Longtermism in an Infinite World

Christian J. Tarsney and Hayden Wilkinson (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2022

GPI Working Paper No . 14-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Longtermism in an Infinite World*

Christian J. Tarsney† & Hayden Wilkinson†

Last updated: September, 2022

Abstract

The case for longtermism depends on the vast potential scale of the future. But that same vastness also threatens to undermine the case for longtermism: If the universe as a whole, or the future in particular, contain *infinite* quantities of value and/or disvalue, then many of the theories of value that support longtermism (e.g., risk-neutral total utilitarianism) seem to imply that none of our available options are better than any other. If so, then even apparently vast effects on the far future cannot in fact make the world morally better. On top of this, some strategies for avoiding this problem of "infinitarian paralysis" (e.g., exponential pure time discounting) yield views that are much less supportive of longtermism. In this chapter, we explore how the potential infinitude of the future affects the case for longtermism. We argue that (i) there are reasonable prospects for extending risk-neutral totalism and similar views to infinite contexts and (ii) many such extension strategies will still support the case for longtermism, since they imply that when we can only effect (or only *predictably* affect) a finite, bounded part of an infinite universe, we can ignore the unaffectable rest of the universe and reason as if the finite, affectable part were all there is.

*For helpful feedback on assorted versions of this chapter, we thank Teru Thomas, Riley Harris, Elliott Thornley, H. Orri Stefánsson, and participants at both the 2nd Oxford Workshop on Global Priorities and the 2019 Workshop on the Economics of Catastrophe in Oxford.

†Global Priorities Institute, University of Oxford. Comments welcome: hayden.wilkinson@philosophy.ox.ac.uk

1 Introduction

Longtermism is, very roughly, the thesis that what it's morally best to do is primarily determined by the potential effects of our actions on the far future.[1] The case for longtermism rests on the potentially *vast* scale of the future: Human-originating civilization could persist for millions of years or more, and could spread across a large portion of the accessible universe, resulting in an enormous number of future people. If we can affect the improve the welfare of those vastly many people (conditional on their existence), or if we can increase or decrease the probability that they come to exist, these effects might well have greater moral significance than the effects of our actions on the near future.

The most straightforward argument (but far from the only argument)[2] for longtermism rests on an axiology—a theory of how to rank outcomes and risky options morally—that is *additive*, *impartial*, and *risk-neutral*. Roughly, *additivity* means that the value of an outcome is a weighted sum of the values of every individual event or person's life in that outcome, and *impartiality* means that all locations receive the same weight in that sum (regardless of, for instance, their spatiotemporal location or relationship to a particular agent). These premises allow us to reason that, since the far future contains a potentially vast number of value locations (e.g., sentient beings or other persons), how things go in the far future potentially makes an enormous difference to the overall value of the outcome. *Risk neutrality* means that the value of a risky option is equal to the *expected value* of its outcome (i.e., a probability-weighted sum of the values of its various potential outcomes). This premise lets us reason that, even if we can only slightly affect the probabilities of a good vs. a bad long-term future for humanity, these small changes in probability can still be the primary determinant of the ranking of our options, since the stakes are so high. We will refer to the conjunction of these three principles as *risk-neutral totalism*.[3] (For a more precise definition, see §3.)

It is also possible, however, that the future is not merely vast but *infinite*—that it contains

[1] Note that this is a claim about what outcomes or options it would be *best* to bring about (an *axiological* claim), rather than about what we *ought* to do (a *deontic* claim). It may be that in some decisions we ought not bring about the best outcome/option, such as those in which doing so involves imposing an injustice or violating someone's rights; such is entirely consistent with longtermism as we've stated it.

[2] There are similarly compelling cases to be made for longtermism from various axiologies that are *averageist*, *egalitarian*, *person-affecting*, and/or *risk-sensitive* (see Thomas, 2019; Tarsney and Thomas, 2020; Buchak, 2022; Pettigrew, 2022; Greaves and MacAskill, 2021, §6).

[3] This label is convenient but potentially misleading: In the context of welfarist axiologies, the category of additive, impartial theories includes not just total utilitarianism but also critical-level and prioritarian axiologies. Risk-neutral totalism, as we are using the term, ranks risky options by their expected sum of *value* at particular locations, which need not be the same as the expected sum of *welfare* at particular locations, even if value is determined entirely by welfare.

infinitely many valuable events and infinite total value and/or disvalue.[4] And while the potential vastness of the future suggests that it is extremely morally important how our actions affect the long-term future, it is much less clear that the potential infinitude of the future carries the same implication. In particular, the possibility of an infinite future threatens to undermine any case for longtermism based on risk-neutral totalism. First, if the future (or the universe as a whole) contains infinite value and/or disvalue, and our actions have only finite effects, then nothing we do can ever affect the total impartially-weighted sum of value in the universe. This might be taken as a reason to reject either additivity or impartiality, since together they implausibly imply that none of our actions matter. Or, if we stand by these principles and bite the bullet on their apparent nihilistic implication, we must give up on longtermism and any other claims about what it's best to do. Second, the mere *possibility* of an infinite future implies that the *expected* total value of all our options is infinite or undefined. This similarly might lead us to reject at least one of additivity, impartiality, or risk neutrality, or alternatively to accept that these principles do not justify longtermism or any other substantive practical conclusion. (But note that, even on most alternative axiologies, analogous problems arise.)

There is a substantial body of research on the moral comparison of infinite worlds, in both philosophy and economics. Most proposals in this literature aim to extend additive theories of the value of outcomes from finite to infinite contexts—that it, to develop views that are additive in finite contexts while also delivering plausible verdicts in infinite contexts. Most, though not all, of these proposals also aim to retain some version of impartiality. And insofar as they consider risk (which many do not), the usual aim is similarly to extend *risk-neutral* theories from finite to infinite contexts. This literature has generated many sophisticated proposals, which show that it is possible to preserve the spirit of risk-neutral totalism while delivering at least some plausible verdicts in infinite contexts. Nevertheless, all such proposals have significant counterintuitive implications. And indeed, there are various impossibility results showing that *any* axiology for infinite worlds (not just those consistent with risk-neutral totalism) must carry some counterintuitive implications, and in particular must give up at least some of the *prima facie* attractive features our axiologies have in finite contexts.

The challenges of infinite axiology thus threaten the case for longtemism in two ways. First, they might lead us to simply give up on risk-neutral totalism, in favour of moral views that are

[4] This is implied by the influential (though still disputed) inflationary paradigm in cosmology (see Knobe et al., 2006, pp. 50-1). It is also implied by at least some versions of the dominant flat-λ cosmological model, by which the universe will persist forever in a state that is capable of generating life through statistical fluctuations (see Carroll, 2020, pp. 11-6). By either view, for any local physical phenomenon, the universe will contain infinitely many near-perfect duplicates, with probability 1.

less favourable to longtermism. (Analogously, they might lead us to give up on the many other axiologies that are favourable to longtermism but, for brevity, we will focus on risk-neutral totalism here.) For instance, we might conclude that the only escape from these challenges is to abandon impartiality, or to abandon the project of axiology entirely in favour of a particularly extreme form of non-consequentialism that recognises no moral reasons to make the world better. Second, if we do find a satisfactory extension of risk-neutral totalism to infinite contexts, it might turn out that when we *apply* this extended view, accounting for the potential infinitude of our actual circumstances, practical conclusions like longtermism that seemed inescapable when we were assuming the world to be finite are no longer supported.

