalent to maximizing risk-weighted expected value relative to a risk avoidant risk function, albeit one that is less extreme in proportion to one's confidence that one ought to adopt a linear risk function instead.

Notably, I argued in section 4.2 that the future risk-avoidance principle is difficult to motivate by appeal to the more fundamental principle governing risk-taking to which Buchak appeals. The future risk-avoidance principle is therefore a principle in which we might not have very much confidence. I contrasted the future risk-avoidance principle with the view that different impartially beneficent agents can permissibly differ in their evaluation of continued human survival by relying on their own idiosyncratic risk function in computing risk-weighted expected values. However, there is a good case to be made that a view of this kind should be ignored under decision-theoretic uncertainty.

[35] More (2004) argues for a *Proactionary Principle*, to serve as a foil for the Precautionary Principle. However, this principle merely directs us to "[f]avor measures that are proportionate to the probability and magnitude of impacts, and that have a high expectation value."

[36] Admittedly, this is not a case of pure decision theoretic uncertainty, since Buchak's view is a moral claim that presupposes a particular decision theory, and not merely a claim about instrumental rationality under conditions of uncertainty.

[37] If we find risk-weighted expected utility theory plausible, we might think that, at under decision-theoretic uncertainty, the right thing is to maximize risk-weighted meta expected value. However, the point of the argument is to suggest that even those who are confident in orthodox decision theory may be required to act so as to maximize risk-weighted expected utility relative to a non-linear risk function given their decision-theoretic uncertainty.

More generally, there is a good case to be made that the view that agents are required to maximize risk-weighted expected utility but rationally permitted to use any risk function that is increasing and has stationary points at 0 and 1 can be ignored under conditions of decision-theoretic uncertainty.

Here is the argument. According to risk-weighted expected utility theory, a linear risk function is rationally permissible, as is any other risk function satisfying the constraints noted above. According to expected utility theory, only a linear risk function is rationally permissible; all others are impermissible. Under decision-theoretic uncertainty, choosing in accordance with a linear risk function weakly dominates choosing in accordance with a non-linear risk function. In terms of conforming to rational requirements, there is everything to lose and nothing to gain by choosing in accordance with a non-linear risk function. Since you ought to prefer weakly dominant options, you ought to align your preferences over gambles with expected utility theory. In particular, you rationally ought not weight potential downsides of a gamble out of proportion to their probability. We can therefore set aside a kind of view that might have led us to especially weight potential downsides of the gamble represented by continued human survival.[38]

[38] Here is one way to challenge the argument. Buchak (2016) notes that the rational requirement to maximize expected utility has historically been interpreted in one of two ways, paralleling contemporary debates about the scope of rational requirements (Way 2010). On the *narrow-scope* interpretation, the norm states that if the agent has a utility function, $u(\cdot)$, then she is rationally required to maximize the expectation of $u(\cdot)$. On the *wide-scope* interpretation, the rational requirement to maximize expected utility instead states that the agent is rationally required to make it the case that there is a utility function $u(\cdot)$ whose expectation she maximizes. The wide-scope interpretation has been dominant since the development of axiomatic decision theory in the middle of the twentieth century, because it does not require us to assume that the agent's utility function has an inherent cardinal structure prior to the conformity of her preferences to the axioms of expected utility theory. A similar distinction can be drawn when considering the norm to maximize risk-weighted expected utility. This could be taken to mean either, that if $r(\cdot)$ is your risk function, then you are required to maximize risk-weighted expected utility relative to $r(\cdot)$, or that you are required to make it the case that there is some risk function $r(\cdot)$ such that you maximize risk-weighted expected utility relative to r. If the requirement to maximize risk-weighted expected utility is interpreted as narrow-scope, then my argument fails. On this interpretation, if the agent's risk function is non-linear, she is required to maximize risk-weighted expected utility relative to a non-linear risk function. Maximizing risk-weighted expected utility relative to a linear risk function is forbidden for her. By contrast, if the norm is wide-scope, the argument goes through.

