The Epistemic Challenge to Longtermism

Christian J. Tarsney*

Version 4, April 2022
(Latest version here.)

Comments welcome: christian.tarsney@philosophy.ox.ac.uk

Abstract

Longtermists claim that what we ought to do is mainly determined by how our actions might affect the very long-run future. A natural objection to longtermism is that these effects may be nearly impossible to predict—perhaps so close to impossible that, despite the astronomical importance of the far future, the expected value of our present actions is mainly determined by near-term considerations. This paper aims to precisify and evaluate one version of this epistemic objection to longtermism. To that end, I develop two simple models for comparing 'longtermist' and 'neartermist' interventions, incorporating the idea that it is harder to make a predictable difference to the further future. These models yield mixed conclusions: if we simply aim to maximize expected value, and don't mind premising our choices on minuscule probabilities of astronomical payoffs, the case for longtermism looks robust. But on some *prima facie* plausible empirical worldviews, the expectational superiority of longtermist interventions depends heavily on these 'Pascalian' probabilities. So the case for longtermism may depend either on plausible but non-obvious empirical claims or on a tolerance for Pascalian fanaticism.

1 Introduction

If your aim is to do as much good as possible, where should you focus your time and resources? What problems should you try to solve, and what opportunities should you try to exploit? One partial answer to this question claims that you should focus mainly on *improving the very long-run future*. Following Greaves and MacAskill (2021) and Ord (2020), let's call this view *longtermism*. The longtermist thesis represents a radical departure from conventional thinking about how to make the world a better place. But it is supported by *prima facie* compelling arguments, and has recently begun to receive serious attention from philosophers.[1]

*Global Priorities Institute, Faculty of Philosophy, University of Oxford

[1]Proponents of longtermism include Bostrom (2003, 2013) (who focuses on the long-term value of reducing existential risks to human civilization), Beckstead (2013, 2019) (who gives a general

The case for longtermism starts from the observation that the far future is *very big*. A bit more precisely, the far future of human-originating civilization holds vastly greater potential for value and disvalue than the near future. This is true for two reasons. The first is *duration*. On any natural way of drawing the boundary between the near and far futures (e.g., 1000 or 1 million years from the present), it is possible that our civilization will persist for a period orders of magnitude longer than the near future. For instance, even on the *extremely* conservative assumption that our civilization must die out when the increasing energy output of the Sun makes Earth too hot for complex life as we know it, we could still survive some 500 million years.[2] Second is *spatial extent* and *resource utilization*. If our descendants eventually undertake a program of interstellar settlement, even at a small fraction of the speed of light, they could eventually settle a region of the universe and utilize a pool of resources vastly greater than we can access today. Both these factors suggest that the far future has enormous potential for value or disvalue.

But longtermism faces a countervailing challenge: the far future, though very big, is also *unpredictable*. And just as the scale of the future increases the further ahead we look, so our ability to predict the future—and to predict the effects of our present choices—decreases. The case for longtermism depends not just on the intrinsic importance of the far future but also on our ability to predictably influence it for the better. So we might ask (imprecisely for now): does the importance of humanity's future grow faster than our capacity for predictable influence shrinks?[3]

There is *prima facie* reason to be pessimistic about our ability to predict (and hence predictably influence) the far future. First, the existing empirical literature on political and economic forecasting finds that human predictors—even well-qualified experts—often perform very poorly, in some contexts doing little better than chance (Tetlock, 2005). Second, the limited empirical literature that directly compares

defense of longtermism and explores a range of potential practical implications), Cowen (2018) (who focuses on the long-term value of economic growth), Greaves and MacAskill (2021) (who, like Beckstead, defend longtermism generally), and Ord (2020) (who, like Bostrom, focuses mainly on existential risks).

[2]This is conservative as an answer to the question, 'How long is it *possible* for human-originating civilization to survive?' It could of course be very optimistic as an answer to the question, 'How long *will* human-originating civilization survive?'

[3]Versions of this epistemic challenge have been noted in academic discussions of longtermism (e.g. by Greaves and MacAskill (2021)), and are frequently raised in conversation, but have not yet been extensively explored. For expressions of epistemically-motivated skepticism toward longtermism in non-academic venues, see for instance Matthews (2015), Johnson (2019), and Schwitzgebel (2022).

Closely related concerns about the predictability of long-run effects are frequently raised in discussions of consequentialist ethics—see for instance the recent literature on 'cluelessness' (e.g. Lenman (2000), Burch-Brown (2014), Greaves (2016)). Going back further, there is this passage from Moore's *Principia*: '[I]t is quite certain that our causal knowledge is utterly insufficient to tell us what different effects will probably result from two different actions, except within a comparatively short space of time; we can certainly only pretend to calculate the effects of actions within what may be called an 'immediate' future. No one, when he proceeds upon what he considers a rational consideration of effects, would guide his choice by any forecast that went beyond a few centuries at most; and, in general, we consider that we have acted rationally, if we think we have secured a balance of good within a few years or months or days' (Moore, 1903, §93). This amounts to a concise statement of the epistemic challenge to longtermism, though of course that was not Moore's purpose.

the accuracy of social, economic, or technological forecasts on shorter and longer timescales consistently confirms the commonsense expectation that forecasting accuracy declines significantly as time horizons increase.[4] And if this is true on the modest timescales to which existing empirical research has access, we should suspect that it is all the more true on scales of centuries or millennia. Third, we know on theoretical grounds that complex systems can be extremely sensitive to initial conditions, such that very small changes produce very large differences in later conditions (Lorenz, 1963; Schuster and Just, 2006). If human societies exhibit this sort of behavior with respect to features that determine the long-term effects of our actions (to put it *very* roughly), then attempts to predictably influence the far future may be insuperably stymied by our inability to measure the present state of the world with arbitrary precision.[5] Fourth and finally, it is hard to find historical examples of anyone successfully predicting the future—let alone predicting the effects of their present choices—even on the scale of centuries, let alone millennia or longer.[6]

If our ability to predict the long-term effects of our present choices is poor enough, then even if the far future is overwhelmingly *important*, the main determinants of what we presently ought to do might lie mainly in the near future. The aim of this paper is to investigate this epistemic challenge to longtermism.

My ambitions are limited, however. One is simply to state the challenge clearly and distinguish some of its possible variants. The other, which occupies most of the paper, is to evaluate one particular version of the challenge, defined and circumscribed by a number of substantive assumptions, and directed at a particular

[4]See for instance Makridakis and Hibon (1979) (in particular Table 10 and discussion on p. 115), Fye et al. (2013) (who even conclude that 'there is statistical evidence that long-term forecasts have a worse success rate than a random guess' (p. 1227)), and Muehlhauser (2019) (in particular fn. 17, which reports unpublished data from Tetlock's Good Judgment Project).

Muehlhauser gives a useful survey of the extant empirical literature on 'long-term' forecasting (drawing heavily on research by Mullins (2018)). For our purposes, though, the forecasts covered by this survey are better described as 'medium-term'—the criterion of inclusion is a time horizon ≥ 10 years. To my knowledge, there is nothing like a data set of truly long-term forecasts (e.g., with time horizons greater than a century) from which we could presently draw conclusions about forecasting accuracy on these timescales. And as Muehlhauser persuasively argues, the conclusions we can draw from the current literature even about medium-term forecasting accuracy are quite limited for various reasons—e.g., the forecasts are often imprecise, non-probabilistic, and hard to assess for difficulty.

[5]For discussions of extreme sensitivity to initial conditions in social systems, see for instance Pierson (2000) and Martin et al. (2016). Tetlock also attributes the challenges of long-term forecasting to chaotic behavior in social systems, when he writes: '[T]here is no evidence that geopolitical or economic forecasters can predict anything ten years out beyond the excruciatingly obvious—'there will be conflicts'—and the odd lucky hits that are inevitable whenever lots of forecasters make lots of forecasts. These limits on predictability are the predictable results of the butterfly dynamics of nonlinear systems. In my [*Expert Political Judgment*] research, the accuracy of expert predictions declined toward chance five years out' (Tetlock and Gardner, 2015). But Tetlock may be drawing too pessimistic a conclusion from his own data, which show that the accuracy of expert predictions declines *toward* chance, while remaining significantly *above* chance—for discussion, see §1.7 of Muehlhauser (2019).