This chapter will consider to what extent the challenges of infinite axiology in fact threaten the case for longtermism—in particular, the case for longtermism based on risk-neutral totalism. Our conclusions will be tentatively positive for longtermism: First, we survey some of the existing proposals for extending risk-neutral totalism and conclude that, while they all face costs, those costs are not severe enough to scuttle the project entirely. Second, we show that most extant extensions of risk-neutral totalism allow us, when we can only predictably affect a finite part of an infinite universe, to simply ignore the infinite unaffectable part of the universe and reason as if the finite affectable part were all that existed. Insofar as this is our actual situation, which it is to a good approximation, the risk-neutral-totalist case for longtermism can still go through even while accounting for the potential infinitude of the future. The possibility that our actions might have *infinite* predictable effects raises further challenges, but tends to strengthen the case for longtermism since those effects are almost certainly located in the far future, and any extension of risk-neutral totalism should regard them as overwhelmingly important. Perhaps the greatest challenge for longtermism in this vicinity, we will suggest, comes not from the world having infinite value *per se* but from the related possibility of options with *unbounded* finite effects and without finite expected values. While there are various promising ideas for extending risk-neutral totalism to evaluate such options, these proposals have yet to be integrated with proposals from the infinite axiology literature.

We proceed as follows. Section 2 describes our formal framework. Section 3 introduces two minimal principles that are implied by almost all extant views in infinite axiology. Section 4 will consider the extent to which these principles allow us to rely on finite ethical reasoning of the sort employed in the risk-neutral-totalist case for longtermism, given the circumstances and choices we actually confront. Section 5 considers how the possibility that our choices have infinite predictable effects on the far future affects the case for longtermism. Section 6 considers

to whether the difficulties of infinite axiology force us to reject risk-neutral totalism, and what implications this might have for the case for longtermism. Section 7 sums up and highlights some especially important questions for future research.

2 Formal framework

Let's first introduce some terminology and notation (adapted from Wilkinson (2021a, 2022b)).

We assume, first, a domain $\mathcal{O}$ of *possible worlds* or *outcomes*. Each world contains some set of *value locations*, or simply *locations*, with which valuable events are associated. A location is a token entity of some common type that can exist (or have counterparts) across different outcomes. Locations might be persons, or person-stages, or positions in space and time, or something else.[5] Whatever locations are, there is an infinite set $\mathcal{L}$ of all possible locations. We assume that the value of an outcome is determined, in one way or another, by which locations exist, the value realised at each location, and perhaps other features of locations (e.g., their relative positions in time). And we assume that the value realised at each location can be represented by a real number, in a way that is order-preserving (i.e., that greater numbers corresponding to greater degrees of value) and unique at least up to positive affine transformation (so that the numbers carry meaningful information about the relative sise of *differences* in value). Let $\mathcal{V} \subseteq \mathbb{R}$ represent the possible degrees of value that can be realised at locations. Then each outcome O_i determines a local value function $V_i : \mathcal{L} \to \mathcal{V} \cup \{\Omega\}$ that specifies the value realised at each location l in outcome O_i, with Ω representing the non-existence of the location in that outcome.

We also wish to compare lotteries (probability distributions) over outcomes, which correspond to the options from which real-world agents must choose under conditions of risk. The set of all possible such lotteries is denoted by $\mathcal{P}$. For any lottery L_i, its probability of resulting in some set $\mathcal{O}'$ of outcomes is given by $L_i(\mathcal{O}')$, and its domain of outcomes with non-zero probability by $\mathcal{O}_i$. To keep the notation in check, we will abbreviate $L(\{O\})$ to $L(O)$ when denoting the probability of a single outcome O. And, when a lottery results in some outcome O with probability 1, we denote both outcome and lottery by O.

An axiology is an evaluative ranking of outcomes and lotteries on outcomes. We assume that

[5] For a defence of adopting persons as the appropriate type, see Askell (2019). For arguments in favour of adopting spacetime positions, see Wilkinson (nd) and Wilkinson (2021b).

these two rankings must be consistent in the sense that one outcome is better than another just in case a lottery yielding the first outcome with probability 1 is better than a lottery yielding the second with probability 1. Thus we use $\succcurlyeq$ (read, "is at least as good as") to represent both the ranking of outcomes and the ranking of lotteries. The relation $\succcurlyeq$ is a preorder: a binary relation that is reflexive and transitive, but not necessarily complete. As usual, $\succ$ is the asymmetric part of $\succcurlyeq$ (representing strict betterness) and $\sim$ is the symmetric part (representing equal goodness).

3 Two consensus principles

In this section we consider principles for extending risk-neutral totalism to infinite contexts. Formally, risk-neutral totalism can be expressed as the following thesis:

> *Risk-neutral Totalism*: For any outcomes O_a and O_b, if O_a has greater total value than O_b, then $O_a \succ O_b$. If they have the same (finite) total value, then $O_a \sim O_b$. Likewise, for any lotteries L_a and L_b, if L_a has greater expected total value than L_b, then $L_a \succ L_b$. If they have the same (finite) expected total value, then $L_a \sim L_b$.

We are looking for principles, then, that extend risk-neutral totalism in the sense of implying these conditionals, while also implying at least some further comparisons—particularly in cases where total value or its expectation are infinite or undefined. And, less formally, we also want a view that preserves the *spirit* of risk-neutral totalism—that is, the spirit of the underlying (albeit imprecisely stated) principles of additivity, impartiality, and risk neutrality.

Our foil in this search is a view we will call *naive* risk-neutral totalism.

> *Naive Risk-Neutral Totalism*: For any outcomes O_a and O_b, $O_a \succcurlyeq O_b$ if and only if O_a has greater total value than O_b. Likewise, for any lotteries L_a and L_b, $L_a \succcurlyeq L_b$ if and only if L_a has greater expected total value than L_b.

Why is it naive? Suppose two outcomes (or lotteries) each have positive infinite (expected) total value—suppose even that both outcomes (lotteries) contain precisely the *same* locations, and *every* location has greater (expected) value in one than in the other. Naive risk-neutral totalism does not allow that either outcome (lottery) is better.

For less naive views, there have been many proposals in the literature for extending the totalist ranking of outcomes and/or the *risk-neutral* totalist ranking of lotteries (e.g., Vallentyne, 1993; Vallentyne and Kagan, 1997; Liedekerke and Lauwers, 1997; Bostrom, 2011; Arntzenius, 2014; Jonsson and Voorneveld, 2018; Wilkinson, 2021b; Clark, nd). But, rather than describe them each in detail, we will examine a pair of uncontroversial principles that almost all of them uphold. As we will see, these principles by themselves go a long way toward rescuing risk-neutral-totalist reasoning from the threat of infinities.

The first of these principles we will call *Sum of Differences*. It says that: we can compare two outcomes by summing up the *differences* in value at each value location, as long as this sum is well-defined.

> *Sum of Differences (SoD)*: For any outcomes O_a and O_b, a sufficient condition for $O_a \succ O_b$ is that
>
> $$\sum_{l \in \mathcal{L}} \Big(V_a(l) - V_b(l) \Big) > 0$$
>
> either by converging unconditionally to a non-negative value, or by diverging unconditionally to $+\infty$, with $\Omega = 0$ (i.e., non-existence of a location is treated as equivalent to existence with value 0). Likewise, if this sum is equal to 0, then $O_a \sim O_b$.[6]

To illustrate, consider the following pair of outcomes, O_a and O_b. And note that naive risk-neutral totalism says that neither is better than the other.

