Longtermism in an Infinite World

Christian J. Tarsney (Population Wellbeing Initiative, University of Texas at Austin) and Hayden Wilkinson (Global Priorities Institute, University of Oxford)

Global Priorities Institute | March 2023

GPI Working Paper No . 4-2023

Longtermism in an Infinite World[*]

Christian J. Tarsney[†] & Hayden Wilkinson[‡]

Last updated: January, 2023

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 future contains *infinite* value, then many theories of value that support longtermism (e.g., risk-neutral total utilitarianism) seem to imply that no available action is better than any other. And some strategies for avoiding this conclusion (e.g., exponential time discounting) yield views that are much less supportive of longtermism. This chapter explores 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 still support standard arguments for longtermism, since they imply that when we can only affect (or only *predictably* affect) a finite part of an infinite universe, we can reason as if only that finite part existed. On the other hand, (iii) there are improbable but not impossible physical scenarios in which our actions can have *infinite predictable effects* on the far future, and these scenarios create substantial unresolved problems for both infinite ethics and the case for longtermism.

Keywords: longtermism, infinite axiology, infinite ethics, infinite aggregation, infinitarian paralysis

[*] 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.

[†] Population Wellbeing Initiative, University of Texas at Austin.

[‡] Global Priorities Institute, University of Oxford.

1 Introduction

Longtermism is, very roughly, the thesis that the moral value of actions available to present day agents is primarily determined by their potential effects 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 or billions of years, and could spread across a large portion of the accessible universe, resulting in an enormous number of future people. If we can affect 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) for longtermism rests on an axiology—a theory of moral value for outcomes and risky prospects—that is *additive*, *impartial*, and *risk-neutral.*[2] Roughly, *additivity* means that the value of an outcome is a weighted sum of values realized at particular "value locations" in that outcome (e.g., welfare realized in the lives of particular persons), 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 (in particular, persons[3]), how things go in the far future potentially makes an enormous difference to the overall value of outcomes. *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

[1] Note that this is a claim about what actions *would be best* (an *axiological* claim), not about what agents *ought* to do (a *deontic* claim). It is possible that we sometimes ought not perform the best available action—for instance, if you can prevent five murders by committing one, a deontologist might concede that *it would be better* if you committed murder (since it's better for one murder to occur rather than five), while still maintaining that you *ought not* commit murder. Thus, in focusing on longtermism as an axiological thesis, we are setting aside the question of whether what one *ought to do* is primarily determined by possible effects on the far future.

[2] Arguments for longtermism can also be made in the context of axiologies that are, for instance, *averageist*, *egalitarian*, *person-affecting*, and/or *risk-sensitive*, though with additional complications. For relevant discussion, see Tarsney and Thomas (2020), Thomas (forthcoming), Buchak (2022), Pettigrew (2022), Greaves and MacAskill (2021), and Wilkinson (n.d.b), among others. Much of what we say in this chapter generalizes to arguments for longtermism based on these alternative axiologies, but we focus on risk-neutral totalism for simplicity.

[3] We use "person" simply as shorthand for "morally considerable welfare subject".

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 value of our actions, since the stakes are so high. We will refer to the conjunction of these three principles as *risk-neutral totalism*.[4] (For a more precise definition, see §3.)

It is also possible, however, that the future is not merely vast but *infinite*, containing infinitely many value locations and infinite total value and/or disvalue.[5] 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 if our actions are guaranteed to 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 seem to imply, implausibly, that none of our actions matter. Or, if we stand by these principles and bite the bullet on their apparent nihilistic implication, it seems 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.[6]

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

[4] 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.

[5] This is implied by the influential (though still disputed) inflationary paradigm in cosmology (see Knobe et al. (2006): 50-51). 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): 11-16). By either view, for any local physical phenomenon, the universe will contain infinitely many near-perfect duplicates, with probability 1.

[6] It is important to note that the challenges of infinite ethics are not unique to risk-neutral totalism—very similar challenges arise for most axiologies, including the sorts of axiologies mentioned in fn. 2 above.

theories of the value of outcomes from finite to infinite contexts—that is, 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 that can be satisfied in finite contexts.