Taking account of normative uncertainty has the potential to impact the way we think about the value of the future in a variety of ways. An agent who aggregates evaluations across theories by calculating expected values can be expected to behave for the most part roughly as if she were certain of a pluralist axiology, and so will act as if she gives some degree of weight to the disvalue of inequality. Insofar as she is uncertain between total utilitarianism and positive critical level utilitarianism, she behaves roughly as if she endorsed a weaker form of the latter view. Both of these adjustments have the potential to alter the ways she assesses the prospect of continued human survival. It is possible that her confidence in average utilitarianism could go some way to cancelling out her confidence in positive critical level utilitarianism, and the large population swamping result due to Greaves and Ord might not tell as strongly against that idea as has been thought. Lastly, normative uncertainty in respect of how to make decisions under uncertainty may push us in the direction of making risk avoidant choices, even if we are confident we ought to be risk neutral with respect to moral value, albeit only insofar as there is a case to be made that risk avoidance is not merely permitted, but mandated.

4.4 Summary

We have considered a number of different theories about how to value uncertain prospects and how they might bear on our evaluation of continued human survival, considered as a gamble. Much of the discussion has focused on whether we might be required to be risk averse in respect of moral value, thus putting especial weight on the possible downsides associated with the persistence of humanity.

Whereas expected utility theory requires that agents are risk neutral in respect of utility, there are reasons to worry that a rational agent cannot have a utility function that is linear in moral value, and that any morally acceptable utility function may require the agent to be risk-seeking or risk-averse in respect of moral value in some imaginable contexts, in ways that might become especially relevant when considering grand questions like the continued survival of humanity. Some decision theories may permit or require risk aversion in relation to moral value even given a linear utility function. We discussed claims that REU theory with a convex risk function should be used in deciding on behalf of future people, and that so deciding favours human extinction in the near-term. We rejected each of these claims.

Lastly, we discussed normative uncertainty. We argued that asymmetries in the space of plausible axiological theories seem to push agents who evaluate possible worlds based on expected moral value aggregated over competing moral theories in the direction of a more pessimistic evaluation of continued human survival, but criticized appeals to large-population swamping results as guides to how to evaluate possible worlds under moral uncertainty, allowing that average utilitarianism and positive critical-level utilitarianism might cancel out in their evaluations of lives near the zero level. Lastly, we considered decision theoretic uncertainty, including an argument that if there is a decision-relevant sense of 'ought' such that our ranking of options ought to be guided by an assessment derived by aggregating evaluations across the different decision theories among which we are uncertain, then a case could be made that we ought to set aside risk avoidant evaluations of continued human survival entirely.

References

Adler, M. 2012. *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford: Oxford University Press.

Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axioms de l'ecole Americaine. *Econometrica 21*(4): 503–546.

Althaus, D. and L. Gloor. 2016. Reducing risks of astronomical suffering: A neglected priority. Center on Long-Term Risk: https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/.

Aristotle. 350 BCE [1980]. *The Nicomachean Ethics*, ed. L. Brown, transl. D. Ross. Oxford: Oxford.

Arnauld, P. and P. Nicole. 1662. *La Logique, ou l'Art de Penser*. Paris: Jean Guignart, Charles Savreux, Jean de Lavnay.

Arntzenius, F., A. Elga, and J. Hawthorne. 2004. Bayesianism, infinite decisions, and binding. *Mind 113*(450): 251–283.

Arrhenius, G. 2000. An impossibility theorem for welfarist axiologies. *Economics and Philosophy 16*(2): 247–266.

Arrhenius, G. 2013. Egalitarian concerns and population change. In *Inequalities in Health: Concepts, Measures, and Ethics*, eds. N. Eyal, S. A. Hurst, O. F. Norheim, and D. Wikler, 74–92. Oxford: Oxford University Press.

Arrhenius, G. 2014. *Population Ethics: The Challenge of Future Generations*. Unpublished manuscript.

Arrhenius, G. and J. Mosquera. 2022. Positive egalitarianism reconsidered. *Utilitas 34*(1): 19–38.

Barnosky, A. D., N. Matzke, S. Tomiya, G. O. U. Wogan, B. Swartz, T. B. Quental, C. Marshall, J. L. McGuire, E. L. Lindsey, K. C. Maguire, B. Mersey, and E. A. Ferrer. 2011. Has the Earth's sixth mass extinction already arrived? *Nature 471*(7336): 51–57.

Beckstead, N. 2013. On the overwhelming importance of shaping the far future. PhD Dissertation: Rutgers.

Beckstead, N. and T. Thomas. 2021. A paradox for tiny probabilities and enormous values. GPI Working Paper No. 7-2021, https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/.

Behroozi, P. and M. S. Peeples. 2015. On the history and future of cosmic planet formation. *Monthly Notices of the Royal Astronomical Society 454*(2): 1811–1817.