[6]There are some arguable counterexamples to this claim—e.g., the founders of family fortunes who may predict with significantly-better-than-chance accuracy the effects of their present investments on their heirs many generations in the future. (Thanks to Philip Trammell for this point.) But on the whole, the history of thinking about the distant future seems more notable for its failures than for its successes.

version of longtermism. This modest approach is, I think, forced on us by the nature of the question. Both the scale of the far future and our ability to predictably affect it are matters of degree, and so assessing the epistemic challenge is inevitably a quantitative exercise. The structure of this exercise, and its quantitative inputs, are sensitive to various background assumptions, including one's choice of epistemic, ethical, and decision-theoretic frameworks, and an open-ended list of empirical assumptions (e.g., technological assumptions about the eventual capabilities of an advanced civilization, and cosmological assumptions about the long-term fate of the universe). We can therefore only assess the epistemic challenge by considering different sets of plausible assumptions one at a time. I have tried to assume no more than is necessary to make the challenge tractable (that is, to bite off as large a piece of the challenge as I can fruitfully evaluate in one paper), and to make assumptions that are substantively plausible and otherwise well-motivated. But I also hope that the exercise I work through in this paper will provide a model that can be adapted to address other variants of the epistemic challenge, under different sets of assumptions.

The most significant assumptions I make in this paper are as follows: first, I assume a precise probabilist epistemic framework. Specifically, I assume that rational agents ought to assign (precise) probabilities to decision-relevant possibilities (e.g., to the world being in a particular state, or a particular action having a particular outcome), in a way that is constrained (though not necessarily uniquely determined) by the agent's evidence. Second, I assume a total welfarist consequentialist normative framework. And third, I assume a decision-theoretic framework of expected value maximization. I will refer to the conjunction of these three assumptions as *expectational utilitarianism*, for short. I will call any challenge to longtermism that does not require rejecting expectational utilitarianism an *empirical challenge*, since it does not rely on normative claims unfavorable to longtermism, and I will call anyone who is skeptical of longtermism even conditional on expectational utilitarianism an *empirical skeptic*. I choose this set of assumptions partly because they represent a widely held package of views, and partly because I find them plausible. But they also serve to screen off various other, non-empirical challenges to longtermism (e.g., ethical and decision-theoretic), so that we can consider the strength of the epistemic challenge in isolation, in a setting otherwise favorable to longtermism.[7]

On the other hand, when it comes to empirical questions (e.g., choosing values for model parameters), I will err toward assumptions unfavorable to longtermism, in order to test its robustness to the epistemic challenge within the normative framework of expectational utilitarianism.

The paper proceeds as follows: in §2, I state the longtermist thesis a bit more precisely, and identify a version of longtermism focused on 'making persistent differences', which will be the focus of our investigation. In §3, I similarly attempt to precisify the epistemic challenge, and to circumscribe the version of the challenge

[7]For a version of the epistemic challenge that arises in an imprecise probabilist framework, see Mogensen (2021). For discussion of axiological and ethical challenges to longtermism, see Beckstead (2013) and Greaves and MacAskill (2021). And for discussion of the decision-theoretic worry that the case for longtermism depends on 'fanatical' application of expected value reasoning, see Tarsney (2020), Beckstead and Thomas (2020), and Wilkinson (2022).

that will be our focus. In §4, I describe the first model for comparing longtermist and neartermist interventions. The distinctive feature of this model is its assumption that humanity will eventually undertake an indefinite program of interstellar settlement, and hence that in the long run, the potential value of human-originating civilization grows as a cubic function of time, reflecting our increasing access to resources as we settle more of the universe. In §5, by contrast, I consider a simpler model which assumes that humanity remains Earth-bound and eventually reaches a 'steady state' of zero growth. §6 considers the effects of higher-level uncertainties—both uncertainty about key parameter values and uncertainty between the two models. Accounting for these uncertainties makes the expectational utilitarian case for longtermism substantially more robust, but in a way that leaves it vulnerable to charges of 'fanaticism' (reliance on small probabilities of extreme outcomes), which I briefly discuss. §7 takes stock, organizes the conclusions of the preceding sections, and surveys several other versions of the epistemic challenge that remain as questions for future research.

2 Longtermism and persistent differences

I will understand *longtermism* as the following thesis.

Longtermism In most choice situations (or at least, most of the most important choice situations) faced by present-day human agents, what the agent ought to do all things considered is mainly determined by the possible effects of her actions on the far future.[8]

Thus stated, the longtermist thesis clearly inherits the vagueness of terms like 'most', 'the most important choice situations', 'present-day', 'mainly determined by', and 'far future'. For the most part, I will not try to say how these terms should be precisified.[9]

Importantly for our purposes, the 'ought' in the longtermist thesis should be understood as 'evidence-relative'—i.e., as expressing what an agent ought to do not in light of the way the world in fact is and the actual consequences of her actions (which she might have no way of knowing), nor in light of her subjective credences (which might be simply irrational), but in light of the epistemic probabilities, the probabilities supported by her evidence. I won't try to say exactly what epistemic probabilities are, but they should be understood as depending not just on 'raw' evidence (e.g., sensory inputs) but also on the agent's cognitive capacity for extracting useful information from that raw evidence. That is, the interesting question for practical purposes is not whether an agent who is exposed to exactly my sensory

[8]This closely resembles the initial, informal statement of 'Deontic Strong Longtermism' in Greaves and MacAskill (2021, p. 3).

[9]The exception is 'far future', which I will later precisify as 'more than 1000 years from the present'. This gives a rough sense of what longtermists mean by 'long-term' or 'far future', though some longtermists think that what we ought to do is mainly determined by considerations much more than 1000 years in the future.

For one attempt to precisify 'mainly determined by', see Greaves and MacAskill's final statement of Deontic Strong Longtermism (Greaves and MacAskill, 2021, p. 26).

inputs, but who also has unlimited computational and reasoning abilities, would be able to foresee important long-term effects from her present choices, but whether *I*, with my much more limited capacities, can do so.

Longtermists hold that our choices can make a significant difference to the value of the far future, but may disagree about what particular strategies we should pursue to make the future go well. One broad class of longtermist strategies, which will be my primary focus in this paper, involve *trying to make a persistent difference*. That is, the longtermist (i) identifies some state S judged to be better *ex ante* (meaning, I will assume, in expectation) than its complement state $\neg S$ and (ii) tries to intervene on the world so that state S is realized when state $\neg S$ would otherwise have been realized, with the intention that (iii) the world will then *remain* in state S when (iv) it would otherwise have remained in state $\neg S$. For instance, they might aim to bring about some improvement in norms, values, or institutions (e.g., recognition of the moral status of some class of beings, constitutional protection of certain rights and liberties, a more democratic government or independent judiciary in a particular country...), with the intention that this improvement will persist for a long period of time in which it might otherwise not have happened at all. Or they might aim to reduce some catastrophic risk to human civilization, with the aim that civilization itself will persist when it would otherwise have been (and remained) destroyed. Let's call longtermist strategies of this form *persistent-difference strategies*.[10]

The expected value of pursuing any longtermist strategy of this kind depends on four factors, corresponding to (i)–(iv) above.

1. **Importance:** It is significantly better in expectation for the world to be in state S rather than $\neg S$ at a given time.

2. **Tractability:** It is possible to significantly increase the probability that the world is in state S at some future time t.

3. **Persistence:** State S is *persistent*, meaning that when S obtains, there is a low probability per unit time of transitioning to $\neg S$ (or more generally, the expected length of time S will continue to obtain is large).

[10]Not all longtermist strategies have this form—or at least, not all are most naturally described as having this form. For instance, 'speed-up' strategies aim to bring about either a permanent acceleration or a one-time forward shift in some positive trend (e.g., economic growth (Cowen, 2018) or space settlement (Bostrom, 2003)). If such a strategy succeeds, then at each future time, we will be more advanced/better off than we otherwise would have been. But we will not (in any intuitive sense) be put in some persistent state that we would otherwise not have been in. Or, more abstractly, one might imagine that on some important dimension (say, the quality of our moral values), human societies follow an unbiased random walk, changing for better or worse with equal probability in each time period. An intervention that moved a society one step in the positive direction on this dimension (say, to +3 instead of +2) would improve our expected position at each future time by one step, without making any persistent difference (e.g., not keeping us persistently at +3 or persistently above +2)

Nevertheless, I will limit my focus to persistent-difference strategies, which seem to capture most (though not all) plausible strategies for improving the far future. Insofar as this one type of longtermist strategy can survive the epistemic challenge (which will be my provisional conclusion), then longtermism itself can survive the epistemic challenge. But it would, of course, be interesting to investigate epistemic worries about non-persistence-based strategies for improving the far future as well.

4. **Complement Persistence:** State $\neg S$ is persistent, in the same sense.

As we will see, each of these factors is a potential target for empirical challenges to longtermism generally, and for epistemic challenges more specifically.

3 'Control challenges' and the 'epistemic challenges' to longtermism

The empirical skeptic of longtermism denies that any persistent-difference strategy has very great long-term expected value—specifically, that any such strategy can improve the far future in expectation enough to outweigh the near-term expected value of available neartermist alternatives. I will assume (as seems to be true of most real-world debates) that the primary disagreement between longtermists and empirical skeptics is not about the expected value of available neartermist interventions (i.e., how much good we can do in the near term) nor about harmful side-effects of longtermist interventions, but rather about *the feasibility of predictably improving the far future*, and therefore the amount of far-future expected value that can be generated by pursuing longtermist objectives. In taking the pessimistic side of this question, the empirical skeptic must make one or both of the following claims:

The Control Challenge (rough) There's simply nothing we can do to substantially improve the far future. That is, even if we were maximally informed, we would still lack the necessary power or influence to make any sufficiently important and persistent difference.