The challenges of infinite axiology thus threaten the case for longtermism in two ways. First, they might lead us to simply give up on risk-neutral totalism, in favour of moral views that are less favourable to longtermism. 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, while extant proposals for extending risk-neutral totalism to infinite contexts all face costs, those costs are not severe enough to scuttle the project entirely. Second, as we will show, most such proposals 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.

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 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.[7] 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-

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

preserving (i.e., greater numbers correspond to greater degrees of value) and unique at least up to positive affine transformation (meaning that the numbers carry meaningful information about the relative size of *differences* in value).[8] 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 $V_i: \mathcal{L} \rightarrow \mathcal{V} \cup \{\Omega\}$ 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. The *total* value of an outcome is the sum of local value at all locations that exist in that outcome (formally, $\sum_{l \in \mathcal{L}: O_i(l) \neq 0} O_i(l)$), which we will abbreviate $Tot(O_i)$. This sum can, of course, be infinite or undefined.

We also wish to compare *prospects* (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 prospects over outcomes is denoted by $\mathcal{P}$. For any prospect P_i, its probability of resulting in an outcome in set $\mathcal{O}'$ is given by $P_i(\mathcal{O}')$. We will abbreviate $P_i(\{O\})$ to $P_i(O)$ to denote the probability of a single outcome O. When a prospect results in some outcome O with probability 1, we allow O to denote the prospect as well as the outcome.

An *axiology* is an evaluative ranking of both outcomes and prospects. We assume that these two rankings must be consistent in the sense that one outcome is at least as good as another just in case a prospect yielding the first outcome with probability 1 is at least as good as a prospect 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 prospects. 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).

[8] The latter assumption follows from additivity, which lets us make sense of the relative size of value differences. For instance, we can say that the difference in value realised at l_1 and l_2 is at least as great as that between l_3 and l_4 if and only if substituting l_1 for l_2 and l_4 for l_3 in an outcome will always yield an outcome that is at least as good as the original. Additivity guarantees that the size ordering of value differences, defined in this way, will be complete.

3 Two consensus principles

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

> *Risk-Neutral Totalism*: For any outcomes O_i and O_j whose total values are finite, $O_i \succcurlyeq O_j$ if and only if its total value is greater.[9] Likewise, for any prospects P_i and P_j whose *expectations* of total value are finite, $P_i \succcurlyeq P_j$ if and only if its expectation is greater.[10]

We are looking for principles, then, that extend risk-neutral totalism in the sense of implying these biconditionals, while also implying at least some further comparisons 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_i and O_j whatsoever, $O_i \succcurlyeq O_j$ if and only if its total value is greater. Likewise, for any prospects P_i and P_j whatsoever, $P_i \succcurlyeq P_j$ if and only if its expected total value is greater.

This view is naive because it generalizes risk-neutral totalism from finite to infinite contexts in a way that is simple and straightforward, but which a little reflection reveals to be implausible. Suppose, for instance, that in outcomes O_a and O_b exactly the same infinite set of locations exists, in the same spatiotemporal arrangement, but that each location has value 1 in O_a and 2 in O_b. Nearly everyone would agree that O_b is better than O_a, but naive risk-neutral totalism implies that they are equally good. More troublingly

[9] Formally: If $Tot(O_i)$ and $Tot(O_j)$ are both finite, then $O_i \succcurlyeq O_j \Leftrightarrow Tot(O_i) \geq Tot(O_j)$.

[10] Formally: If $\mathbb{E}(Tot(P_i))$ and $\mathbb{E}(Tot(P_j))$ are both finite, then $P_i \succcurlyeq P_j \Leftrightarrow \mathbb{E}(Tot(P_i)) \geq \mathbb{E}\left(Tot(P_j)\right)$, where $\mathbb{E}(Tot(P) = \sum_{O \in \mathcal{O}} Tot(O)P(O)$.

for practical purposes, naive risk-neutral totalism implies that if there is already infinite value and/or disvalue in the world, then no finite change (e.g., saving a life) can ever make things better or worse overall. Since the effects of the actions actually available to us appear to be finite, this strongly suggests that it doesn't matter what we do (at least from an axiological point of view), even in cases where it clearly *does* matter (e.g., when we have the opportunity to save a life).[11]