Benatar, D. 2006. *Better Never To Have Been: The Harm of Coming Into Existence.* Oxford: Oxford University Press.

Bentham, J. 1780 [1823]. *An Introduction to the Principles of Morals and Legislation.* London: W. Pickering.

Bernoulli, D. 1738. Specimen theoriae novae de mensura sortis. *Papers of the Imperial Academy of Sciences in Petersburg* 5: 175–192.

Blackorby, C., W. Bossert, and D. J. Donaldson. 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics.* Cambridge: Cambridge University Press.

Blessenohl, S. 2020. Risk attitudes and social choice. *Ethics 130*(4): 485–513.

Bostrom, N. 2003. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas 15*(3): 308–314.

Bostrom, N. 2013. Existential risk prevention as global priority. *Global Policy 4*(1): 15–31.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bradford, G. 2021. Perfectionist bads. *Philosophical Quarterly 71*(3): 586–604.

Bradley, B. 2010. Benatar and the logic of betterness. *Journal of Ethics and Social Philosophy 4*(2): 1–6.

Broome, J. 1991. Utility. *Economics and Philosophy 7*(1): 1–12.

Broome, J. 1999. *Ethics Out of Economics*. Cambridge University Press.

Broome, J. 2004. *Weighing Lives*. Oxford: Oxford University Press.

Broome, J. 2005. Should we value population? *Journal of Political Philosophy 13*(4): 399–413.

Browning, H. and W. Veit. 2021. Positive wild animal welfare. PhilSci Archive Preprint, http://philsci-archive.pitt.edu/19608/.

Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Buchak, L. 2016. Decision theory. In *The Oxford Handbook of Probability and Philosophy*, eds. C. Hitchcock and A. Hajek, 789–814. Oxford: Oxford University Press.

Buchak, L. 2017. Taking risks behind the veil of ignorance. *Ethics 127*(3): 610–644.

Buchak, L. 2019. Weighing the risks of climate change. *The Monist 102*(1): 66–83.

Buchak, L. 2022. How should risk and ambiguity affect our charitable giving? Global Priorities Institute Working Paper No. 8 - 2022, https://globalprioritiesinstitute.org/lara-buchak-how-should-risk-and-ambiguity-affect-our-charitable-giving/.

Bykvist, K. 2007. The benefits of coming into existence. *Philosophical Studies 135*(3): 335–362.

Cahen, H. 1988. Against the moral considerability of ecosystems. *Environmental Ethics 10*(3): 195–216.

Callicott, J. B. 1980. Animal liberation: A triangular affair. *Evnrionmental Ethics 2*(4): 311–338.

Callicott, J. B. 1989. *In Defense of the Land Ethic: Essays in Environmental Philosophy*. New York, NY: SUNY Press.

Carlson, E. 2007. Higher values and non-Archimedean additivity. *Theoria 73*(1): 3–27.

Carruthers, P. 2019. *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford: Oxford University Press.

Carvalho, M. R., C. Jaramillo, F. de la Parra, D. Caballero-Rodríguez, F. Herrera, S. Wing, B. L. Turner, C. D'Apolito, M. Romero-Báez, P. Narváez, C. Martínez, M. Gutierrez, C. Labandeira, G. Bayona, M. Rueda, M. Paez-Reyes, D. Cárdenas, Álvaro Duque, J. L. Crowley, C. Santos, and D. Silvestro. 2021. Extinction at the end-Cretaceous and the origin of modern neotropical rainforests. *Science 372*(6537): 63–68.

Caviola, L., D. Althaus, A. L. Mogensen, and G. P. Goodwin. 2022. Population ethical intuitions. *Cognition* 218: 104941.

Chalmers, D. J. 2022. *Reality +: Virtual Worlds and the Problems of Philosophy*. London: Allen Lane.

Cohen, G. A. 2012. Rescuing conservatism: A defense of existing value. In *Finding Oneself in the Other*, 143–174. Princeton, NJ: Princeton University Press.

Cooney, N. 2014. *Veganomics: The Surprising Science on What Motivates Vegetarians, from the Breakfast Table to the Bedroom*. New York, NY: Lantern Books.

Cowie, C. 2017. Does the repugnant conclusion have any probative force? *Philosophical Studies 174*(12): 3021–3039.