The Epistemic Challenge (rough) Even if there are actions available to us that would substantially improve the far future, we lack the epistemic capacities necessary to distinguish those actions from actions that would either worsen the far future or have no substantial effect. As a result, none of the actions available to us substantially improve the far future *in expectation*.

The difference between these challenges concerns the counterfactual question of what we *would* be able to do from an ideal epistemic position—if we were maximally informed about the consequences of each available action. I will understand 'maximally informed' to imply not knowledge of what the actual consequences of each action would be (since there may be no fact of the matter about the consequences of actions we won't in fact take, due to either indeterministic physics or counterfactual underdetermination (Hare, 2011; Portmore, 2016)), but rather perfect knowledge of the objective chances of particular actions having particular consequences. We can then state the two challenges more precisely as follows:

The Control Challenge (less rough) At least in most choice situations faced by present-day human agents, there are no available actions that substantially increase the chance-expected value of the far future (i.e., the expected value taken relative to objective chances).

The Epistemic Challenge (less rough) At least in most choice situations faced by present-day human agents, even if there are actions available that would substantially increase the *chance-expected* value of the far future, there are no available actions that substantially increase the *epistemic* expected value of the far future (i.e., the expected value relative to epistemic probabilities).

A further distinction: with respect to any particular persistent-difference strategy, the disagreement between the longtermist and the empirical skeptic may focus on any of the four factors identified above (importance, tractability, persistence, complement persistence). My primary interest in this paper, however, will be in the last two factors (persistence and complement persistence) and in challenges to longtermism that focus on them. *Persistence skepticism* toward longtermism is a form of empirical skepticism according to which persistence and/or complement persistence are the major points of failure among the persistent-difference strategies advocated by longtermists—that is, the factors that longtermists most seriously overestimate and that play the largest role in cutting naive estimates of the expected value of longtermist interventions down to size.[11] My focus on persistence skepticism means, in particular, that the models developed in §§4–5 below are focused on modelling how different assumptions about persistence/complement persistence affect the expected value of longtermist interventions, while relying on simplified treatments of importance and tractability.

I adopt this focus because persistence skepticism strikes me as the most plausible (though not the *only* plausible) form of empirical skepticism toward persistent-difference strategies, for two reasons. First, naive back-of-the-envelope estimates of the expected value of longtermist interventions (particularly existential risk mitigation) typically generate numbers many orders of magnitude larger than any effects we can plausibly have on the near future (see, for instance, Bostrom (2003, 2013)). As we will see, because the possibility of persistence failures acts like a discount rate on the expected value of persistent-difference strategies, pessimistic assumptions about persistence or complement persistence can reduce those naive estimates by many orders of magnitude, potentially below the expected value of neartermist alternatives. It seems more difficult, though, for plausible pessimism about the importance or tractability of longtermist objectives to have a similarly dramatic effect. Second, persistence and complement persistence are the factors that distinguish persistent-difference strategies from neartermist alternatives. That is, any general skepticism about the tractability of altruistically motivated actors substantially changing the world (e.g., on the grounds that existing arrangements represent stable equilibria that are generally difficult or harmful to disrupt) or about the importance of apparently-desirable changes (e.g., because of hedonic adaptation) will apparently deflate the expected value of neartermist and longtermist strategies

[11] The persistence skeptic need not claim that it is impossible to make *any* highly persistent difference the world. They might, for instance, concede that some trivial differences can be extremely persistent—e.g., if I bury a corrosion-resistant object deep underground, or launch Russell's teapot into a stable orbit around the Sun, I can be reasonably confident that even a million years from now, my object (i) will be where I put it and (ii) would not have been there, if not for my action. The objection to these actions as strategies for improving the far future is, of course, that the difference they make is unimportant.

alike, leaving the comparative merits of neartermism and longtermism unaffected. But general skepticism about the *persistence* (or complement persistence) of altruistically motivated improvements to the world does provide an obvious reason to favor neartermism over longtermism.

Just like empirical skepticism generally, persistence skepticism can take either a 'control' or an 'epistemic' form. The persistence-based control challenge asserts that the differences targeted by longtermists pursuing persistent-difference strategies are *all*, *inevitably* only weakly persistent. Even if we manage, for instance, to effect a positive change in human institutions or values, either there are insuperable factors that will almost certainly (with high objective chance) cause things to revert to the *status quo ante* in not too long, or else it is extremely likely that the same positive change would have happened anyway, without our intervention, in not too long.

The persistence-based epistemic challenge, on the other hand, asserts that while some persistent-difference strategies might have very high chance-expected value (which we could identify, if we knew the objective chances), we lack the epistemic capacities needed to identify those strategies—either to identify the particular differences that have high persistence and complement persistence or, with respect to a particular difference, to identify the particular, fine-grained intervention (sequence of actions) that would be necessary to make it highly persistent. For instance, perhaps some features of human institutions or values exhibit multiple equilibria in a way that makes them subject to strong lock-in effects (where, after a certain time, it becomes very unlikely that they will exit whichever equilibrium they have found themselves in); if we knew which features those were, we would have a chance to influence them while they are still malleable, and thereby make a persistent positive difference; but our understanding of the dynamics of human societies is too impoverished to let us identify those features. Or, alternatively, it might be that many positive changes *can* be made persistent, with sufficient skill—e.g., good institutions can be made persistent by protecting them with constitutional arrangements that are both strong enough to resist change by transiently powerful bad actors, and flexible enough to adapt to new circumstances when needed; good values can be made persistent when buttressed by the right combination of ideas, traditions, and institutions (as in the case of long-lived religions). But we lack the strategic understanding to identify the very particular sequence of actions necessary to buttress our positive changes and make them persistent. In either case, the greater the degree of persistence to which we aspire, the harder it will be to identify suitbable objectives and interventions, and in this way our epistemic limitations make it harder to predictably influence the further future.

The models developed in §§4–5 are meant to illuminate the effects of our assumptions about persistence and complement persistence on the expected value of persistent-difference strategies. They do not formally distinguish, however, between 'control' and 'epistemic' versions of persistence skepticism: they involve only a single probability measure, which could be interpreted as either epistemic or objective, so the expected value estimates they generate could be interpreted as either epistemic expectations or chance-expectations. In applying the models, I will interpret the probabilities as epistemic, with the aim of estimating the epistemic expected value of a hypothetical longtermist intervention. Even under this interpretation,

the results of the exercise do not distinguish between control worries and epistemic worries about persistent-difference strategies: if our conclusions are favorable to longtermism, then both challenges are mistaken; if unfavorable, either challenge could be the correct one. (Which challenge is correct depends on what our conclusions would have been if we had done the same exercise using objective chances.) And for practical purposes, there is not much need to distinguish the two challenges: what matters is how we should rank our options given our *actual* epistemic position, not how things might change if we were in the radically idealized epistemic position of chance-omniscience.[12]

Nevertheless, I conceive this exercise primarily as an investigation of the epistemic challenge, because the epistemic challenge seems much more plausible than the control challenge. Omniscience—even mere chance-omniscience—is extremely powerful. An agent freed from most epistemic limitations (apart from foreknowledge of their own choices, of the outcomes of indeterministic events, and of indeterminate counterfactuals) would almost certainly have enormous power, even if they had only the non-epistemic capacities of an ordinary human. They could consider each element of a vast space of possible strategies (all sequences of actions they have the ability to perform), and evaluate the consequences of each of those strategies with vastly greater accuracy than our own epistemic capacities permit. It is hard to believe that such an agent could not find any way of having a persistent, positive influence on the far future. They could, for instance, write an optimific book containing (i) a list of precise instructions for key individuals and institutions to put human civilization on a positive long-term trajectory, along with (ii) several revolutionary mathematical theorems, new scientific discoveries, and surprising-but-accurate predictions, impressive enough to convince those key actors to follow the book's advice. Though such an agent would have vast influence on both the near and the far future, the vastly greater scale/stakes of the far future mean that longtermism would almost certainly be true for them (assuming expectational utilitarianism). If longtermism is false with respect to you or me, then, it is because of our epistemic deficiencies. That is, if there is a successful empirical challenge to longtermism, it is an epistemic challenge, not a control challenge.