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	$\cdots$
O_a :	0	1	1	1	1	1	1	$\cdots$
O_b :	1	0	0	1	1	1	1	$\cdots$
$V_a - V_b$:	-1	1	1	0	0	0	0	$\cdots$

Sum of Differences says that we can compare O_a and O_b by summing the numbers in the bottom row, as long as that sum is well-defined. In this case, since the sum of positive, we can conclude that O_a is strictly better than O_b. Importantly for our purposes, in cases like this where two outcomes differ at only finitely many locations, Sum of Differences implies that we can equally well compare the two outcomes by comparing their subtotals of value *at just those*

[6] This principle is presented and defended by Vallentyne and Kagan (1997, p. 11), Lauwers and Vallentyne (2004, p. 21), and Basu and Mitra (2007).

locations where they differ. Thus, from the fact that the subtotal value of O_a from l_1 to l_3 is 2, and the corresponding subtotal for O_b is 1, we can conclude that O_a is strictly better than O_b.

While Sum of Differences compares only some pairs of infinite outcomes, the comparisons it does imply are all highly plausible, insofar as one finds additivity and impartiality plausible in finite contexts. Unsurprisingly, then, almost every proposal for extending impartial additive axiologies to infinite contexts implies Sum of Differences (with respect to its preferred kind of value locations, e.g. persons or spacetime positions).[7]

But Sum of Differences says nothing about how to compare lotteries. For that purpose, we can extend it to a principle we will call *Sum of Value-Probability Differences* (SVPD). To state this principle, let $L\big(V(l) = v\big)$ denote lottery L's probability of yielding an outcome with value v at location l.[8]

> *Sum of Value-Probability Differences (SVPD)*: For any lotteries L_a and L_b, $L_a \succ L_b$ if
>
> $$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} v \times \Big(L_a\big(V(l) = v\big) - L_b\big(V(l) = v\big)\Big) > 0$$
>
> either by converging unconditionally to a positive value or by diverging unconditionally to $+\infty$, with $\Omega = 0$ (i.e., non-existence of a location is treated as equivalent to existence with value 0). Likewise, if this sum is equal to 0, then $L_a \sim L_b$.

Informally, this principle tells us to consider, for each pair of a degree of value and a possible location, the difference in the probability of that degree of value being realised at that location if L_a is chosen vs. if L_b is chosen. We then multiply these probability differences by the degree of value concerned, and sum these terms across both locations and degrees of value to obtain an overall ranking of the lotteries. Importantly, however, SVPD yields a comparison only if this sum converges unconditionally, i.e., regardless of the order in which the terms are summed.

The infinite axiology literature doesn't contain as many proposals for comparing lotteries as it does for comparing outcomes. But every such proposal, if combined with SoD, implies SVPD.[9]

[7] The only exceptions we know of are the proposals of Liedekerke and Lauwers (1997), Clark (nd), and Bader (nd), which all violate the Pareto principle (see §5) with respect to any possible kind of location.

[8] We assume for simplicity that lotteries are discrete, and that the set $\mathcal{V} \subseteq \mathbb{R}$ of possible degrees of value at particular locations is countable.

[9] In fact, if combined with SoD, every such proposal strengthens SVPD by constraining the order of summation. The proposals of Arntzenius (2014, pp. 55-6), Bostrom (2011, pp. 27-30), and Meacham (2020) strengthen it to satisfy what could be called *Sum of Differences in Expectations*: that two lotteries can be compared by first finding the difference in expected value at each location, then SoD over the locations' expected values.

Like SoD, then, SVPD is a relatively weak principle that should be mostly uncontroversial insofar as our goal is to extend risk-neutral totalism to infinite contexts.

To illustrate SVPD, consider the following pair of lotteries:

	$L_1(O_i)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	$\cdots$
L_1	$\frac{1}{2}$	O_1 :	2	2	0	1	1	1	1	1	1	$\cdots$
	$\frac{1}{2}$	O_2 :	2	0	0	0	0	0	0	0	0	$\cdots$

	$L_2(O_i)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	$\cdots$
L_2	$\frac{1}{2}$	O_3 :	2	0	2	1	1	1	1	1	1	$\cdots$
	$\frac{1}{2}$	O_4 :	0	0	0	0	0	0	0	0	0	$\cdots$

Here, L_1 and L_2 yield the same prospects for all locations except l_1–l_3. Importantly, this need not imply that these locations are *unaffected* by the choice of lottery. For instance, perhaps the value realised at these locations depends on a fair coin flip, and choosing L_1 will cause these locations to have value 1 iff the coin lands heads, while choosing L_2 will cause them to have value 1 iff the coin lands tails. But the choice of lottery does not affect the *probability distribution* over outcomes for any of these locations. Thus, for all $i > 3$ and all v, $L_a\big(V(l_i) = v\big) - L_b\big(V(l_i) = v\big) = 0$. This allows us, in applying SVPD, to simply ignore all these locations, and compare L_a with L_b by comparing their expected subtotals of value at locations l_1–l_3, as long as these are finite. Since L_1 has an expected subtotal value of 3 at these locations, and L_2 has an expected subtotal value of 2, ICPD tells us that $L_1 \succ L_2$. And in general, when two lotteries yield the same local prospects at all but finitely many locations, SVPD allows us to compare those lotteries by comparing their expected subtotals at that finite set of locations.

Wilkinson (2022b, p. 14) strengthens SVPD to satisfy what could be called *Expected Sum of Differences*: that two lotteries can be compared by the expectation of their sum of differences (i.e., the expectation of the sum in SoD). It turns out that these two stronger principles are incompatible: the first principle implies *ex ante* Pareto while the second implies statewise dominance; but, in infinite contexts, these two principles can conflict (Wilkinson, 2022b, p. 8). Because SVPD requires unconditional convergence, it is neutral in these hard cases, and compatible with either *ex ante* Pareto or statewise dominance.

4 Do the consensus principles let us ignore real-world infinities?

We now have two modest and plausible principles for comparing options in an infinite world. Each of them yields at least some verdicts in infinite cases that naive risk-neutral totalism can't handle. Specifically, we have seen that these principles let us compare pairs of outcomes (resp., lotteries) in which local outcomes (resp., prospects) differ at only finitely many locations by ignoring all the locations where there's no difference and applying naive risk-neutral-totalist reasoning to the finite remainder, as if only those locations existed. In this section, we will consider whether this is enough to recover the real-world practical implications that we would expect from naive risk-neutral totalism if the world were finite, including in particular the risk-neutral-totalist case for longtermism as described earlier.