Crisp, R. and T. Pummer. 2020. Effective justice. *Journal of Moral Philosophy 17*(4): 398–415.

Cuddington, K. 2019. Insect herbivores, life history and wild animal welfare. https://rethinkpriorities.org/publications/insect-herbivores-life-history-and-wild-animal-welfare.

Darimont, C. T., S. M. Carlson, M. T. Kinnison, P. C. Paquet, T. E. Reimchen, and C. C. Wilmers. 2009. Human predators outpace other agents of trait change in the wild. *Proceedings of the National Academy of Sciences 106*(3): 952–954.

Darwall, S. 1996. *The Second Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

DeGrazia, D. 1996. *Taking Animals Seriously: Mental Life and Moral Status*. Cambridge: Cambridge University Press.

DeGrazia, D. 2010. Is it wrong to impose the harms of human life? A reply to Benatar. *Theoretical Medicine and Bioethics 31*(4): 317–331.

Dirzo, R., G. Ceballos, and P. Ehrlich. 2022. Circling the drain: the extinction crisis and the future of humanity. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377: 20210378.

Dirzo, R., H. S. Young, M. Galetti, G. Ceballos, N. J. B. Isaac, and B. Collen. 2014. Defaunation in the Anthropocene. *Science 345*(6195): 401–406.

Dreier, J. 1996. Rational preference: Decision theory as a theory of practical rationality. *Theory and Decision 40*(3): 249–276.

Easwaran, K. 2014. Decision theory without representation theorems. *Philosophers' Imprint 14*(27): 1–30.

Ewin, R. E. 1970. On justice and injustice. *Mind 79*(314): 200–216.

Faria, N. R., M. A. Suchard, A. Rambaut, D. G. Streicker, and P. Lemey. 2013. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B: Biological Sciences 368*(1614): 20120196.

Feldman, F. 1991. Some puzzles about the evil of death. *Philosophical Review 100*(2): 205–227.

Fishburn, P. 1970. *Utility Theory for Decision Making*. New York, NY: John Wiley Sons.

Francis, T. 2019. The implications of non-totalist population ethics for longtermism. Unpublished manuscript.

Franck, S., C. Bounama, and W. von Bloh. 2006. Causes and timing of future biosphere extinctions. *Biogeosciences 3*(1): 85–92.

Frick, J. 2015. Contractualism and social risk. *Philosophy & Public Affairs 43*(3): 175–223.

Frick, J. 2017. On the survival of humanity. *Canadian Journal of Philosophy 47*(2-3): 344–367.

Frick, J. 2022. Context-dependent betterness and the mere addition paradox. In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 232–263. Oxford: Oxford University Press.

Fukuyama, F. 1992. *The End of History and the Last Man*. Free Press.

Goodpaster, K. E. 1978. On being morally considerable. *Journal of Philosophy 75*(6): 308–325.

Goodsell, Z. 2021. A St Petersburg paradox for risky welfare aggregation. *Analysis 81*(3): 420–426.

Greaves, H. 2017. Population axiology. *Philosophy Compass 12*(11): e12442.

Greaves, H. and T. Ord. 2017. Moral uncertainty about population axiology. *Journal of Ethics and Social Philosophy 12*(2): 135–167.

Groff, Z. and Y.-K. Ng. 2019. Does suffering dominate enjoyment in the animal kingdom? An update to welfare biology. *Biology and Philosophy 34*(4): 40.

Gustafsson, J. 2020. The levelling-down objection and the additive measure of the badness of inequality. *Economics and Philosophy 36*(3): 401–406.

Hammond, P. J. 1998. Objective expected utility: A consequentialist perspective. In *Handbook of Utility Theory, Vol. 1*, eds. S. Barberà, P. J. Hammond, and C. Seidl, 143–211. Dordrecht: Kluwer.

Hanson, R. 2016. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford: Oxford University Press.

Hansson, B. 1988. Risk aversion as a problem of conjoint measurement. In *Decision, Probability and Utility: Selected Readings*, eds. P. Gärdenfors and N.-E. Sahlin, 136–158. Cambridge: Cambridge University Press.

Harman, E. 2009. Critical study: David Benatar. *Better Never to Have Been: The Harm of Coming Into Existence* (Oxford: Oxford University Press, 2006). *Noûs 43*(4): 776–785.

Harman, E. 2015. The irrelevance of moral uncertainty. *Oxford Studies in Metaethics* 10: 53–79.