In search of a more plausible view, there have been many alternative proposals for extending the totalist ranking of outcomes and/or the risk-neutral totalist ranking of prospects to infinite contexts (e.g., Vallentyne 1993; Vallentyne and Kagan 1997; Liedekerke and Lauwers 1997; Bostrom 2011; Arntzenius 2014; Jonsson and Voorneveld 2018; Wilkinson 2021b; Clark n.d.). But, rather than describe these (often rather intricate) proposals 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_i and O_j, a sufficient condition for $O_i \succ O_j$ is that
>
> $$\sum_{l \in \mathcal{L}} (V_i(l) - V_j(l)) > 0$$
>
> 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

[11] Along with the challenge of evaluating outcomes with infinite or undefined value and prospects over those outcomes, risk-neutral totalism (and many other views) also face difficulties evaluating prospects that have infinite or undefined value, even though all their possible outcomes have finite value. Such prospects include, for instance, the St. Petersburg game (Bernoulli 1738) and the Pasadena game (Nover and Hájek 2004). There are notable parallels between these two challenges, in theory (both challenges involve trying to rank divergent sums) and in practice (both threaten to create widespread incomparability between our options, particular in situations where our choices might affect the very far future). But in this chapter, to keep things manageable, we focus exclusively on the first challenge (of *outcomes* with infinite or undefined value).

as equivalent to existence with value 0). Likewise, if this sum is equal to 0, then $O_i \sim O_j$.[12]

To illustrate, consider the following pair of outcomes. (In this array, columns represent possible locations and rows represent possible outcomes. Each number in the array gives the local value at a particular location in a particular outcome.)

$$\begin{array}{r|cccccccc} & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & \cdots \\ O_1: & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \cdots \\ O_2: & 1 & 0 & 0 & 1 & 1 & 1 & 1 & \cdots \\ V_1 - V_2: & -1 & 1 & 1 & 0 & 0 & 0 & 0 & \cdots \end{array}$$

Naive risk-neutral totalism says that neither of these outcomes is better than the other, since the sum of values in each outcome is infinite. But Sum of Differences lets us compare O_1 and O_2 by summing the numbers in the bottom row, as long as that sum is well-defined. In this case, since the sum is positive, we can conclude that O_1 is strictly better than O_2. 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 locations where they differ*. Thus, from the fact that the subtotal value of O_1 from l_1 to l_3 is 2, and the corresponding subtotal for O_2 is 1, we can conclude that O_1 is strictly better than O_2.

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).[13]

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

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

But Sum of Differences says nothing about how to compare prospects. For that purpose, we can extend it to a principle we will call *Sum of Value-Probability Differences* (SVPD). To state this principle, let $P(V(l) = v)$ denote prospect *P*'s probability of yielding an outcome with value v at location l.[14]

> *Sum of Value-Probability Differences (SVPD)*: For any prospects P_i and P_j, $P_i \succ P_j$ if
>
> $$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} v \times (P_i\,(V(l) = v) - P_j(V(l) = v)) > 0$$
>
> 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 $P_i{\sim}P_j$.

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 P_i is chosen vs. if P_j 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 prospects. In cases of only finitely many value locations and finite expected local value at each, SVPD agrees with risk-neutral totalism. And it can many infinitary prospects too, as we illustrate below. But, importantly, SVPD does not always yield a comparison—it only does so if the sum in the definition converges (or diverges to $+/-\infty$) 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 prospects as it does for comparing outcomes. But every such proposal, if combined with SoD, implies SVPD.[15] Like SoD, then, SVPD is a relatively weak principle that should be

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

[15] In fact, if combined with SoD, every such proposal strengthens SVPD by constraining the order of summation. The proposals of Arntzenius (2014: 55-56), Bostrom (2011: 27-30), and Meacham (2020) strengthen it to satisfy what could be called *Sum of Differences in Expectations*: that two prospects can be compared by first summing the relevant terms over all values of v *at* each location l, and only then summing over all locations l. Effectively, they perform the Sum of Differences over expectations at each location. Meanwhile, Wilkinson (2022b: 14) strengthens SVPD to satisfy what could be called *Expected Sum*

mostly uncontroversial insofar as our goal is to extend risk-neutral totalism to infinite contexts.