Let's say that an infinite axiology $\succcurlyeq$ is *practically equivalent* (to naive risk-neutral totalism in a finite world) if and only if there is some finite set of possible locations such that, for any pair of actual lotteries L_1, L_2 faced by an agent in a real-world choice situation, if L_1 has a greater expected subtotal of value than L_2 *at those locations*, then $L_1 \succ L_2$. We have already seen that any axiology satisfying SVPD will satisfy practical equivalence if the pairs of lotteries we are called upon to compare in practice yield different prospects at only finitely many locations.[10]

Assuming (as we will) that the correct infinite axiology satisfies SVPD, whether it is practically equivalent to naive risk-neutral totalism then depends on whether the lotteries corresponding to our real-world options actually have this feature. The answer to this question will depend in part on what kind of probability is morally relevant—the same real-world option might be associated with one distribution of objective chances over outcomes, a different distribution of evidential probabilities, and yet another distribution of subjective credences. We will focus on evidential probabilities: how probable that outcome is to result from a given option, on the present evidence of the agent deciding between that option and others (see Williamson, 2000, p. 209). (But much of what we say will carry over to subjective credences.) So, we want to know: Is our real-world situation such that the pairs of options we are required to evaluate beget different evidential probability distributions over local value at only finitely many locations?

[10] Of course, it is possible that every individual real-world option pair yields different local prospects at only finitely many locations, but that the set of possible locations whose prospects are affected by *some* real-world choice is infinite. This is ruled out, however, if we assume that the number of real-world choices and options is finite—or limit our focus to a finite set of real-world choices each involving only finitely many options (e.g., all of the choices faced by human agents in the 21st century).

The answer to this question depends on our empirical evidence concerning how much of the universe we can affect and in what ways. There is, on the one hand, substantial empirical reason to believe that, even if the universe is infinite, we can only affect a finite part of it. In particular, the impending heat death of the universe seems to promise an end to life as we know it, only finitely far in the future. And given that causal signals cannot travel faster than the speed of light, there is only finitely much room in our causal future for value to occupy, before heat death overtakes us.

On the other hand, there are various live hypotheses that do allow for infinite quantities of moral value in our causal future: For instance, some multiverse hypotheses imply that events in one 'universe' affect events in other universes, or even bring new universes into existence (as in Lee Smolin's 'cosmological natural selection' model; see Smolin, 1992), which suggests that we can affect the moral value of events beyond the heat death of our local 'universe'. Other hypotheses suggest that a future civilization may someday be able to perform infinite computations, potentially simulating an infinite number of minds, within the finite spatiotemporal limits of our pre-heat death future light cone (Earman and Norton, 1993; Tipler, 1994). Finally, and perhaps most straightforwardly, some versions of the dominant cosmological model (called the flat-λ model) imply that morally valuable life will not cease upon heat death: individual brains, civilisations, and even galaxies will continue to be generated by random fluctuations, sometimes called *Boltzmann brains* or *Boltzmann universes* (Carroll, 2020, p. 10). And the manner and timing of those fluctuations may be affected (albeit likely not in a predictable way) by our present actions—such fluctuations can be altered by even subtle changes in gravity (as by the Hawking effect) and electric field strength (as by the Casimir effect).

None of these hypotheses represent established physics and, in general, the claim that there are infinitely many value locations in our causal future seems on a much weaker footing than the claim that there are infinitely many value locations in the universe as a whole. Our own impression is that, of the hypotheses surveyed above, the Boltzmann brain hypothesis is by a significant margin the closest thing to a plausible implication of established physical theories.[11] And if the only source of infinite value and disvalue in our causal future is Boltzmann brains that arise by random fluctuations after the heat death of the universe, this seems to leave us in the happy condition where an infinite axiology satisfying SVPD will be practically equivalent to naive risk-neutral totalism: While our present choices may determine which Boltzmann brains

[11] But, as Carroll (2020) explains, a universe eternally capable of generating Boltzmann brains is only implied by *some* versions of the flat-λ model, and we may well have reason to reject these versions exactly *because* they imply the existence of infinitely many future Boltzmann brains.

come to exist and what experiences they have, the evidential probability of any particular event after the heat death of the universe (e.g., a particular Boltzmann brain existing at a given position in spacetime) does not vary from option to option. At least, it is very hard to see how our evidence could distinguish our options in this way. Thus, any two options in present-day choice situations will yield the same local prospects for all possible locations after the heat death of the universe, and SVPD therefore allows us to simply ignore these locations.[12]

On the other hand, hypotheses on which our descendants may be able intentionally create new universes (as in cosmological natural selection) or perform computational supertasks do allow us to *predictably* affect infinitely many locations (i.e., affect their prospects). For instance, by increasing the probability that humanity survives the coming century, we increase the probability that our descedants will someday deploy these technologies, and thereby increase the probability of existence for infinite numbers of potential persons. Similarly, attempts to change the institutions or future values of human-originating civilization might increase or decrease the probability that a civilization with these infinitary capacities would choose to use them. It is debatable whether our ordinary choices have any effect on the evidential probability of humanity's long-term survival or of particular values prevailing in the far future. But these possibilities of indirect but predictable infinite effects are particularly relevant to the case for longtermism, since any choices that *do* affect the long-term future of humanity (e.g., by affects its odds of survival or its future values) seem like to have at least some effect on the evidential probability of our descendants developing and deploying technologies that could create infinite value or disvalue. So it seems that, at least in the present-day choice situations to which the longtermist thesis applies (where our choices do affect the long-term future of humanity), our choices probably make a non-zero—though perhaps *extremely* small—difference to the evidential prospects of infinitely many potential value locations.[13]

[12]Similar things can be said of certain multiverse hypotheses—for instance, if our "universe" interacts gravitationally with other universes in a higher-dimensional space, assuming we are not in a position to know anything about the empirical details of our effects on other universes.

[13]Another factor besides our cosmological evidence that determines what lotteries we face is whether the correct decision theory is causal or evidential. We have so far implicitly assumed a causal decision theory on which our choices only make a difference to the outcomes and prospects of locations in our causal future. But if evidential decision theory or some other non-causal decision theory is correct, then our options can yield different local prospects at locations outside our causal future (for instance, because our choices give us evidence about the choices of our doppelgängers in distant parts of the universe). So this is another way in which our choices might make a difference to infinitely many local prospects. Indeed, even non-zero credence in evidential decision theory might have this consequence, if we treat our uncertainty between causal and evidential decision theory in the same way as empirical uncertainty (see MacAskill (2016), MacAskill et al. (2021)). This would reinforce the conclusion in the main text that our choices may affect infinitely many local prospects, but perhaps only very slightly (if our credence in non-causal decision theories is only slight).

5 Infinite and unbounded effects

Recall that the challenge of infinite axiology threatens the case for longtermism in two ways: 1) because it may force us to abandon risk-neutral totalism and with it the risk-neutral-totalist case for longtermism (and similarly, despite our focus here, it may force us to abandon various other axiologies that support longtermism too); and 2) because, if we do find a satisfactory way of extending risk-neutral totalism to infinite contexts, the practical implications of this extended view might deviate from the implications of risk-neutral totalism in a finite universe. The last two sections have gone some way toward mitigating both worries: We have seen that there are existing proposals for extending risk-neutral totalism that, in virtue of implying SoD and SVPD, can deliver at least some plausible verdicts in infinite contexts, rescuing us from universal infinitarian paralysis. And we have seen that when we can only affect the prospects of finitely many locations in an infinite universe, these principles yield the same practical implications that we would get by simply applying naive, finitary risk-neutral totalism to that finite part of the universe.