Haybron, D. M. 2008. *The Pursuit of Unhappiness: The Elusive Psychology of Well-Being.* Oxford: Oxford University Press.

Hedden, B. 2016. Does MITE make right? Decision-making under normative uncertainty. *Oxford Studies in Metaethics* 11: 102–128.

Herzog, H. 2010. *Some We Love, Some We Hate, Some We Eat: Why it's So Hard to Think Straight About Animals.* New York, NY: HarperCollins.

Holtug, N. 2004. Person-affecting moralities. In *The Repugnant Conclusion: Essays on Population Ethics*, eds. T. Tännsjö and J. Ryberg, 129–161. Dordrecht: Kluwer Academic Publishers.

Holtug, N. 2010. *Persons, Interests, and Justice.* Oxford: Oxford University Press.

Horta, O. 2010. Debunking the idyllic view of natural processes: Population dynamics and suffering in the wild. *Telos: Revista Iberoamericana de Estudios Utilitaristas 17*(1): 73–90.

Hurka, T. 1982. Value and population size. *Ethics 93*(3): 496–507.

Hurka, T. 1993. *Perfectionism.* Oxford: Oxford University Press.

Hurka, T. 2010. Asymmetries in value. *Noûs 44*(2): 199–223.

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, B. Puranen, et al. (Eds.) 2014. *World Values Survey: Round Six - Country-Pooled Datafile Version*, https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. Madrid: JD Systems Institute.

Jaworska, A. and J. Tannenbaum. 2014. Person-rearing relationships as a key to higher moral status. *Ethics 124*(2): 242–271.

Kagan, S. 2014. An introduction to ill-being. *Oxford Studies in Normative Ethics* 4: 261–88.

Kagan, S. 2019. *How to Count Animals, More or Less.* Oxford: Oxford University Press.

Kant, I. 1785 [1998]. *Groundwork of the Metaphysics of Morals,* transl. M. Gregor. Cambridge: Cambridge University Press.

Kavka, G. 1978. The futurity problem. In *Obligations to Future Generations,* eds. R. I. Sikora and B. M. Barry, 186–203. Winwick: White Horse Press.

Kemp, L., C. Xu, J. Depledge, K. L. Ebi, G. Gibbins, T. A. Kohler, J. Rockström, M. Scheffer, H. J. Schellnhuber, W. Steffen, and T. M. Lenton. 2022. Climate endgame: Exploring catastrophic climate change scenarios. *Proceedings of the National Academy of Sciences 119*(34): e2108146119.

Kitcher, P. 1999. Essence and perfection. *Ethics 110*(1): 59–83.

Kitcher, P. 2000. Parfit's puzzle. *Noûs 34*(4): 550–577.

Knutsson, S. 2019. The world destruction argument. *Inquiry: An Interdisciplinary Journal of Philosophy* (10): 1004–1023.

Korsgaard, C. M. 1996. *Creating the Kingdom of Ends.* Cambridge: Cambridge University Press.

Kosonen, P. 2022. Tiny probabilities of vast value. D.Phil. thesis, Oxford University.

Kreps, D. M. 1988. *Notes on the Theory of Choice.* Abingdon: Routledge.

Kurzweil, R. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York, NY: Viking.

Lattimore, R. 1951. *The Iliad of Homer.* Chicago, IL: The University of Chicago Press.

Leopold, A. 1949. *A Sand County Almanac.* New York, NY: Oxford University Press.

Lockhart, T. 2000. *Moral Uncertainty and its Consequences.* Oxford University Press.

Lynas, M. 2020. *Our Final Warning: Six Degrees of Climate Emergency.* London: 4th Estate.

MacAskill, W. 2014. Normative uncertainty. D.Phil. thesis, Oxford University.

MacAskill, W. 2016. Smokers, psychos, and decision-theoretic uncertainty. *Journal of Philosophy 113*(9): 425–445.

MacAskill, W., K. Bykvist, and T. Ord. 2020. *Moral Unertainty*. Oxford University Press.

MacAskill, W. and T. Ord. 2020. Why maximize expected choice-worthiness?1. *Noûs 54*(2): 327–353.

MacAskill, W., A. Vallinder, C. Oesterheld, C. Shulman, and J. Treutlein. 2021. The evidentialist's wager. *Journal of Philosophy 118*(6): 320–342.

Marx, K. 1844 [2007]. *Economic and Philosophical Manuscripts*, transl. M. Milligan. Mineola, NY: Dover.