To illustrate SVPD, consider the following pair of prospects. (As before, columns represent possible locations and rows represent possible outcomes. The probability of a particular outcome under a particular prospect is given in the first column.)

$$P_1 \left\{ \begin{array}{ccccccccccccc} P_1(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 2 & 2 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & \dots \\ 0.5 & O_2: & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \end{array} \right.$$

$$P_2 \left\{ \begin{array}{ccccccccccccc} P_2(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_3: & 2 & 0 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & \dots \\ 0.5 & O_4: & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \end{array} \right.$$

Here, P_1 and P_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 prospect. For instance, perhaps the value realised at these locations depends on a fair coin flip, and choosing P_1 will cause these locations to have value 1 iff the coin lands heads, while choosing P_2 will cause them to have value 1 iff the coin lands tails. But the choice of prospect does not affect the *probability distribution* over local outcomes for any of these locations. So, intuitively, we would like to ignore them. Fortunately, SVPD lets us do this: for all $i > 3$ and all v, $P_a(V(l_i) = v) - P_b(V(l_i) = v) = 0$. This means that, in applying SVPD, we can simply ignore all these locations, and compare P_1 with P_2 by comparing only their differing expectations of local value at locations l_1 to l_3, as long as these are finite. Here, P_1 has an expected (subtotal) value of 3 across these locations, while P_1 has expected value of 2, so SVPD tells us that $P_1 \succ P_2$. And, in general, when two prospects yield the same local prospects at all but finitely many locations, SVPD allows us to

of Differences: that two prospects can be compared by first summing those differences for the pairs (v, l) corresponding to each outcome, and only then summing over all outcomes. Effectively, the latter method takes the expectation of the Sum of Differences. And 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: 8). Because SVPD requires unconditional convergence, it is neutral in these hard cases, and compatible with either *ex ante* Pareto or statewise dominance.

compare those prospects by comparing their expected subtotals at that finite set of locations.

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 satisfactorily. Specifically, we have seen that these principles let us compare pairs of outcomes (resp., prospects) 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.

Whether we can indeed do so depends on what prospects we actually face, and so 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 epistemic probabilities, and yet another distribution of subjective credences. We will focus on epistemic probabilities, i.e., the probabilities that it is epistemically rational for an agent to assign to particular outcomes on the supposition that she takes a given action. (But much of what we say will carry over to subjective credences.)

So, is our real-world situation such that the pairs of options we are required to evaluate beget different epistemic probability distributions over local value at only finitely many locations? 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 (e.g., Smolin's 'cosmological natural selection' model; see Smolin 1992) imply that events in one 'universe' affect events in other universes, which would suggest 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: 10). And the manner and timing of those fluctuations may be affected (albeit likely not in any 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 represents 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.[16] 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 allow us to simply apply naive risk-neutral totalism to the part of the world we predictably affect: While our present choices may determine

[16] 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.

which Boltzmann brains come to exist and what experiences they have, the epistemic probability of any particular event after the heat death of the universe (e.g., a particular Boltzmann brain existing at a particular spaccetime position) 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.[17]

On the other hand, hypotheses on which our descendants may be able to 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 descendants 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.

Thus, it may be that the really difficult problems of infinite ethics arise in precisely those cases to which the longtermist thesis is supposed to apply—namely, choices that affect the epistemic probabilities of humanity's long-term survival or other important long-term outcomes (e.g., particular values prevailing in the far future). Arguably, many of our choices are not like this—for instance, your decision what to eat for breakfast may make no difference at all to the epistemic probability of humanity's long-term survival. In that case, minimal principles like SoD and SVPD can straightforwardly protect us from the paralyzing effects of infinitary ethical considerations in these ordinary cases. But in the more consequential situations where our choices do have some predictable effect on the long-term future, they plausibly also make a non-zero—though perhaps *extremely* small—difference to the epistemic prospects of infinitely many potential value locations.