Nonetheless, both worries remain live. While extant proposals for extending risk-neutral totalism have some attractive features, they also have significant drawbacks, some of which (as we will see) are inescapable. And since we cannot rule out hypotheses that would allow us to predictably affect infinitely many value locations, it is not *quite* true that our actions only affect finitely many local prospects, so SVPD alone does not guarantee that the true infinity axiology will be practically equivalent to naive risk-neutral totalism. In this section and the next we will consider these remaining worries, in reverse order.

First, then, suppose that (an extension of) risk-neutral totalism is true, and more specifically that SVPD is true. But suppose also that we could conclude that our choices do affect the local prospects of *infinitely* many locations, at least slightly. What practical implications does this have, particularly with respect to the case for longtermism?

This is a hard question to answer in general, partly because there are many importantly distinct ways in which our choices might affect infinitely many prospects. But examination of a few particular cases will be enough to illustrate three general points: First, infinite predictable effects (i.e., affecting infinitely many local prospects) are not always problematic—in some cases, it is possible to rank pairs of lotteries with this feature in a way that is principled, intuitively plausible, and in the spirit of risk-neutral totalism (as illustrated below). Second, insofar as we

can make comparisons in these situations, the possibility of infinite predictable affects will tend to *strengthen* the risk-neutral-totalist case for longtermism, since (i) risk-neutral totalists should generally give absolute priority to infinite effects over finite effects and (ii) these infinite effects will tend to be located in the future. But, third, there are some kinds of infinite predictable effects, which we plausibly face in real-world choice situations, where it is intuitively unclear how to rank our options, where no ranking is given by modest principles like SVPD, and where it is at least conceivable that our options are simply incomparable. The primary way in which infinite predictable effects might threaten the risk-neutral-totalist case for longtermism, then, is by implying that, at least in those situations where our choices affect the long-term future, we face widespread incomparability, with no available option being better or worse than any other.[14]

To illustrate these points, let's start with the easy cases of infinite predictable effects, and work our way toward the harder cases. First, there are cases of infinite predictable effects that SVPD ranks easily. For instance, suppose that there is some potential future population at infinitely many locations, each of whom will certainly have positive value if they exist, and that you can increase the probability that they come to exist without changing their prospects conditional on existence.

$$L_1 \left\{ \begin{array}{c|cccccccccc} L_1(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.5 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ 0.5 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

$$L_2 \left\{ \begin{array}{c|cccccccccc} L_2(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.51 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ 0.49 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

In this case, the sum from the definition of SVPD—of $v \times \Big(L_2\big(V(l) = v\big) - L_1\big(V(l) = v\big)\Big)$ for each location l and possible local value v—diverges unconditionally to $+\infty$ (bearing in mind that we treat Ω as 0). So, SVPD implies that $L_2 \succ L_1$.

There are other cases in which it seems clear which of two lotteries is better, that are not ranked by SVPD, but that can be handled by natural and plausible strengthenings of SVPD. For

[14] While this conclusion might either refute longtermism or make it trivially true, depending on how the longtermist thesis is formulated, it would clearly violate the spirit of longtermism to conclude that we can never improve (the prospects of) the world as a whole by improving (the prospects of) the long-term future.

instance, consider L_3 versus L_4, which are similar to the above but, this time, you can improve their prospects of local value conditional on existence, without changing the probability that they come to exist in the first place.

L_3

$Pr(O)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	$\cdots$
0.05	O_2 :	2	2	2	2	2	2	2	2	2	$\cdots$
0.05	O_1 :	1	1	1	1	1	1	1	1	1	$\cdots$
0.9	O_Ω :	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	$\cdots$

L_4

$Pr(O)$		l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9	$\cdots$
0.051	O_2 :	2	2	2	2	2	2	2	2	2	$\cdots$
0.049	O_1 :	1	1	1	1	1	1	1	1	1	$\cdots$
0.9	O_Ω :	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	$\cdots$

Clearly L_4 is better than L_3. But SVPD is silent: Because L_4 increases each location's probability of realising value 2 while decreasing each location's probability of realising value 1, the value-weighted sum of location-outcome probability differences is non-convergent. But this can be remedied by strengthening SVPD, allowing outcomes at each location to be compared to a "baseline" outcome for that location.

> *Baseline-Adjusted Sum of Value-Probability Differences*: For any lotteries L_a and L_b, , $L_a \succ L_b$ if there exists an outcome $O_b \in \mathcal{O}$ such that
>
> $$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} \Big(v - V_b(l)\Big)\Big(L_a\big(V(l)=v\big) - L_b\big(V(l)=v\big)\Big) > 0,$$
>
> either by converging unconditionally to a positive value, or by diverging unconditionally to $+\infty$ (with $\Omega = 0$). Likewise, if there is an outcome O_b for which this sum is equal to 0, then $L_a \sim L_b$.

If we choose O_1 as the baseline outcome O_b, and substitute L_4 and L_3 for L_a and L_b respectively, we find that the above sum diverges unconditionally to $+\infty$. So we can conclude that $L_4 \succ L_3$. And this baseline-adjusted principle, while slightly more complicated than SVPD, is similarly modest and uncontroversial.[15]

[15] In particular, like SVPD, it follows from the proposals in Wilkinson (2022b, p. 14), Arntzenius (2014, pp. 55-6), Bostrom (2011, pp. 27-30), and Meacham (2020), together with Sum of Differences.

In both these cases, the principles we have appealed to imply that one lottery is "infinitely better" than another in the sense that no finite improvement of the worse lottery (or worsening of the better lottery) could affect the comparison. (For instance, if we add any finite number of locations that will realise value 1 for sure under L_1 or L_3, and 0 for sure under L_2 or L_4, the ranking would be unchanged.) Correctly evaluating this sort of infinite improvement requires some extension of naive risk-neutral totalism.[16] But, in general, the possibility of such unambiguous infinite improvements *strengthens* the risk-neutral-totalist case for longtermism. Why? First, any infinite axiology in the spirit of risk-neutral totalism should be *fanatical* about infinite improvements: Shifting any amount of probability from an infinitely worse outcome to an infinitely better outcome should take precedence over any finitary considerations, in the evaluation of lotteries (see Beckstead and Thomas, nd; Wilkinson, 2022a). And second, if there is any evidential probability of our choices having such infinite effects, it is almost all in the far future: The infinitely-better and infinitely-worse trajectories whose probabilities we can affect will, presumably, either unfold over infinite future time, or require far-future technology (e.g., computers that can perform supertasks in finite time), or both.[17]

More generally, it seems to us that the possibility that we face choices between lotteries that yield different local prospects at infinitely many locations does not threaten the case for longtermism *as long as these lotteries can be compared*. As a rough argument: One version of the longtermist thesis is that our options typically differ more in far-future value than in near-future value. Suppose we believed this thesis while assuming that our choices only (predictably) affect finitely many value locations in the far future, but then come to believe that our choices affect infinitely many locations in the far future (without changing our beliefs about their effects on the near future). It seems unlikely (though not impossible) that this realization should *reduce* the typical differences in far-future value between our options. This leaves two possibilities: One is that it amplifies those differences (or at least leaves them unchanged), thereby strengthening the case for longtermism (or at least leaving it unweakened). The other, however, is that we find that we can no longer compare the far-future effects of our options.

There are, unfortunately, many hard cases in infinite axiology that are not resolved by simple principles like SVPD, where it is not obvious how we should rank two outcomes or lotteries, and

[16] In both cases, the expected total value of both lotteries if $+\infty$ (or undefined, if we do not countenance infinite expectations), so naive risk-neutral totalism rules that the two lotteries are equally good (or simply fails to compare them).