May, T. 2018. Would human extinction be a tragedy? *The New York Times*. 17 Dec 2018, https://www.nytimes.com/2018/12/17/opinion/human-extinction-climate-change.html.

Mayerfeld, J. 1996. The moral asymmetry of happiness and suffering. *Southern Journal of Philosophy 34*(3): 317–338.

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

McCauley, D. J., M. L. Pinsky, S. R. Palumbi, J. A. Estes, F. H. Joyce, and R. R. Warner. 2015. Marine defaunation: Animal loss in the global ocean. *Science 347*(6219): 1255641–1–1255641–7.

McGee, V. 1999. An airtight dutch book. *Analysis 59*(4): 257–265.

McMahan, J. 1981. Problems of population theory. *Ethics 92*(1): 96–127.

McMahan, J. 1988. Death and the value of life. *Ethics 99*(1): 32–61.

McMahan, J. 1996. Cognitive disability, misfortune, and justice. *Philosophy and Public Affairs 25*(1): 3–35.

McMahan, J. 2002. *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.

McMahan, J. 2013. Causing people to exist and saving people's lives. *The Journal of Ethics 17*(1-2): 5–35.

Menzel, C. 2021. Possible Worlds. In *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.)., ed. E. N. Zalta. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/possible-worlds/.

Mill, J. S. 1863. *Utilitarianism*. London: Parker, Son, and Bourn.

Monton, B. 2019. How to avoid maximizing expected utility. *Philosophers' Imprint* 19.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

More, M. 2004. The proactionary principle. The Extropy Institute, http://www.extropy.org/proactionaryprinciple.htm.

Moynihan, T. 2020. *X-Risk: How Humanity Discovered Its Own Extinction*. Falmouth: Urbanomic.

Mòzǐ. 5th-3rd century BCE [2020]. *The Essential Mòzǐ: Ethical, Political, and Dialectical Writings*, transl. Fraser. Oxford: Oxford University Press.

Murphy, M. C. 2001. *Natural Law and Practical Rationality*. Cambridge: Cambridge University Press.

Nagel, T. 1970. Death. *Noûs 4*(1): 73–80.

Nebel, J. M. 2019. Asymmetries in the value of existence. *Philosophical Perspectives 33*(1): 126–145.

Nebel, J. M. 2021. Rank-weighted utilitarianism and the veil of ignorance. *Ethics 131*(1): 87–106.

Nebel, J. M. 2022. Totalism without repugnance. In *Ethics and Existence: The Legacy of Derek Parfit*, eds. J. McMahan, T. Campbell, J. Goodrich, and K. Ramakrishnan, 200–231. Oxford: Oxford University Press.

Newcome, S. 1728 [1732]. *An Enquiry Into the Evidence of the Christian Religion*. London: William Innys.

Ng, Y.-K. 1989. What should we do about future generations? Impossibility of Parfit's theory X. *Economics and Philosophy 5*(2): 235–253.

Ng, Y.-K. 1995. Towards welfare biology: Evolutionary economics of animal consciousness and suffering. *Biology and Philosophy 10*(3): 255–285.

Nietzsche, F. 1901 [2017]. *The Will to Power*, transl. R. K. Hill and M. A. Scarpitti. London: Penguin Books.

Norwood, F. B. and J. L. Lusk. 2011. *Compassion, by the Pound: The Economics of Farm Animal Welfare*. Oxford University Press.

Nover, H. and A. Hájek. 2004. Vexing expectations. *Mind 113*(450): 237–249.

Nozick, R. 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.

Nozick, R. 1989. *The Examined Life: Philosophical Meditations*. New York, NY: Simon & Schuster.

Nussbaum, M. C. 2000. *Women and Human Development: The Capabilities Approach*. Cambridge University Press.

O'Neill, O. 1989. *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: Cambridge University Press.

Ord, T. 2021. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.

Pallies, D. 2022. Attraction, aversion, and asymmetrical desires. *Ethics 132*(3): 598–620.

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

Parfit, D. 1986. Overpopulation and the quality of life. In *Applied Ethics*, ed. P. Singer, 145–164. Oxford: Oxford University Press.

Parfit, D. 1991. *Equality or Priority?* The University of Kansas: The Lindley Lecture.

Parfit, D. 2011. *On What Matters, Volume 2*. Oxford: Oxford University Press.