Just as, for practical purposes, not much hangs on the distinction between the control and epistemic challenges, so, for purposes of this paper, not much hangs on whether you accept my assessment of the relative plausibility of these challenges. You can understand the following investigation as addressing either challenge or both, according to which seems most in need of address. But it's important to recognize that worries about the *long-term persistence* of the effects of our actions are not necessarily distinct from worries about our ability to *predict* the long-term effects of our actions—the former can be a manifestation of the latter. (The positive differences we make to the world have low expected persistence, so the worry goes, because we lack the epistemic capacity to identify the differences that are potentially persistent, and the specific actions needed to make them persistent.) Note also that insofar as our investigation turns up conclusions favorable to longtermism (which it

[12]Though the latter question is not entirely idle: it tells us something about the expected value of improving our epistemic position, bringing our epistemic probabilities into closer alignment with the objective chances.

will, in the main), this serves *a fortiori* to address the epistemic version of persistence skepticism, since it disconfirms persistence skepticism in general.

To sum up, the focus of the following investigation is on a particular piece of the epistemic challenge to longtermism, though one that I take to be particularly significant: *epistemic persistence skepticism*. This is the worry, with respect to persistent-difference strategies, that it is prohibitively difficult to identify potentially persistent differences and strategies for ensuring their persistence; that as a result the epistemic prospects for any particular intervention having extremely-long-lasting positive effects are meager; and that duly accounting for these facts will dramatically deflate naive estimates of the expected value of longtermist interventions. There are other pieces to the epistemic challenge to longtermism—both epistemic challenges to longtermist strategies other than persistent-difference strategies, and challenges to persistent-difference strategies that focus more on the importance or tractability of longtermist objectives rather than their persistence or complement persistence. While these other challenges won't be my focus, I'll return to them briefly in §7.

4 The cubic growth model

In this section and the next, I set out two models for the expected value of a longtermist intervention that aims to make a persistent positive difference to the world. Central to both models is the idea that we can influence the probabilities of alternative states of the world less at more remote times. They thereby allow us to evaluate persistence skepticism by quantifying this long-term 'fade-out' of our capacity for predictable influence and seeing how it affects the expected value of longtermist interventions.

There are some precedents for this sort of modeling exercise—in particular, models in a similar spirit to mine are sketched in the appendix of Ng (2016), Appendix E of Ord (2020), and Sittler (ms). The most important contrast between these models and mine is that Ng, Ord, and Sittler are primarily interested in general analytical insights (concerning, e.g., whether a reduction in existential risk should increase or decrease our willingness to pay for further reductions), rather than numerical estimation of the expected value of longtermist interventions. Compared to their models, therefore, the models I develop below will sacrifice some mathematical elegance and simplicity for empirical realism and detail.

The 'cubic growth model' described in this section assumes that our civilization eventually undertakes a program of interstellar expansion, while the 'steady state model' in the next section assumes that we remain Earth-bound. In §4.1, I introduce and motivate the cubic growth model. §4.2 fills in values for its various parameters, with the exception of the crucial parameter that determines how fast our capacity for predictable influence deteriorates. §4.3 presents and discusses the results of the model for a range of values of that crucial parameter.

4.1 Introducing the model

I assume that an agent is faced with a choice between two options, N and L. N is a neartermist 'benchmark' intervention whose expected value lies mainly in the

near future. L is a longtermist intervention that aims to positively influence the far future. In explaining and applying the model, it will be useful to have a working example on which to focus. As our working example, let's suppose that the agent works for a philanthropic organization with a broad remit, and is choosing between two ways of granting \$1 million. N would spend the \$1 million on public health programs in the developing world. L would spend the \$1 million on mitigating existential risks to human civilization, say by supporting research on pandemic risks from novel pathogens.[13]

The neartermist benchmark N, I will assume, has an expected value that is specified exogenously to the model. In the working example, where N represents spending on global public health programs, $\mathrm{EV}(N)$ might be estimated by a standard cost-effectiveness evaluation of the sort produced by charity evaluators like GiveWell. Extrapolating from GiveWell's highest current cost-effectiveness estimate for any global public health intervention, I will assume that N yields 10,000 quality-adjusted life years (QALYs) in expectation.[14]

For simplicity, let's normalize our value scale so that the expected value of the 'status quo' (doing nothing with the \$1 million, or burning it, or spending it in some non-philanthropic way) is 0, and $\mathrm{EV}(N)$ is 1. Let's call a unit of value on this scale a *valon* (abbreviated V)—that is, one valon is a unit of value equivalent to 10,000 QALYs. Thus, L has greater expected value than N in the working example iff $\mathrm{EV}(L) > 1$ V.

I assume that L is equivalent to the status quo in the near future—i.e., its benefits

[13]I choose existential risk mitigation as the working example of a persistent-difference strategy mainly because it's especially easy to quantify—that is, it's easier to make empirically motivated estimates of the various model parameters for this application than for others. But the model is meant to describe persistent-difference strategies in general, so it could also be applied, for instance, to efforts to persistently improve political institutions or moral values.

[14]For interventions whose primary benefit is saving lives, GiveWell estimates an average cost per life saved. In its most recent cost-effectiveness estimates (GiveWell, 2021), vitamin A supplementation (as implemented by Helen Keller International) had the lowest estimated cost per life saved, at approximately \$3000. (For more details, see GiveWell's cost-effectiveness models at `https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models`.) Assuming constant returns (in line with our practice of making empirical assumptions unfavorable to longtermism), this implies that \$1 million in funding for vitamin A supplements will save $333\frac{1}{3}$ lives in expectation. To allow comparison with our longtermist intervention L, it is useful to convert this to QALYs. I will therefore assume that expected value of saving a life is 30 QALYs, meaning that \$1 million spent on vitamin A supplements has an expected value of 10,000 QALYs.

It is not obvious, of course, that public health interventions aimed at saving lives in the developing world are the *most* cost-effective neartermist intervention. Some interventions to benefit poor people in the developing world have other primary benefits (e.g., direct cash transfers and deworming treatments), but are arguably competitive with life-saving interventions in terms of value per dollar spent. And interventions focused on the welfare of non-human animals (e.g., to promote veganism or improve conditions for farmed animals) are arguably more cost-effective than any neartermist interventions primarily benefiting humans. I focus on life-saving public health interventions to avoid tendentious and highly uncertain value comparisons between very different altruistic payoffs. But also, as we will see, the main qualitative conclusions we reach below would not be changed very much by an adjustment of one or two orders of magnitude in the expected value of the benchmark neartermist intervention, so as long the estimate we're using is not *vastly* too low, it should be adequate for our purposes.

(if any) lie in the far future. More specifically, L aims to increase the probability that the world is in some target state S in the far future. In the working example, where L aims to mitigate existential risk, S can be interpreted as something like: 'The accessible region of the universe contains an intelligent civilization.'[15]

The model aims to estimate the expected value of L that accrues in the far future. So we will designate the boundary between the near future and the far future as $t = 0$. What distinguishes the 'far' future, for our purposes, is our lack of any fine-grained information that might enable detailed causal models of the effects of our interventions. When thinking about the far future, the model assumes, we may be able to predict some general trend lines (e.g., that the spatial extent of human-originating civilization will increase with time), but cannot predict local fluctuations around those trend lines (as we can do in the near future, e.g., for economic growth, crime rates, etc.) or other particular events. In the working example, I will assume that the boundary between the near and far future is 1000 years from the present (i.e., in the year 3022). Time in the far future is measured from this boundary, so for instance $t = 6$ years corresponds to the year 3028.

Let S_0 designate the event of the world being in the target state S at $t = 0$, and $\neg S_0$ designate its complement. More generally, S_t is the event of the world being in state S at time t, and $\neg S_t$ is its complement. In the working example, where S means 'The accessible universe contains an intelligent civilization', S_0 means 'The accessible universe contains an intelligent civilization in the year 3022', which is *roughly* equivalent to 'Humanity survives the next thousand years.'

We will model persistence worries by the possibility of what I will call *exogenous nullifying events* (ENEs). These come in two flavors:

- **Negative ENEs** are events in the far future (i.e., after $t = 0$) that put the world into state $\neg S$. In the working example, where S represents the existence of an intelligent civilization in the accessible universe, a negative ENE is any existential catastrophe that might befall such a civilization: e.g., a self-destructive war, a lethal pathogen or meme, or some cosmic catastrophe like vacuum decay.

- **Positive ENEs** are events in the far future that put the world into state S. In the working example, this is any event that might bring a civilization into existence in the accessible universe where none existed previously. The obvious ways this could happen include the evolution of another intelligent

[15] 'The accessible region of the universe' or 'accessible universe' refers to our future light cone, that is, the region of spacetime that it is possible to reach from Earth today travelling at or below the speed of light.