[17] Another conceivable source of infinite stakes that are not clearly located in the far future is supernatural—in particular, affecting the probabilities that particular individuals achieve infinitely good vs. infinitely bad afterlives. Whether these considerations count for or against longtermism depends on whether these possible afterlives are temporal, and whether they stand in temporal relations to the present.

where incomparability is plausible. Here are two examples. First, suppose your choice affects that probability that some infinite future population will come to exist (say, within an infinite simulation or a "baby universe" of the sort envisioned by cosmological natural selection), and you know that if it does exist, that population will contain both infinitely many locations with positive value (e.g., persons with lives worth living) and infinitely many locations with negative value (e.g., persons with lives worth not living).

$$L_5 \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.5 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \cdots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

$$L_6 \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.51 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \cdots \\ 0.49 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

Second, suppose that your choice does not affect the probability that such an infinite future population comes to exist, you believe it to be *identity-affecting*: that is, which particular locations will compose that population depends on your chioce.

$$L_7 \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.5 & O_1: & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & 1 & \cdots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

$$L_8 \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.5 & O_1: & \Omega & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & \cdots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \cdots \end{array} \right.$$

Third and finally, suppose you can affect the probability that this infinite future population will be governed by one set of norms, institutions, or values rather than another, e.g. by having a potentially persistent effect on present-day values. For instance, you might affect the likelihood that, in that future population, greater weight is given to the welfare of individuals with greater cognitive capacities, which might result in a wider range of individual welfare levels.

$$L_9 \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.5 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \cdots \\ 0.5 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \cdots \end{array} \right.$$

$$L_{10} \left\{ \begin{array}{c|cccccccccc} Pr(O) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \cdots \\ 0.51 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \cdots \\ 0.49 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \cdots \end{array} \right.$$

None of these three cases are resolved by SVPD (or by the stronger, baseline-adjusted version discussed above). Nor is it there an intuitively clear right answer in any of these cases.[18] That doesn't mean that these are necessarily genuine cases of incomparability—it's possible to articulate principles that deliver verdicts in cases like these, especially if those principles are allowed to take account of the spatiotemporal arrangement of locations (see, e.g., Wilkinson, 2021b). But it is at least plausible that the kinds of tradeoffs involved in these cases do create incomparability (especially since, as we will see in the next section, there are compelling formal arguments that there must be at least some incomparability in infinite axiology). And it is plausible that we face tradeoffs like these in our real-world choices—at least, in those choices that have some predictable effect on the long-run future, which may affect the probabilities of infinite future populations coming to exist (e.g., by affecting the probability that our civilization survives long enough to create them) or the prospects faced by those populations (e.g., by helping to shape the values and institutions that govern the far future).

We conclude then that, if the true axiology extends risk-neutral totalism, it will *either* leave the risk-neutral-totalist case for longtermism unscathed, *or* undermine it by implying widespread incomparability in real-world choice situations. Which of these possibilities is more plausible depends on at least three factors.

1. Our real-world epistemic situation—in particular, which hypotheses about the long-term effects of our actions we think deserve evidential probabilities that are non-zero and non-symmetric (i.e., not cancelled out by equal probabilities of opposite effects, so that they create net differences in the probabilities of particular outcomes between options).

[18] In the second case, it is tempting to conclude that the two lotteries are equally good, but natural ways of generalising this judgement can get us into trouble. For instance, as we will discuss in the next section, the principle of unrestricted anonymity (which says that two outcomes with the same cardinality of locations realising each degree of value are equally good) is incompatible with a weak Pareto principle and so, *a fortiori*, with SoD.

2. The strength of our infinite axiology—for instance, how often it is able to make comparisons between pairs of lotteries where each is better at infinitely many locations, or where each is infinitely better in some states of nature. The former question might depend particularly on whether our axiology is sensitive to the spatiotemporal arrangement of locations, which can help us evaluate tradeoffs between infinite sets of locations.

3. The criteria of identity or counterparthood for locations across different outcomes (which can, for instance, determine whether the choice between two lotteries has the same effect at every location, or creates tradeoffs between locations).[19]

Since all of these factors depend on difficult and unresolved philosophical questions, we unfortunately cannot yet decide with confidence between these two possible conclusions.

Many will judge, however, that *if* the most theoretically plausible extensions of risk-neutral totalism to infinite settings imply widespread incomparability in real-world choice situations, we should not embrace this conclusion but should rather abandon risk-neutral totalism.[20] For this reason, it seems that the most likely way in which the challenges of infinite axiology might undermine the risk-neutral-totalist case for longtermism is not by changing the practical implications of risk-neutral totalism, but by motivating its rejection. So let's next consider that possibility.

6 Giving up risk-neutral totalism

Why might finding ourselves in the infinite setting leads us to not only reject risk-neutral totalism but also its various possible extensions? There are various impossibility results showing that some attractive features of risk-neutral totalism in finite contexts must be given up in infinite contexts. Depending on what principles one takes to be core commitments of risk-neutral totalism, these results might be taken to show that its core commitments are simply

[19]For instance, suppose you face a choice that will influence the probability that some future population containing both humans (with better lives) and non-human animals (with worse lives) will come to exist. Or suppose your choice influences the values that will determine the relative treatment of humans and non-human animals in the far future. If possible locations represent possible persons, and every possible welfare subject is neither necessarily human or necessarily non-human, then these choices will involve tradeoffs between locations. But if every possible location has the same odds of being human and non-human, conditional on its existence, then all locations may face the same local prospect conditional on any option you choose, so that your choice does not require any tradeoffs between locations.

[20]In the literature on infinite aggregation, the conclusion that *no* real-world option is better than any other is typically treated as a *reductio*, to be avoided at all costs. The exception is Smith (2003), who argues for *de facto* moral nihilism on the basis of broadly totalist moral assumption coupled with the infinitude of the future.

inconsistent, or that they are implausible since they are incompatible with other principles that, while not core commitments of risk-neutral totalism, are independently plausible. We will briefly mention four such results.

The first such result is that it cannot be true that both: increasing the value at every location always makes the world better (what is known as the *(Weak) Pareto* principle); and the comparison of outcomes is entirely independent of the identities (and other, qualitative properties) of the locations obtaining each such value (also known as *Unrestricted Anonymity*).[21] But both principles seem to be core commitments of any theory that aims to extend risk-neutral totalism—Pareto is an extreme weakening of SoD, and Unrestricted Anonymity reflects a commitment to impartiality.[22]

The second such result is that the (even Weak) Pareto principle cannot hold for different types of locations (at least for any two types of locations for which their counterpart relations are not essentially dependent on each other). For instance, it cannot be that an outcome is always made better by increasing the value obtained by each *person*, while also that an outcome is always made better by increasing the value obtained at each *position* in spacetime (Cain, 1995; Wilkinson, 2021b, pp. 1925-8). Again, risk-neutral totalism upholds Pareto for all types of locations in the finite context, and each such version of Pareto may seem like a core commitment of totalist theories.