Parfit, D. 2016. Can we avoid the Repugnant Conclusion? *Theoria 82*(2): 110–127.

Persson, I. 2001. Equality, priority and person-affecting value. *Ethical Theory and Moral Practice 4*(1): 23–39.

Persson, I. 2003. The badness of unjust inequality. *Theoria 69*(1-2): 109–124.

Peterson, M. 2004. From outcomes to acts: a non-standard axiomatization of the expected utility principle. *Journal of Philosophical Logic 33*(4): 361–378.

Pettigrew, R. 2022. Effective altruism, risk, and human extinction. Global Priorities Institute Working Paper No. 2-2022. https://globalprioritiesinstitute.org/wp-content/uploads/Pettigrew-Effective-altruism-risk-and-human-extinction-2.pdf.

Popper, K. R. 1961 [2011]. *The Open Society and Its Enemies.* Abingdon: Routledge.

Portmore, D. W. 1999. Does the Total Principle have any repugnant implications? *Ratio 12*(1): 80–98.

Powell, R. 2020. *Contingency and Convergence: Toward a Cosmic Biology of Body and Mind.* Cambridge, MA: MIT Press.

Quiggin, J. 1982. A theory of anticipated utility. *Journal of Economic Behavior Organization 3*(4): 323–343.

Rabinowicz, W. 2003. The size of inequality and its badness: Some reflections around Temkin's *Inequality*. *Theoria 69*(1-2): 60–84.

Rabinowicz, W. 2009. Broome and the intuition of neutrality. *Philosophical Issues 19*(1): 389–411.

Rabinowicz, W. and G. Arrhenius. 2015. The value of existence. In *The Oxford Handbook of Value Theory*, eds. I. Hirose and J. Olson, 424–444. Oxford: Oxford University Press.

Rachels, J. 2004. Drawing lines. In *Animal Rights: Current Debates and New Directions*, eds. C. R. Sunstein and M. C. Nussbaum, 162–174. Oxford: Oxford University Press.

Rachels, S. 1998. Is it good to make happy people? *Bioethics 12*(2): 93–110.

Ramsey, F. P. 1926. Truth and probability. In *Foundations of Mathematics and Other Logical Essays*, 156–198. Abingdon: Routledge.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political Liberalism*. New York, NY: Columbia University Press.

Riedener, S. 2020. An axiomatic approach to axiological uncertainty. *Philosophical Studies 177*(2): 483–504.

Roberts, M. A. 2011. An asymmetry in the ethics of procreation. *Philosophy Compass 6*(11): 765–776.

Rolston, H. 1988. *Environmental Ethics*. Philadelphia, PA: Temple University Press.

Ross, J. 2006. Rejecting ethical deflationism. *Ethics 116*(4): 742–768.

Russell, J. S. 2021. On two arguments for fanaticism. Global Priorities Institute Working Paper No . 17-2021, https://globalprioritiesinstitute.org/wp-content/uploads/Jeffrey-Sanford-Russell_On-two-arguments-for-Fanaticism.pdf.

Russell, J. S. and Y. Isaacs. 2021. Infinite prospects. *Philosophy and Phenomenological Research 103*(1): 178–198.

Russell, S. 2019. *Human Compatible: AI and the Problem of Control*. London: Penguin.

Saad, B. and A. Bradley. forthcoming. Digital suffering: Why it's a problem and how to prevent it. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Sagan, C. 1994. *Pale Blue Dot: A Vision of the Human Future in Space*. New York, NY: Random House.

Sandberg, A. forthcoming. *Grand Futures: Thinking Truly Long-Term*.

Savage, L. J. 1972. *Foundations of Statistics, Second Edition*. New York, NY: Dover.

Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge, MA: Harvard University Press.

Scheffler, S. 2013. *Death and the Afterlife*. Oxford University Press.

Scheffler, S. 2018. *Why Worry About Future Generations?* Oxford University Press.

Schopenhauer, A. 1850 [1970]. *On The Suffering of the World*, transl. R. J. Hollingdale. London: Penguin.

Sebo, J. 2018. The moral problem of other minds. *The Harvard Review of Philosophy* 25: 51–70.

Segall, S. 2016. *Why Inequality Matters: Luck Egalitarianism, Its Meaning and Value*. Cambridge: Cambridge University Press.

Segall, S. 2019. Why we should be negative about positive egalitarianism. *Utilitas 31*(4): 414–430.