For the sake of conservatism, I will assume throughout the paper that we are in fact limited by the speed of light, and cannot reach or exploit the resources of regions outside our future light cone. Likewise, I set aside various other physical and technological possibilities that might greatly expand the reach or increase the capacities of future civilization: e.g., that we live in a Gödel spacetime containing closed timelike curves, or can construct computers capable of computational supertasks in finite time, or can persist as a civilization for infinite time (as in some cyclic cosmological models). In general, accounting for such possibilities is only likely to strengthen our qualitative conclusions, by increasing the potential scale of the far future and thereby making the expectational case for longtermism even more robust under uncertainty, but also exacerbating worries about Pascalian fanaticism (assuming we assign these scenarios low probability).

species on Earth (at a time when all previously existing intelligent species have died out) or somewhere else in the accessible universe, or the arrival of another expanding civilization from outside the accessible universe.

What negative and positive ENEs have in common is that they 'nullify' the intended effect of the longtermist intervention. After the first ENE occurs, it no longer matters (at least in expectation) whether the world was in state S at $t = 0$, since the current state of the world no longer depends on its state at $t = 0$. If a negative ENE has occurred, the world will immediately thereafter be in state $\neg S$, regardless of what state it was in at $t = 0$, and its subsequent state will depend only on the pattern of future ENEs, not on the state of the world at $t = 0$. And similarly for positive ENEs. Thus, if the longtermist intervention L succeeds in making a difference by putting the world into state S at $t = 0$, this difference will persist until the first ENE occurs.

Calling ENEs 'exogenous' means simply that they are exogenous to the *model*—they need not be exogenous to the civilization they affect (e.g., they include events like self-destructive wars). More precisely, we assume that ENEs are probabilistically independent of the choice between L and N, from the agent's perspective.

The possibility of ENEs is the first key assumption of the cubic growth model. The second is that (conditional on survival) human-originating civilization will eventually begin to settle other star systems, and that this process will (on average over the long run) proceed in all directions at a constant speed. Further, the model assumes that the expected value of a civilization in state S at time t is proportionate to its resource endowment at t, which grows (not necessarily linearly) with the spatial volume it occupies. A civilization's resource endowment (in particular, the quantities of raw materials and usable energy at its disposal) determine how large a population it can support, which it turn determines its value in total welfarist consequentialist terms. I assume that in the long run and on a large enough scale, an interstellar civilization will convert resources into population and welfare at some roughly constant average rate.[16]

We have now described the main features of the model informally. The next step is to formalize the model as an equation for the expected value of the longtermist intervention L. I will first introduce the parameters that figure in this equation, then state the equation itself.

The model parameters are as follows:

[16]By assuming a constant speed of space settlement, the cubic growth model neglects two effects that are important over very long timescales: first, the assumption of a constant speed of space settlement *in comoving coordinates* (implicit in taking spatial volume as a proxy for resources) ignores cosmic expansion, which becomes significant when we consider timescales on the order of billions of years or longer (Armstrong and Sandberg, 2013, pp. 8–9). Second, it ignores the declining density (even in comoving coordinates) of resources like usable mass and negentropy predicted by thermodynamics, which becomes significant on even longer timescales. If we were using the model to make comparisons between longtermist interventions, these considerations would be significant and would have to be accounted for. But for our purpose of comparing a longtermist with a neartermist intervention, these effects can be safely ignored: as we will see, if events a billion years or more in the future make any non-trivial difference to $\mathrm{EV}(L)$, then L has already handily defeated N on the basis of nearer-term considerations.

1. t_f is what I will call the 'eschatological bound': the time after which the universe can no longer support intelligent life and beyond which, we will assume, there is no longer any difference in expected value between L and N.[17] The most natural candidate for an eschatological bound is the heat death of the universe, though as we will see, it does not matter very much which of the various plausible bounds we select.

2. p is the amount by which the longtermist intervention changes the probability of being in the target state S at $t = 0$, relative to the neartermist benchmark. Formally, $p = Pr(S_0|L) - Pr(S_0|N)$.[18]

3. v_e is the difference between states S and $\neg S$ in *expected value realized on Earth* per unit time. (As we will see, separating value realized on Earth from value realized in the rest of the universe increases the accuracy of the model when the rate of ENEs is high.)

4. v_s is the difference in expected value between states S and $\neg S$ *per star in the region of settlement* per unit time, excluding value realized on Earth. In the working example, this is the difference in expected value between the existence and non-existence of an intelligent civilization in the accessible universe, per available star per unit time.

5. t_l is the time at which interstellar settlement commences, relative to the near future/far future boundary.

6. s is the speed of interstellar settlement.

7. n is a function that gives the number of stars within a sphere of radius x centered on Earth, and hence the number of stars that will be available at a given time in the process of space settlement. Since stars (and mass/energy resources in general) are many orders of magnitude more abundant in our immediate environment than in the universe as a whole, the early years of space settlement will be unusually fruitful, and we will be badly misled if we do not account for this. Since our aim in this paper requires only order-of-magnitude accuracy, however, I will use a relatively crude density function, characterized by just two parameters: d_g, the number of stars per unit volume within 130,000 light years of Earth (a sphere that safely encompasses the Milky Way) and d_s, the number of stars per unit volume in the Virgo Supercluster (which contains the Milky Way).[19]

[17] Of course, there may be no such time. (For instance, in a cyclic cosmology like the Steinhardt-Turok model, a civilization might be able to persist indefinitely if it can transmit information and therefore perpetuate itself from one cycle to the next.) But I assume for the sake of conservatism that there is such a bound.

[18] Note that p is not a probability but a *difference* of probabilities, and can therefore be negative. But of course an agent will only entertain L as a strategy for ensuring that the world is in state S at $t = 0$ if she judges that $p = Pr(S_0|L) - Pr(S_0|N) > 0$.

[19] I use the star density of the Virgo Supercluster rather than the accessible universe as a whole because whether L or N has greater expected value in the model is almost entirely determined by the 'early' period of space settlement—on the order of tens to hundreds of millions of years—during which we remain confined to the supercluster.

8. r is the *rate* of ENEs, i.e., the expected number of ENEs (positive or negative) per unit time. For now, we assume that this rate is constant (an assumption I will defend shortly), though in §6 we will consider the effects of uncertainty about r, which introduces a form of time-dependence (see fn. 37).

We can now state the model itself:

Cubic growth model

$$\mathrm{EV}(L) = p \times \int_{t=0}^{t_f} (v_e + v_s n((t - t_l)s)) \times e^{-rt}\, dt$$

where n is a function from distance to number of stars (roughly estimating the number of stars within that distance of Earth), defined as:

$$n(x) = \begin{cases} 0 & x \leq 0 \\ \frac{4}{3}\pi x^3 d_g \text{ stars} & 0 \leq x \leq 1.3 \times 10^5 \text{ ly} \\ \frac{4}{3}\pi(x^3 d_s + (1.3 \times 10^5)^3(d_g - d_s)) \text{ stars} & x \geq 1.3 \times 10^5 \text{ ly} \end{cases}$$

Intuitively, the model equation can be understood as follows: L is an intervention in service of a persistent-difference strategy, aiming to increase the probability that the world is in state S at the near future/far future boundary ($t = 0$), in the hope that it will thereafter remain in state S. L increases the overall probability of the world being in state S at $t = 0$ by some amount p. This is then multiplied by the expected value of starting off in state S rather than $\neg S$, which is given by the integral over all future times up to the eschatological bound of the product of two terms:

- $(v_e + v_s n((t - t_l)s))$, which represents the expected value of being in state S rather than $\neg S$ at time t. v_e represents the expected betterness of state S on Earth, v_s represents the expected betterness of S per settled star system, and $n((t - t_l)s))$ gives the number of stars we will have settled by time t. Cubic growth in the model comes from the fact that this number grows cubically as function of time, via the x^3 terms in n. Importantly, though, there is not one steady cubic growth trajectory; rather, we transition from one cubic growth trajectory to another, slower cubic growth trajectory once we have finished settling the star-dense Milky Way.

- e^{-rt}, which represents the probability that no ENE occurs before time t. We care about this probability, rather than the probability that the world is in state S at time t, because we are interested not in the absolute expected value of the far future conditional on L, but in the *difference* L makes to the expected value of the far future compared to N. And if an ENE has occurred before time t then the state of the world (S or $\neg S$) will be the same regardless of whether L or N was chosen.

The two most notable features of the model are (1) the cubic growth term $v_s n((t-t_l)s))$ and (2) the assumption of a constant probability of ENEs, which amounts to an exponential discount on the stream of expected value associated with L. As we will see, plausible values of r yield discount rates that are quite small relative to the rates typically used in economic models. Nevertheless, the combination of polynomial growth and *any* positive exponential discount rate, however small, means that the discount rate eventually 'wins': after some point, the integrand of $\mathrm{EV}(L)$ will go monotonically to zero, and quickly enough that $\mathrm{EV}(L)$ is guaranteed to be finite even without the presence of the eschatological bound.[20,21]

The model involves several significant simplifications. I will discuss three of them here, and give a more complete accounting in the appendix. First is the relatively crude approximation of the rate at which our resource endowment grows as we expand into the cosmos. For our purposes (namely, comparing a neartermist and a longtermist intervention, rather than making comparisons among longtermist interventions), it is the early years of space settlement that are crucial. So I have tried to capture the two most important inhomogeneities in the growth of our resource endowment during those early years: the fact that Earth is settled to begin with, and the relative abundance of stars in the Milky Way. Still, the function n short-changes the case for longtermism more than a little, since stars are still more abundant within (say) 100 light years of Earth than within 130,000 light years. This is partly offset, however, by the model's generous assumption that once a star is settled, it immediately begins producing value at its 'mature' rate. It is plausible that, especially in the first years of space settlement, there will be a 'ramp-up period' or learning curve that prevents us from immediately converting our abundant local resource endowment into value.