The third deals with lotteries. In the infinite setting, it cannot be true that both: increasing the expectation of value at every location always makes the lottery better (known as the *ex ante (Weak) Pareto* principle); and replacing every outcome in a lottery with a better outcome (in the same state and with the same probability) always makes the lottery better.[23] Again, both principles are upheld by risk-neutral totalism in the finite context and seem like core commitments of such a theory.[24]

The fourth and final result is that it is impossible to give a *complete* and *constructive*[25] ordering $\succcurlyeq$ of infinite outcomes (and, by extension, of lotteries) that satisfies both (Weak) Pareto and even a restricted form of anonymity—namely, *Finite Anonymity*, which says that

[21] The result comes from Liedekerke (1995) originally. See also Hamkins and Montero (2000, p. 237).

[22] Note that it is contested that impartiality requires Unrestricted Anonymity—see Wilkinson (2021b, pp. 1928-31). In the literature, nearly all proposals to extend risk-neutral totalism opt to violate Unrestricted Anonymity to uphold Pareto (for at least some type of locations) and indeed SoD as well (e.g., Vallentyne, 1993; Vallentyne and Kagan, 1997; Jonsson and Voorneveld, 2018; Wilkinson, 2021b).

[23] This result comes from Wilkinson (2022b, §4).

[24] Indeed, stronger versions of both principles feature in the classic theorem of Harsanyi (1955) that is often taken to support risk-neutral utilitarianism.

[25] This means that the ordering must have an explicit description.

permutations of the local values at *finitely many* locations cannot change the value of an outcome.[26] This is a far weaker principle than Unrestricted Anonymity, and still clearly tied to the ideal of impartiality. Putting this result differently, if there is a complete ordering of infinite outcomes that respects Pareto and Finite Anonymity, then it is impossible to give a finite statement of the criterion for one world being better than another, much less to effectively decide for any possible pair of outcomes whether one is better than the other.

A natural response to this last result is to abandon the completeness of $\succcurlyeq$—arguably the least compelling of the various principles in conflict. But this draws our attention to the second way in which infinite axiology still threatens risk-neutral totalism: Any extension of risk-neutral totalism that satisfies certain powerful theoretical desiderata may leave us with *too much* incompleteness in practice. For instance, it has been argued that any infinite axiology must generate widespread incomparability in practice if it satisfies Pareto for persons (*ibid.*, §3.2; Askell, 2019)) or is insensitive to the spatiotemporal arrangement of persons, which is arguably a requirement of impartiality (see Wilkinson, nd, §3.2-3.4). Suppose that many of our choices turn out to have very small effects on the prospects of infinitely many potential value locations, with each option improving the prospects of infinitely many locations while worsening the prospects of infinitely many others, in such a way that our options are incomparable. Even if this is only true of some choices, and therefore does not amount to complete practical paralysis, it might nevertheless be seen as an unacceptable practical implication that forces us to give up risk-neutral totalism for some alternative normative worldview that offers greater practical guidance.

Suppose we conclude that these difficulties of extending risk-neutral totalism to infinite contexts are too great, and that risk-neutral totalism must therefore be given up. Importantly, the challenges of infinite axiology are not unique to risk-neutral totalism, and many ways of abandoning risk-neutral totalism would do little to ease these challenges—for instance, average utilitarian, prioritarian, and egalitarian views face similarly great difficulties. So what alternatives to risk-neutral totalism might we adopt if our main concern is to escape this sort of difficulty altogether? Here are four possibilities.

1. **Pure time discounting**: Value and disvalue arising in the further future contributes less to our overall evaluation of outcomes merely because of its position in time. If our discount schedule is sufficiently severe (e.g., exponential) and value at a time is bounded,

[26] See Zame (2007, Theorem 4) and the more general result given in Lauwers (2010).

this implies that the total discounted value of the future is finite, even if the future contains infinitely many value locations.[27]

2. **Agent-relative consequentialism or strong non-consequentialism**: There is no such thing as the impartial or agent-neutral value of outcomes; or, if there is, it is largely irrelevant to what we should do and plays no essential role in guiding our practical decisions (*cf* Taurek, 1977). Perhaps the value of outcomes is agent-relative, incorporating strong partiality toward the agent and their nearest-and-dearest with little if any weight given to far-off strangers, or depends entirely on the agent's subjective preferences. Or perhaps outcomes don't even have agent-relative value, and which of your options you should prefer in a given choice situation is determined by thoroughly non-consequentialist considerations.

3. **Narrow person-affecting views**: The overall value of an outcome, from the perspective of a particular choice situation, depends only on those locations that exist *necessarily* with respect to that choice situation, i.e., regardless of the agent's choice (see, e.g., Temkin, 1987, pp. 166-7).[28]

4. **Ignoring small probabilities**: Sufficiently low-probability states or outcomes should simply be ignored in ranking lotteries; lotteries should be valued at their expected total value, conditional on such low-probability events (or outcomes) not occurring.[29] Suitably formulated (which is no small challenge—see Kosonen, nd), this policy of small-probability neglect might allow us to ignore both small probabilities of infinite non-random effects and tail risks without finite expectations, which arguably constitute the greatest obstacles to finding an extension of risk-neutral totalism with acceptable real-world implications.

Compared to risk-neutral totalism, on any of these views, the case for longtermism appears weaker. But each of these views has serious drawbacks—in our view, greater than those of

[27]This constitutes a rejection of both Unrestricted and Finite Anonymity, and so clearly abandons risk-neutral totalism's commitment to impartiality. This sort of partiality toward nearer locations has been defended as necessary for the evaluation of infinite futures—see for instance Koopmans (1960). But note that *time* discounting alone does not avoid the problems associated with a *spatially* infinite universe; so to avoid all of the difficulties of the infinite setting, one might need a spatial as well as a temporal discount rate. For a survey of arguments against pure time discounting, see Greaves (2017a, §7).

[28]This is a species of agent-relative consequentialism, but an especially notable one for present purposes. A very similar view could be articuled in *time-relative* rather than *agent-relative* fashion: the goodness of outcomes is time-relative, depending only on those locations that necessarily exist as of that time. Such time-relative views violate both Unrestricted and Finite Anonymity, as well as providing deeply counterintuitive implications (see Greaves, 2017b, pp. 8-9).

[29]This view is dubbed *Nicolausian discounting* by Monton (2019), who defends it. For objections, see for instance Wilkinson (2022a), and Beckstead and Thomas (nd).

the various proposed extensions of risk-neutral totalism in the infinite setting. But no doubt some will disagree, and it is undeniable that these challenges do count somewhat in favour of normative worldviews less favourable to longtermism.

7 Conclusion

We set out to investigate whether the axiological challenges of infinite worlds undermine the risk-neutral-totalist case for longtermism. The results of this investigation are, unfortunately, mixed and uncertain.

Our own provisional conclusions are as follows. First, any plausible extension of risk-neutral totalism to infinite contexts can rank lotteries in any decision where our choices affect only finitely many local prospects. In such decisions, any such view preserves the risk-neutral-totalist case for longtermism by letting us ignore all those locations whose prospects are unaffected. And many of our real-world decisions have this nice character since, even those physical hypotheses that put infinitely many value locations in our causal future mostly suggest that our choice affect only finitely many local prospects.

Second, we should assign some non-zero probability to physical hypotheses that let us predictably affect infinitely many locations. This means that our choices—at least those that affect the long-run future—do have at least some small effect on infinitely many local prospects.