Sepielli, A. 2009. What to do when you don't know what to do. *Oxford Studies in Metaethics* 4: 5–28.

Shepherd, J. 2018. *Consciousness and Moral Status*. New York, NY: Routledge.

Shulman, C. and N. Bostrom. 2021. Sharing the world with digital minds. In *Rethinking Moral Status*, eds. S. Clarke, H. Zohny, and J. Savulescu, 306–326. Oxford: Oxford University Press.

Sidgwick, H. 1906 [1981]. *The methods of ethics*, 7th. ed. Cambridge: Hacking.

Sikora, R. 1978. Is it wrong to prevent the existence of future generations? In *Obligations to Future Generations*, eds. R. I. Sikora and B. M. Barry, 112–166. Winwick: White Horse Press.

Singer, P. 1993. *Practical Ethics*, 2nd. ed. Cambridge: Cambridge University Press.

Smart, R. N. 1958. Negative utilitarianism. *Mind 67*(268): 542–543.

Sumner, W. 2020. The worst things in life. *Grazer Philosophische Studien 97*(3): 419–432.

Tännsjö, T. 2002. Why we ought to accept the repugnant conclusion. *Utilitas 14*(3): 339–359.

Tarsney, C. 2020. Exceeding expectations: Stochastic dominance as a general decision theory. GPI Working Paper No. 3-2020, https://globalprioritiesinstitute.org/christian-tarsney-exceeding-expectations-stochastic-dominance-as-a-general-decision-theory/.

Tarsney, C. and T. Thomas. 2020. Non-additive axiologies in large worlds. GPI Working Paper No. 9-2020, https://globalprioritiesinstitute.org/christian-tarsney-and-teruji-thomas-non-additive-axiologies-in-large-worlds/.

Tarsney, C. J. 2021. Vive la différence? Structural diversity as a challenge for metanormative theories. *Ethics 131*(2): 151–182.

Taylor, C. 1992. *The Ethics of Authenticity*. Harvard University Press.

Taylor, P. W. 1981. The ethics of respect for nature. *Environmental Ethics 3*(3): 197–218.

Temkin, L. S. 1993. *Inequality*. Oxford: Oxford University Press.

Thoma, J. 2019. Risk aversion and the long run. *Ethics 129*(2): 230–253.

Thoma, J. and J. Weisberg. 2017. Risk writ large. *Philosophical Studies 174*(9): 2369–2384.

Thomas, C. D. 2015. Rapid acceleration of plant speciation during the anthropocene. *Trends in Ecology Evolution 30*(8): 448–455.

Thomas, C. D. 2017. *Inheritors of the Earth: How Nature is Thriving in an Age of Extinction*. London: Allen Lane.

Thomas, T. 2018. Some possibilities in population axiology. *Mind 127*(507): 807–832.

Tomasik, B. 2015. The importance of wild-animal suffering. *Relations: Beyond Anthropocentrism 3*(2): 133–152.

Vallentyne, P. 2005. Of mice and men: Equality and animals. *The Journal of Ethics 9*(3-4): 403–433.

Varner, G. E. 1998. *In Nature's Interests: Interests, Animal Rights, and Environmental Ethics*. Oxford: Oxford University Press.

von Neumann, J. and O. Morgenstern. 1947. *Theory of Games and Economic Behavior, Second Edition*. Princeton, NJ: Princeton University Press.

Watkins, J. W. N. 1977. Towards a unified decision theory: A non-Bayesian approach. In *Foundational Problems in the Special Sciences: Part Two of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975*, eds. R. E. Butts and J. Hintikka, 345–379. Dordrecht: Springer Netherlands.

Way, J. 2010. The normativity of rationality. *Philosophy Compass 5*(12): 1057–1068.

Weatherson, B. 2014. Running risks morally. *Philosophical Studies 167*(1): 141–163.

Webb, S. 2002. *If the Universe Is Teeming with Aliens - Where Is Everybody? Fifty Solutions to Fermi's Paradox and the Problem of Extraterrestrial Life*. New York, NY: Copernicus Books.

Wells, H. G. 1895. *The Time Machine*. New York, NY: H. Holt and Company.

Williams, E. G. 2015. The possibility of an ongoing moral catastrophe. *Ethical Theory and Moral Practice 18*(5): 971–982.

Williamson, P. 2021. A new argument against critical-level utilitarianism. *Utilitas 33*(4): 399–416.