A second important simplification is the assumption that the longtermist intervention L only aims to affect one feature of the far future, namely, whether we are in state S or state $\neg S$. In reality, of course, actions affect the world in multiple ways. Research on AI value alignment, for instance, might simultaneously increase the probability that our civilization survives the next 1000 years and increase the

[20]To illustrate both the significance and the limitations of these observations, consider an analogy. Why, if we accept longtermism, should we not accept 'ultra-longtermism', which holds that what we ought to do is mainly determined by the potential consequences of our actions more than (say) Graham's Number years in the future? One apparently very good reason is *proton decay*: it is widely believed (though not yet confirmed) that protons eventually decay into lighter particles, with a half-life on the order of 10^{30} years or longer (Langacker, 1981). If proton decay occurs, we might think of it as imposing a sort of exponential discount rate on our projects, since the resources with which we might eventually reap the rewards of those projects are literally evaporating at an exponential rate. But if protons have a half life of 10^{30} years, then the implied annual discount rate is approximately $-\frac{\ln(0.5)}{10^{30}} \approx 7 \times 10^{-29}$. This discount rate is bound to *eventually* overwhelm any polynomial rate of growth, and therefore provides a sufficient (though probably not necessary) reason why most of our practical concern should be kept within some finite temporal limit. At the same time, it illustrates that a small enough exponential discount rate can still be completely irrelevant to the 'moderate' longtermist thesis we are considering here.

[21]The assumption of merely-polynomial growth may seem revisionary relative to the exponential growth assumed in standard economic models. But this latter is growth in *consumption*, whereas we are concerned with growth in *total welfare*, which is not standardly assumed to be exponential (since individual welfare or utility is treated as a concave function of consumption, often even assumed to be bounded above).

probability, conditional on survival, that the denizens of our civilization 1000 years from now have high rather than low average welfare. I adopt the stylized assumption of a single, binary objective mainly for simplicity and tractability. But it also seems plausible that, in most cases, there will be order-of-magnitude differences between the increments of expected value generated by the various objectives of a given longtermist intervention, in which case we can safely focus on the *most* important objective without much loss of accuracy.

A third important simplification is treating r as time-independent. In the context of the working example, for instance, many people believe that we live in a 'time of perils' (Sagan, 1994) and that the risk of existential catastrophes (i.e., negative ENEs) is likely to decline over time, especially as we begin settling the stars and so hedge our bets against the sort of local catastrophes that might befall a single planet or star system.

I make the assumption of time-independence, again, partly for simplicity and tractability. But it is also in keeping with the principle of making the empirical assumptions that are least favorable to longtermism, within reason. While the 'time of perils' hypothesis is plausible, it is of course still highly speculative.[22] With respect to the protection afforded by space settlement, the existential risks against which space settlement is clearly protective (e.g., asteroids, climate change, supervolcanoes) are arguably minor compared to other risks (e.g., from artificial superintelligence, engineered pathogens, nuclear war) against which it offers limited if any protection (Ord, 2020, pp. 194, 167). And even if you think that these particular risks will subside with time (due to space settlement, improved defensive technologies, or for some other reason), it is notable that the existential risks that most worry us today were mostly unimaginable even 100 years ago, arising as unforeseen consequences of humanity's increasing technological and industrial capabilities. So even if *these particular* risks will decline over time, this provides at best weak evidence that *total* existential risk will decline over time—we might instead simply discover new risks as our capabilities increase further (Ord, 2020, p. 162). (The apparent pattern of increasing existential risk in the 20th and 21st centuries even gives some *prima facie* evidence that those future risks will be *greater* than the risks we face today.) Finally, remember that the cubic growth model applies only to the far future, taken to begin 1000 years from the present. Even if we do live in a time of perils, it is plausible that this period will have largely subsided by 3022 (assuming we survive that long in the first place) and that the risk of existential catastrophe is roughly time-independent thereafter.

4.2 Parameter values for the working example

I will next fill in parameter values for the working example of existential risk mitigation. This serves two purposes. First, it illustrates the model and clarifies the intended interpretations of its parameters. But second, and more importantly, it lets us (partially) assess the challenge of epistemic persistence skepticism toward longtermism. If, under conservative empirical assumptions, a particular longtermist intervention (existential risk mitigation) yields more expected value for the same

[22] For a case against the hypothesis, see Thorstad (2022).

resources than a plausibly-optimal neartermist intervention, then this is substantial evidence that longtermism can survive the challenge. If, on the other hand, there are plausible empirical assumptions under which this particularly promising longtermist strategy yields less expected value than the neartermist alternative, this would bolster the case for epistemic persistence skepticism.

In the cubic growth model, r is both the most consequential parameter and the hardest to estimate. So my approach will be to decide on values for the other parameters and then, using those values, compute $\mathrm{EV}(L)$ for a wide range of possible values of r.

t_f is the easiest parameter to estimate, because it turns out not to matter very much (though it becomes more significant in the steady state model below, for which we will need a revised estimate). I will use the most conservative reasonable basis for t_f, namely, the time at which the last stars are expected to burn out. This gives us $t_f = 10^{14}$ years (Adams and Laughlin, 1997). But the value of t_f is comparatively unimportant because if L is still yielding any significant expected value after roughly $t = 10^8$ years, then it has already accumulated vastly greater expected value than N. That is, bounding the integral anywhere after $t = 10^8$ years will almost never affect whether $\mathrm{EV}(L) > \mathrm{EV}(N)$.

p is more consequential, and harder to estimate. I will make a lower-bound estimate based on the details of the working example, that is almost certainly far too pessimistic, but nevertheless informative. The estimate proceeds in two stages: first, how much could humanity as a whole change the probability of S_0 (i.e., roughly, the probability that we survive the next thousand years), relative to the status quo, if we committed all our collective time and resources solely to this objective for the next thousand years? 'One percent' seems like a very safe lower bound here (remembering that we are dealing with epistemic probabilities rather than objective chances). Now, if we assume that each unit of time and resources makes the same marginal contribution to increasing the probability of S_0, we can calculate p simply by computing the fraction of humanity's resources over the next thousand years that can be bought for $1 million, and multiplying it by 0.01. This yields $p \approx 2 \times 10^{-14}$.[23]

This is a *very* conservative lower bound. First, resources committed to any objective tend to have diminishing marginal impact. And the status quo seems to represent a very early margin with respect to any longtermist objective—that is, we should expect only a small fraction of humanity's resources over the next thousand years to be committed to any given longtermist objective like mitigating existential risks. So we should expect that the marginal impact of a given unit of resources is greater than the average impact of that same unit would be on the assumption that we invest all our resources in that objective. Second, resources committed at an earlier time should have greater impact, all else being equal. (If nothing else, this is true because resources that might be committed to existential risk mitigation, say, 500 years from now can do nothing to prevent any of the existential catastrophes that

[23] Assume a working population of 5 billion, working 40 hours a week, 50 weeks a year. This yields a total of $40 \times 50 \times 5 \times 10^9 = 10^{13}$ work hours per year, or 10^{16} work hours over the next 1000 years. Assume that $1 million is enough to hire ten people for a year (or two people for five years, etc), for a total of 20,000 work hours. This amounts to 2×10^{-12} (two trillionths) of humanity's total labor supply over the next thousand years, and yields $p = 2 \times 10^{-14}$.

might occur in the next 500 years, while resources committed today are potentially impactful any time in the next thousand years.) Thus, I think it would be justifiable to adjust p upward from this lower-bound estimate by a several-order-of-magnitude 'fudge factor', if we were so inclined. But in the spirit of making things hard for longtermism, I will stick with $p = 2 \times 10^{-14}$.[24]

Estimating v_s presents a different puzzle: it is easy to come up with empirically motivated estimates, but different scenarios compatible with the cubic growth model yield *vastly* different estimates of v_s. I will highlight two scenarios in particular.