[17] Similar things can be said of certain multiverse hypotheses where we can affect other "universes" (for instance, through gravitational interactions in a higher-dimensional space) but are not in a position to know anything about the empirical details those effects.

So we must ask whether SVPD, or plausible extensions thereof, can handle these situations.[18]

5 Infinite predictable 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 what risk-neutral totalism would recommend if the universe were merely finite. 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

[18] A different way in which our choices might affect the prospects of infinitely many locations is if the correct decision theory is non-causal (e.g., evidential). We have so far implicitly assumed a causal decision theory on which our choices can in principle 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). If the universe is spatially infinite and contains infinitely many situations identical or arbitrarily similar to ours, then it is guaranteed that the number of such locations will be infinite as well. Indeed, even non-zero credence in non-causal decision theory might have this upshot, if we treat our uncertainty between causal and non-causal 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).

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 infinite axiology will allow us to apply naive risk-neutral totalism as described above. 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 prospects 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 effects 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 far 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, in situations where our choices affect the long-term future, we face widespread incomparability, with no available option being better or worse than any other.[19]

[19] 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.

To illustrate these points, let's start with the easy cases of infinite predictable effects, and work our way toward the harder cases.[20] 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 you can increase the probability that they come to exist without changing their prospects conditional on existence.

$$P_3 \left\{ \begin{array}{ccccccccccccc} P_3(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \ldots \\ 0.5 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \ldots \\ 0.5 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \ldots \end{array} \right.$$

$$P_4 \left\{ \begin{array}{ccccccccccccc} P_4(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \ldots \\ 0.51 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \ldots \\ 0.49 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \ldots \end{array} \right.$$

In this case, the sum from the definition of SVPD—of $v \times (P_4(V(l) = v) - P_3(V(l) = v))$ 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 $P_4 \succ P_3$.

There are other cases in which it seems clear which of two prospects is better, that are not ranked by SVPD, but that can be handled by natural and plausible strengthenings of SVPD. For instance, consider the following case, where you can improve the prospects of infinitely many locations conditional on existence, without affecting their probabilities of existence.

$$P_5 \left\{ \begin{array}{ccccccccccccc} P_5(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \ldots \\ 0.05 & O_2: & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \ldots \\ 0.05 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \ldots \\ 0.9 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \ldots \end{array} \right.$$

[20] We note in passing that effects on the outcomes or prospects of infinitely many locations can have finite sums—for instance, if we improve l_1 by $\frac{1}{2}$, l_2 by $\frac{1}{4}$, l_3 by $\frac{1}{8}$, and so on. These ultra-easy cases are handled adequately by SoD and SPVD, and in some cases by (unextended) risk-neutral totalism. But if our choices do predictably affect infinitely many locations, the effects are unlikely to be this well-behaved. We focus, therefore, on situations involving infinite sums.

$$P_6\left\{\begin{array}{cccccccccccc} P_6(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.051 & O_2: & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & \dots \\ 0.049 & O_1: & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & \dots \\ 0.9 & O_\Omega: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{array}\right.$$

Clearly P_6 is better than P_5. But SVPD is silent: Because P_6 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 prospects P_i andP_j, $P_i \succ P_j$ if there exists an outcome $O_b \in \mathcal{O}$ such that
>
> $$\sum_{(v,l)\in\mathcal{V}\times\mathcal{L}} (v - V_b(l)(P_i(V(l) = v) - P_j(V(l) = v)) > 0$$
>
> 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 $P_i \sim P_j$.

If we choose the outcome O_1 above (in which every location has value 1) as our baseline outcome O_b, and substitute P_6 and P_5 for P_i and P_j respectively, we find that the above sum diverges unconditionally to $+\infty$. So we can conclude that $P_6 \succ P_5$. And this baseline-adjusted principle, while slightly more complicated than SVPD, is similarly modest and uncontroversial.[21]

In both these cases, the principles we have appealed to imply that one prospect is "infinitely better" than another in the sense that no finite improvement of the worse prospect (or worsening of the better prospect) could affect the comparison. (For instance, if we add any finite number of locations that will realise value 1 for sure under P_3 or P_5, and 0 for sure under P_4 or P_6, the ranking would be unchanged.) Correctly evaluating this sort of infinite improvement requires some extension of naive risk-neutral totalism.[22]

[21] In particular, like SVPD, it follows from the proposals in Wilkinson (2022b: 14), Arntzenius (2014: 55-56), Bostrom (2011: 27-30), and Meacham (2020), given Sum of Differences.