Third, in those circumstances, any otherwise plausible extension of risk-neutral totalism that makes comparisons (rather than implying widespread incomparability) will very likely preserve the risk-neutral-totalist case for longtermism. Indeed, it seems that it would even strengthen that case by implying that the long-term stakes of our actions are infinite.

Fourth, if otherwise plausible extensions of risk-neutral totalism instead imply widespread incomparability in practice, then we plausibly have good reason to reject risk-neutral totalism. And various impossibility results in infinite axiology might also be taken to motivate the rejection of risk-neutral totalism, since they imply that at least some of its attractive features in finite contexts must be given up in infinite contexts.

We ourselves are inclined to think that risk-neutral totalism remains more plausible than each of the alternatives raised above, despite the impossibility results.[30] And we hold out hope

[30]We both incline at least somewhat towards totalism. One of us (HW) also inclines toward risk neutrality,

that the correct extension of risk-neutral totalism to infinite contexts, while it may countenance some incomparability between outcomes and lotteries, will not imply very widespread incomparability in real-world choice situations. But this hope has not yet been fully vindicated—it is not yet clear what the correct extension is. (Nor has it been vindicated, nor the correct extension identified, for the many axiologies other than risk-neutral totalism that are also favourable to longtermism.) Until that correct extension is found, while infinitary worries about the case for longtermism can be mitigated, they cannot be totally allayed.[31]

References

Arntzenius, F. (2014). Utilitarianism, decision theory and eternity. *Philosophical Perspectives 28*(1), 31–58.

Askell, A. (2019). *Pareto Principles in Infinite Ethics*. Ph. D. thesis, New York University.

Bader, R. (n.d.). *Person-Affecting Population Ethics*. unpublished manuscript.

Basu, K. and T. Mitra (2007). Utilitarianism for infinite utility streams: A new welfare criterion and its axiomatic characterization. *Journal of Economic Theory 133*(1), pp. 350–73.

Beckstead, N. and T. Thomas (n.d.). A paradox for tiny probabilities and enormous values. unpublished manuscript.

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics 10*, 9–59.

Buchak, L. (2022). How should risk and ambiguity affect our charitable giving? Technical report, GPI Working Paper No. 8-2022.

Cain, J. (1995). Infinite utility. *Australasian Journal of Philosophy*, pp. 401–4.

Carroll, S. M. (2020). Why Boltzmann brains are bad. In S. Dasgupta, R. Dotan, and B. Weslake (Eds.), *Current Controversies in Philosophy of Science*. Taylor & Francis.

while the other (CT) does not, but thinks that the correct principles for evaluation of risky prospects will have similar implications in practice.

[31] For interested readers, we have two suggestions for future research. First, compared to the extensive literature on the evaluation of infinite outcomes, there has been relatively little work in infinite-world axiology on the evaluation of lotteries. More such work, exploring possible strengthenings of principles like SVPD and their practical implications, could be very useful. Second, most views in infinite-world axiology make use of an identity or counterpart relation across possible outcomes, and the practical implications of these views depend on the nature of that relation. But most work in this area does not incorporate a full theory of the relevant relation or think through what it implies about our real-world circumstances. This sort of work also seems essential to fully understanding the practical implications of an infinite axiology (see note 19 above).

Clark, M. (n.d.). Infinite ethics, intrinsic value, and the Pareto principle. Unpublished manuscript, July 2019.

Earman, J. and J. D. Norton (1993). Forever is a day: Supertasks in Pitowsky and Malament-Hogarth spacetimes. *Philosophy of Science 60*(1), 22–42.

Greaves, H. (2017a). Discounting for public policy: A survey. *Economics & Philosophy 33*(3), 391–439.

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

Greaves, H. and W. MacAskill (2021). The case for strong longtermism. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 5-2021.

Hamkins, J. D. and B. Montero (2000). Utilitarianism in infinite worlds. *Utilitas 12*(01), 91–.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy 63*(4), 309–21.

Jonsson, A. and M. Voorneveld (2018). The limit of discounted utilitarianism. *Theoretical Economics 13*(1), 19–37.

Knobe, J., K. D. Olum, and A. Vilenkin (2006). Philosophical implications of inflationary cosmology. *The British Journal for the Philosophy of Science 57*(1), 47–67.

Koopmans, T. C. (1960). Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society 28*(2), 287–309.

Kosonen, P. (n.d.). Tiny probabilities and the value of the far future. unpublished manuscript.

Lauwers, L. (2010). Ordering infinite utility streams comes at the cost of a non-Ramsey set. *Journal of Mathematical Economics 46*(1), 32–37.

Lauwers, L. and P. Vallentyne (2004). Infinite utilitarianism: More is always better. *Economics and Philosophy 20*(2), 307–330.

Liedekerke, L. V. (1995). Should utilitarians be cautious about an infinite future? *Australasian Journal of Philosophy 73*(3), 405–7.

Liedekerke, L. V. and L. Lauwers (1997). Sacrificing the patrol: Utilitarianism, future generations and infinity. *Economics and Philosophy 13*(2), 159–174.

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

MacAskill, W., A. Vallinder, C. Shulman, C. Österheld, and J. Treutlein (2021). The evidentialistâs wager. *Journal of Philosophy 118*(6), pp. 320–42.

Meacham, C. (2020). Too much of a good thing: Decision-making in cases with infinitely many utility contributions. *Synthese*, pp. 1–41.

Monton, B. (2019). How to avoid maximizing expected utility. *Philosophers 19*.

Pettigrew, R. (2022). Effective altruism, risk, and human extinction. Technical report, GPI Working Paper No. 2-2022.

Smith, Q. (2003). Moral realism and infinite spacetime imply moral nihilism. In *Time and Ethics: Essays at the Intersection*, pp. 43–54. Springer.

Smolin, L. (1992). Did the universe evolve? *Classical and Quantum Gravity 9*(1), 173.

Tarsney, C. and T. Thomas (2020). Non-additive axiologies in large worlds. *arXiv preprint arXiv:2010.06842*.

Taurek, J. M. (1977). Should the numbers count? *Philosophy & Public Affairs*, pp. 293–316.

Temkin, L. S. (1987). Intransitivity and the mere addition paradox. *Philosophy & Public Affairs*, 138–187.

Thomas, T. (2019). The asymmetry, uncertainty, and the long term. Technical report, GPI Working Paper No. 11-2019.

Tipler, F. J. (1994). *The Physics of Immortality: Modern Cosmology, God, and the Resurrection of the Dead*. New York: Anchor Books.

Vallentyne, P. (1993). Utilitarianism and infinite utility. *Australasian Journal of Philosophy 71*(2), 212–217.

Vallentyne, P. and S. Kagan (1997). Infinite value and finitely additive value theory. *The Journal of Philosophy 94*(1), 5–26.

Wilkinson, H. (2021a). *Infinite Aggregation*. Ph. D. thesis, Australian National University.

Wilkinson, H. (2021b). Infinite aggregation: Expanded addition. *Philosophical Studies 178*(6), pp. 1917–49.

Wilkinson, H. (2022a). In defence of fanaticism. *Ethics 132*, pp. 445–77.

Wilkinson, H. (2022b). Infinite aggregation and risk. *Australasian Journal of Philosophy*.

Wilkinson, H. (n.d.). Chaos, add infinitum. unpublished manuscript.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

Zame, W. R. (2007). Can intergenerational equity be operationalized? *Theoretical Economics 2*(2), 187–202.