The 'Space Opera' scenario In this scenario, the settlement of space takes the form of human beings (or broadly human-like organisms) living on Earth-like planets at familiar population densities. In this scenario, we might estimate that the average star can support 300 million people at a time, living lives roughly equivalent to present-day happy lives, with a value of one QALY per year. (The 300 million figure is more than a little arbitrary, and chosen partly for convenience, but is meant to reflect the fact that not all stars have particularly Earth-like planets, and those that do may have planets that are smaller and less hospitable to human or post-human life than Earth. It is worth remembering that the large majority of stars are red dwarfs.) Since our unit of value is 10,000 QALYs, this means that $v_s = 3 \times 10^4$ valons per star per year.

The 'Dyson Spheres' scenario In this scenario, space settlement involves high-efficiency conversion of mass and energy into value-bearing entities. A straightforward version of this scenario involves the construction of Dyson spheres or Matrioshka brains (computers housed in a set of concentric Dyson spheres, meant to convert as much of the star's energy output as possible into computation) around each settled star, which are then used to power simulated

[24]For comparison, Millett and Snyder-Beattie (2017) estimate that the risk of human extinction in the next century from accidental or intentional misuse of biotechnology is between 1.6×10^{-6} and 2×10^{-2}, and that $250 billion in biosecurity spending could reduce this risk by at least 1%. Again assuming that spending on existential risk mitigation has either constant or diminishing marginal returns, and ignoring the difference between the 100 and 1000 year timeframes (which means ignoring both potential benefits of risk reduction in the next century on risk in later centuries, but also the possibility that despite averting an existential catastrophe in the next 100 years, we fail to survive the next 1000 years), this implies $p \geq 6.4 \times 10^{-14}$ (using the lowest estimate of extinction risk from biotechnology), though this could increase to as much as $p \geq 8 \times 10^{-10}$ if we took a higher estimate of *status quo* risk levels. (Note two points: first, if the risk of extinction from biotechnology is much below 1% in the next century, then there are probably other, more pressing existential risks on which our notional philanthropist could more impactfully spend her $1 million. Second, the numbers from Millett and Snyder-Beattie are model-based estimates of objective risk, whereas p is meant to capture a change in the epistemic probability of extinction. Given our uncertainties, the epistemic probability of extinction from biotechnology is likely to be orders of magnitude greater than our lower-bound estimate of the objective risk.)

As another point of comparison, Todd (2017) estimates that $100 billion spent on reducing extinction risk could achieve an *absolute* risk reduction of 1% (e.g., reducing total risk from 4% to 3%). Again assuming constant or diminishing marginal returns and ignoring the difference in timeframes, this implies $p \geq 10^{-7}$. None of these numbers should be taken too seriously, but they indicate the wide range of plausible values for p.

minds with happy (or otherwise valuable) experiences. Bostrom (2003) estimates that in this setup, the average star could support the equivalent of 10^{25} happy human lives at a time—i.e., 10^{25} QALYs per year. This implies $v_s = 10^{20}$ valons per star per year.[25]

For now, I will adopt the more conservative figure, $v_s = 3 \times 10^4$ valons per star per year. In §6, we will consider what happens when we incorporate uncertainty between the Space Opera and Dyson Sphere scenarios in our estimate of v_s.

v_e in principle presents the same puzzles as v_s: the amount of value realized annually on Earth (or, more broadly, in our Solar System) might be many orders of magnitude greater in the future than it is today. But still taking conservatism as our watchword, we will assume that human civilization on Earth simply continues to generate the same level of value it does today, which we can estimate at 6 billion QALYs per year, yielding $v_e = 6 \times 10^5$ valons per year.[26]

The parameters d_g and d_s, which characterize the function n, are more or less known quantities: the Milky Way contains approximately 200 billion stars (and the contribution of nearby dwarf galaxies is trivial in comparison), so d_g (stars per unit volume within 130,000 light years of Earth) is approximately $\frac{2\times10^{11}}{\frac{4}{3}\pi(1.3\times10^5)^3} \approx \frac{2\times10^{11}}{9.2\times10^{15}} \approx 2.2\times10^{-5}$ stars per cubic light year. The Virgo Supercluster contains approximately 200 trillion stars, and has a radius of approximately 55 million light years, which implies $d_s = \frac{2\times10^{15}}{\frac{4}{3}\pi(5.5\times10^7)^3} \approx 2.9 \times 10^{-9}$ stars per cubic light year.

The next parameter is s, the long-run average speed of space settlement. This parameter is reasonably consequential, since it is cubed in the model (by the star density function n; intuitively, if we travel twice as fast, we can reach eight times as many stars by a given time—though this is complicated by the transition from the Milky Way to the wider supercluster). But fortunately its range of plausible values is fairly constrained (assuming our descendants will not find some technological workaround that lets them settle the universe at superluminal speeds). I will adopt the fairly conservative assumption that $s = 0.1c$.[27]

[25]Bostrom's estimate is conservative in a number of ways, relative to the assumptions of the Dyson Sphere scenario. It assumes that we would need to simulate all the computations performed by a human brain (as opposed to, say, just simulating the cerebral cortex, while simulating the rest of the brain and the external environment in a much more coarse-grained way, or simulating minds with a fundamentally different architecture than our own) and that the minds we simulate would have only the same welfare as the average present-day healthy human being. There may also be other ways of converting mass and energy into computation that are orders of magnitude more efficient than Matrioshka brains (Sandberg et al., 2016). But the conservative estimate is enough to illustrate the point.

[26]This estimate sets aside the welfare of non-human animals on Earth, or rather, implicitly assumes that in the far future, the total welfare of non-human animals on Earth will be roughly the same whether or not an intelligent civilization exists on Earth. One could argue for either a net positive or net negative effect of far future human civilization on non-human animal welfare on Earth. (And, particularly conditional on a 'space opera' scenario for space settlement, one could argue for positive or negative adjustments to v_s to account for non-human welfare.) But I set these considerations aside for simplicity.

[27]The main constraint on s appears to be the density of the interstellar medium and the consequent risk of high-energy collisions. In terms of the mass requirements of a probe capable of settling new star systems and the energy needed to accelerate/decelerate that probe, Armstrong and Sandberg (2013) argue convincingly that speeds well above $0.9c$ are achievable. On an in-

t_l (the time at which we begin interstellar settlement, relative to the near future/far future boundary) is hard to estimate on any empirical basis, but fortunately is not terribly consequential. I will choose $t_l = 0$ (implying, on our interpretation of the working example, that we begin settling the stars in the year 3022). Other reasonable guesses would not qualitatively change our results.[28]

This leaves the parameter that is both most consequential and hardest to estimate: r, the combined rate of negative and positive ENEs. What makes r particularly difficult to estimate? In the context of the working example, r is (to a good approximation) the sum of three factors, each of which is individually hard to estimate. First is the rate of negative ENEs, i.e., far future existential catastrophes. There are plausible, though inconclusive, arguments for thinking that this will be quite small (and will decline with time): if we survive the next thousand years, this by itself suggests that the existential threats we face are not extremely severe. And once we begin settling the stars, our dispersion should make us immune from all or nearly all natural catastrophes, and provide at least some defense-in-depth against anthropogenic catastrophes. But while these considerations suggest that the hazard rate for far future human-originating civilization should be small, they don't tell us *how* small—and over the long run, even small hazard rates can be extremely significant. Moreover, as I argued at the end of §4.1, we should not be too sanguine about the assumption of low/declining existential risk in the far future.

The second and third components of r come from positive ENEs. To begin with, there is the possibility that a civilization arising elsewhere will attempt to settle our region of the universe and, if we have disappeared, step in to fill the gap left by our absence. How likely this is per unit time depends on the rate at which intelligent civilizations arise in the sufficiently nearby part of the universe (plus some additional uncertainty about how much of the universe an average interstellar civilization will manage to settle, and how quickly). And this is a matter of extreme uncertainty: according to Sandberg et al. (2018) (who perform a resampling analysis on estimates of the various parameters in the Drake equation from the recent scientific literature), plausible estimates for the rate at which intelligent civilizations arise in the universe span more than 200 orders of magnitude!

Finally, there is the possibility that, if we go extinct, another intelligent species

---

tergalactic scale, such speeds may be feasible *tout court* (Armstrong and Sandberg, 2013, p. 9). But there may be a lower speed limit on *intra*galactic settlement, given the greater density of gas and dust particles. The Breakthrough Starshot initiative aims to launch very small probes toward nearby star systems at $\sim 0.2c$, which appears to be feasible given modest levels of shielding (Hoang et al., 2017). Though larger probes will incur greater risk of collisions, this probably will not greatly reduce achievable velocities, since probes can be designed to minimize cross-sectional area, so that collision risk increases only modestly as a function of mass.

Admittedly, $s = 0.1c$ still seems to be less conservative than the other parameter values I have chosen. It is hard to identify a most-conservative-within-reason value for s, but we could for instance take the speed of Voyager 1, currently leaving the Solar System at $\sim 0.000057c$. But using such a small value for s would make the cubic growth model essentially identical to the steady state model (in which interstellar settlement simply never happens; see §5), except for very small values of r. So a less-than-maximally-conservative value of s is in line with the less-than-maximally-conservative assumption of the cubic growth model itself that interstellar settlement will eventually be feasible.