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

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 prospects (see Beckstead and Thomas, n.d.; Wilkinson, 2022a). And second, if there is any epistemic 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.[23]

More generally, it seems to us that the possibility that we face choices between prospects that yield different local prospects at infinitely many locations does not threaten the case for longtermism *as long as these prospects 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 nearfuture 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 prospects, and where incomparability is plausible. Here are three examples. First, suppose your choice affects the probability that some infinite future population will come

[23] 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.

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).

$$P_7\left\{\begin{array}{ccccccccccccc} P_7(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{array}\right.$$

$$P_8\left\{\begin{array}{ccccccccccccc} P_8(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.51 & O_1: & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & \dots \\ 0.49 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{array}\right.$$

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

$$P_9\left\{\begin{array}{ccccccccccccc} P_9(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & 1 & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{array}\right.$$

$$P_{10}\left\{\begin{array}{ccccccccccccc} P_{10}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & \Omega & 1 & \Omega & -1 & \Omega & 1 & \Omega & -1 & \Omega & \dots \\ 0.5 & O_2: & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \Omega & \dots \end{array}\right.$$

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 probability that future societies are inequality-averse, and thereby the probability of more versus less unequal distributions of welfare being realised in those societies.

$$P_{11}\left\{\begin{array}{ccccccccccccc} P_{11}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \dots \\ 0.5 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \dots \\ 0.5 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \dots \end{array}\right.$$

$$P_{12}\left\{\begin{array}{lccccccccccc} P_{12}(O_i) & & l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & \ldots \\ 0.51 & O_1: & 2 & 6 & 2 & 6 & 2 & 6 & 2 & 6 & 2 & \ldots \\ 0.49 & O_2: & 3 & 4 & 3 & 4 & 3 & 4 & 3 & 4 & 3 & \ldots \end{array}\right.$$

None of these three cases are resolved by SVPD (or by the stronger, baseline-adjusted version discussed above). Nor is there an intuitively clear right answer in any of these cases.[24] 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 epistemic probabilities that are non-zero and nonsymmetric (i.e., not cancelled out by equal probabilities of opposite effects, so that they create net differences between possible actions in the probabilities of particular long-term outcomes).

[24] In the second case, it is especially tempting to conclude that the two prospects 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 prospects 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 prospects has the same effect at every location, or creates tradeoffs between locations).[25]

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.[26] 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.

[25] For instance, suppose you face a choice that affects the probability that some future population containing both humans (with good lives) and non-human animals (with bad lives) will come to exist. Or suppose your choice influences the relative treatment of humans and non-human animals in the far future. If each possible location is either necessarily human or necessarily non-human, then these choices will involve tradeoffs between locations—one option being better in expectation for some locations and worse in expectation for others. But if the identity or counterpart relation for locations is indifferent to species (as is plausible, for instance, if locations are spacetime positions rather than persons), then your choice might have the same or very similar effects on the local prospects of each location it affects, and so involve no tradeoffs between locations. All else being equal, it will be easier to rank infinite prospects that do not involve tradeoffs between two infinite sets of locations.

[26] 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 assumptions coupled with the infinitude of the future.

6 Giving up risk-neutral totalism

Why might infinite ethics lead us to reject risk-neutral totalism (and its various possible extensions)? One important reason is the existence of 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 inconsistent, or that they are implausible since they conflict with other principles that, while not core commitments of risk-neutral totalism, are independently plausible. We will briefly mention four such results.

The first is that even weak versions of the Pareto principle are incompatible with an unrestricted anonymity principle. *Weak Pareto* says that, if two worlds have the same locations but each location has greater value at one of the world, then that world is better. *Unrestricted Anonymity* says that if two worlds have the same locations, and have the same number (i.e., cardinality) of locations at each value level, then they are equally good—in other words, it doesn't matter *which* locations realise which degrees of value.[27] But both these 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 intuitively reflects the totalist commitment to impartiality.[28]

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: 1925-1928). 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.

[27] The result comes from Liedekerke (1995) originally. See also Hamkins and Montero (2000: 237).

[28] Note that it is contested that impartiality requires Unrestricted Anonymity—see Wilkinson (2021b: 1928-1931). 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).

The third result concerns prospects. In the infinite setting, it cannot be true that both: increasing the expectation of value at every location always makes the prospect better (known as *ex ante (Weak) Pareto*); and replacing every outcome in a prospect with a better outcome (in the same state and with the same probability) always makes the prospect better.[29] Again, both principles are upheld by risk-neutral totalism in the finite context and seem like core commitments of such a theory.[30]

The fourth and final result is that any ordering of infinite outcomes that satisfies both Weak Pareto and Finite Anonymity must be either incomplete or non-constructive.[31] Finite Anonymity says that permuting the local values at *finitely many* locations does not change the value of an outcome. This is much weaker than Unrestricted Anonymity, is consistent with Weak Pareto, and seems to straightforwardly reflect the ideal of impartiality. An ordering $\succcurlyeq$ is *non-constructive* if it does not have an explicit, finite description—that is, we cannot write down a true formula of the form "For all O_i, O_j, $O_i \succ O_j$ if and only if φ", where the right-hand side of the formula does not contain $\succcurlyeq$ or anything defined in terms of it. While the viability of non-constructive axiological or normative principles has not been substantially explored (and we would find such exploration very valuable), it seems to us that any non-constructive axiology would be problematically arbitrary. Its specification would require infinitely many independent "choices" to rank one outcome over another, without any unifying principle to explain those choices—there would be an infinity of brute axiological facts.[32]

This fourth result, therefore, leaves us with a fairly strong argument for incompleteness. But this draws our attention to the second way in which infinities still threaten risk-neutral totalism: An infinite axiology that extends risk-neutral totalism, even if it satisfies all the theoretical desiderata we deem essential, may yield an

[29] This result comes from Wilkinson (2022b: §4).

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

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

[32] For instance, one way of getting a complete axiology that satisfies both Weak Pareto and Finite Anonymity is to invoke an *ultrafilter*, a particular kind of non-constructive object that in the present context would tell us which infinite subsets of the set of possible locations should be treated as "large" and which as "small". (The resulting principle will then always prefer, for instance, to provide a given benefit to a "large" set of locations rather than a "small" set.) But it is very hard to imagine what could single out any particular ultrafilter, among the uncountably many that can be imposed on the infinite set of possible locations, to play this privileged axiological role. What could ground or explain its special status?

unacceptable amount of incompleteness in practice. Apart from the axiomatic argument for incompleteness just described, and the cases in the last section where risk-neutral totalist commitments do not suggest any obvious ranking of alternative prospects, there are also general arguments for expecting fairly widespread incomparability in infinite extensions of risk-neutral totalism. For instance, it has been argued that any infinite axiology must generate widespread incomparability in practice if it satisfies Pareto for persons (Wilkinson, 2021b, §3.1; Askell 2019)) or is insensitive to the spatiotemporal arrangement of persons, which is arguably a requirement of impartiality (see Wilkinson, n.d.a, §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 by risk-neutral totalist lights. Even if this is only true of some choices, and therefore does not leave us completely adrift in deciding what to do, it might nevertheless be seen as an unacceptable failure of the risk-neutral totalist worldview to offer 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, this implies that the total discounted value of the future is finite, even if the future contains infinitely many value locations.[33]

[33] 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

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 depending 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-axiological considerations.

3. **Narrow person-affecting views**: The overall value of an outcome, from the perspective of an agent in 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: 166-167).[34]

4. **Ignoring small probabilities**: Sufficiently low-probability states or outcomes should simply be ignored in ranking prospects; prospects should be valued at their expected total value, conditional on such low-probability events (or outcomes) not occurring.[35] Suitably formulated (which is no small challenge—see Kosonen n.d.), this policy of small probability neglect might allow us to ignore the speculative hypotheses (like cosmological natural selection and future supertask computers) that allow our actions to predictably affect infinitely many locations, and thereby rescue us from widespread incomparability in real-world decision situations.

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

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).

[34] This is a species of agent-relative consequentialism, but an especially notable one for present purposes. A very similar view could be articulated in *time-relative* rather than *agent-relative* fashion: the overall value of an outcome, from the perspective of a particular moment in time, depends only on those locations that exist at that time, or on those locations whose existence is nomologically necessary given the state of the universe at that time. Such time-relative views violate both Unrestricted and Finite Anonymity, and have some deeply counterintuitive implications (see Greaves 2017b: 8-9).

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

infinite setting. But no doubt some will disagree, and it is undeniable that the challenges of infinite axiology 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 prospects 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—particularly those involving no predictable long-term effects—will plausibly have this nice character.

Second, we should assign some non-zero probability to physical hypotheses that let us predictably affect infinitely many locations in certain decisions. This means that our choices—at least those that affect the long-run future—can 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.[36] And we hold out hope that the correct extension of risk-neutral totalism to infinite contexts, while it may countenance some incomparability between outcomes and prospects, 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.[37]

References

Arntzenius, F. (2014), 'Utilitarianism, decision theory and eternity', in *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 Mitra, T. (2007), 'Utilitarianism for infinite utility streams: A new welfare criterion and its axiomatic characterization', in *Journal of Economic Theory* 133/1: 350–373.

[36] We both incline at least somewhat towards totalism. One of us (HW) also inclines toward risk neutrality, while the other (CT) does not, but thinks that the correct principles for evaluation of risky prospects will have similar implications in practice.

[37] 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 prospects. 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 24 above).

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

Bernoulli, D. (1738/1954), 'Exposition of a new theory on the measurement of risk', in *Econometrica: Journal of the Econometric Society* 22(1):23-36.

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

Buchak, L. (2022), 'How should risk and ambiguity affect our charitable giving?' *Global Priorities Institute – Working Paper*.

Cain, J. (1995), 'Infinite utility', in *Australasian Journal of Philosophy*: 401–404.

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).

Clark, M. (n.d.), *Infinite ethics, intrinsic value, and the Pareto principle* (unpublished manuscript).

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

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

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

Greaves, H. and W. MacAskill (2021), 'The case for strong longtermism', *Global Priorities Institute - Working Paper*.

Hamkins, J. D. and Montero, B. (2000), 'Utilitarianism in infinite worlds', in *Utilitas* 12/01: 91–96.

Harsanyi, J. C. (1955), 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility', in *Journal of Political Economy* 63/4: 309–321.

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

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

Koopmans, T. C. (1960), 'Stationary ordinal utility and impatience', in *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', in *Journal of Mathematical Economics* 46/1: 32–37.

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

Liedekerke, L. V. (1995), 'Should utilitarians be cautious about an infinite future?', in *Australasian Journal of Philosophy* 73/3: 405–407.

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

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

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

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

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

Nover, N. and Hájek, A. (2004), 'Vexing expectations', in *Mind* 113(450):237-249.

Pettigrew, R. (2022), 'Effective altruism, risk, and human extinction', *Global Priorities Institute – Working Paper.*

Smith, Q. (2003), 'Moral realism and infinite spacetime imply moral nihilism' in H. Dyke (ed.), *Time and Ethics: Essays at the Intersection* (Springer), 43–54.

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

Tarsney, C. and Thomas, T. (2020), 'Non-additive axiologies in large worlds', *arXiv preprint arXiv:2010.06842*.

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

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

Thomas, T. (forthcoming), 'The asymmetry, uncertainty, and the long term', in *Philosophy and Phenomenological Research*.

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', in *Australasian Journal of Philosophy* 71/2: 212–217.

Vallentyne, P. and Kagan, S. (1997), 'Infinite value and finitely additive value theory', in *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', in *Philosophical Studies* 178/6: 1917–1949.

Wilkinson, H. (2022a), 'In defence of fanaticism', in *Ethics* 132: 445–477.

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

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

Wilkinson, H. (n.d.b), *How to neglect the long term* (unpublished manuscript).

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