[28]For instance, if we instead used $t_l = 500$ years, the crucial value of r below which L overtakes N in expected value would only decrease from ~ 0.000135 to ~ 0.000133.

and civilization will arise on Earth to take our place. There is considerably less uncertainty here than with respect to alien civilizations (since we don't need to worry about the early steps on the road to civilization, like abiogenesis). Still, we have very little data on the transition from typical mammalian intelligence to human intelligence (and the most important datum we do have—namely, our own existence—may be contaminated by observer selection effects). In any event, I am not aware of any research that strongly constrains this component of r.

It seems safe to assume that the rate of positive ENEs in the working example (that is, the rate at which 'replacements' for human civilization emerge, from either terrestrial and extraterrestrial sources) is not greater than 10^{-6} per year. If r is significantly greater than 10^{-6} per year, therefore, it will be in virtue of negative ENEs. With respect to negative ENEs in the working example (i.e., exogenous existential catastrophes), it seems plausible their rate will not be greater in the far future than it is today, and that it is today not greater than 10^{-2} per year.[29]

Thus we can venture with reasonable confidence that $r < 10^{-2}$ per year. But r could plausibly be *much* smaller than this, if advanced civilizations are extremely stable and if the evolution of intelligence is sufficiently difficult. I cannot see any clear reason for ruling out values of r small enough to be negligible over the next 10^{14} years (say, $r = 10^{-20}$ per year or less).

Rather than attempting to decide what the value of r should be, therefore, I will simply report the results of the model for a wide range of possible values, and leave it for the reader to decide what parts of this range are most plausible. (In §6, we will consider what happens when we account for our uncertainty about r.)

4.3 Results

Table 1 shows the results of the model for the parameter values specified above, combined with a wide range of values of r. I will mainly leave the discussion of these results for §7, but a few points are worth noting immediately.

First, the headline result is that $\mathrm{EV}(L) > \mathrm{EV}(N)$ iff r is less than ~ 0.000135 (a little over one-in-ten-thousand) per year. This is on the high end of plausible long-term values of r (or so it seems to me), but within the range of reasonable speculation. Thus, our initial conclusion is mixed: the combination of polynomial growth and an exponential discount rate does not automatically sink the case for longtermism, but does leave it open to question.

Second, the last three columns in the table report numbers intended to illustrate the timeframe within which most of the expected value of L is realized. The general purport of these numbers is that, even when L is expectationally superior to N, its impact may be concentrated in a timeframe that is long but not astronomical. For

[29]To my knowledge, the most pessimistic estimate of near-term existential risk in the academic literature belongs to Rees (2003), who gives a 0.5 probability that humanity will not survive the next century. Assuming a constant hazard rate, this implies an annual risk of roughly 6.9×10^{-3}. Sandberg and Bostrom (2008) report an informal survey of 19 participants at a workshop on catastrophic risks in which the highest estimate for the probability of human extinction by the year 2100 was also 0.5 (as compared to a median estimate of 0.19). Other estimates, though more optimistic, generally imply an annual risk of at least 10^{-4}. For a collection of such estimates, see Tonn and Stiefel (2014, pp. 134–5).

r	EV(L)	Integrand peak	99th percentile	Remaining EV = 1
10^{-1}	$\sim 1.2 \times 10^{-7}$	0 years	$\sim$ 46 years	–
10^{-2}	$\sim 1.23 \times 10^{-6}$	0 years	$\sim$ 542 years	–
10^{-3}	$\sim 3.44 \times 10^{-4}$	$\sim$ 2975 years	$\sim$ 9997 years	–
1.35×10^{-4}	~ 1	$\sim$ 22,221 years	$\sim$ 74,407 years	0 years
10^{-4}	~ 3.32	$\sim$ 30,000 years	$\sim$ 100,451 years	$\sim$ 68,412 years
10^{-5}	$\sim 3.32 \times 10^{4}$	$\sim$ 300,000 years	$\sim$ 1 million years	$\sim$ 1.8 million years
10^{-6}	$\sim 4.74 \times 10^{7}$	1.3 million years	$\sim$ 5.6 million years	$\sim$ 20 million years
10^{-7}	$\sim 1.54 \times 10^{9}$	1.3 million years	$\sim$ 83 million years	$\sim$ 296 million years
10^{-8}	$\sim 4.39 \times 10^{12}$	1.3 million years	$\sim$ 1 billion years	$\sim$ 4 billion years
10^{-9}	$\sim 4.37 \times 10^{16}$	$\sim$ 3 billion years	$\sim$ 10 billion years	$\sim$ 49 billion years
10^{-10}	$\sim 4.37 \times 10^{20}$	$\sim$ 30 billion years	$\sim$ 100 billion years	$\sim$ 592 billion years

Table 1: The cubic growth model for $p = 2 \times 10^{-14}$, $t_f = 10^{14}$ years, $v_e = 6 \times 10^5$ valons per year, $v_s = 3 \times 10^4$ valons per star per year, $d_g = 2.2 \times 10^{-5}$ stars per cubic light year, $d_s = 2.9 \times 10^{-9}$ stars per cubic light year, $s = 0.1c$, and $t_l = 0$. 'Integrand peak' gives the time at which the integrand of EV(L) is maximized. '99th percentile' gives the time at which 99% of EV(L) has been realized. 'Remaining EV = 1' gives the time after which the remainder of EV(L) is equal to 1 valon (the EV of the neartermist option in the working example).

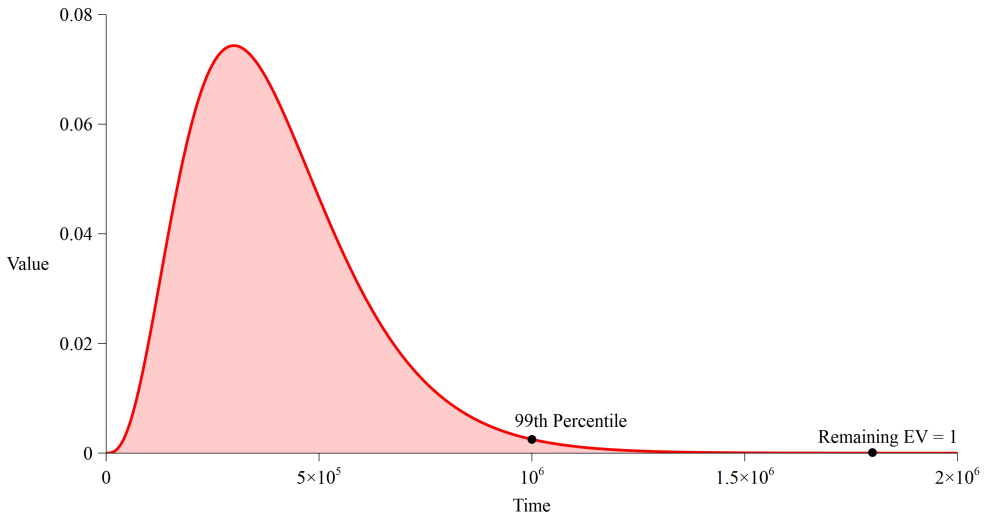

Figure 1: Integrand of EV(L) for $r = 10^{-5}$ (1 ENE per 100,000 years).

instance, Figure 1 illustrates the integrand of of EV(L) for $r = 10^{-5}$, showing that nearly all the expected value of the longtermist intervention is realized within the next 1–2 million years. If the rates of ENEs affecting the most promising persistent-difference strategies are typically of this order of magnitude, then, it would suggest that we should be longtermists, but only on the scale of thousands or millions of years, rather than billions or trillions.

Third, the challenge to longtermism in the cubic growth model comes from a conspiracy of factors, primarily p, v_s, and r, but with r playing in an important sense the greatest role. EV(L) is linear in p and nearly linear in v_s (for small enough values of r). So setting $p = 1$ would raise EV(L) by nearly 14 orders of magnitude, and optimistic-but-reasonable values (e.g., the 10^{-7} implied by Todd (2017)—see fn. 24) could still raise EV(L) by six or seven orders of magnitude, enough to make the case for L over N extremely robust in the model. Replacing the 'space opera' value of v_s with the 'Dyson spheres' value would have a similarly powerful effect (increasing EV(L) by more than 15 orders of magnitude, except when combined with the largest values of r), and more powerful if combined with a commensurate increase in v_e. But, at least in crude quantitative terms, r is even more impactful: even using the conservative values for other parameters adopted above, $r = 0$ would yield EV(L) $\approx 10^{57}$ valons![30] And as shown in Table 1, even the difference between $r = 10^{-2}$ and $r = 10^{-8}$ affects EV(L) by nearly 19 orders of magnitude. Analytically, while EV(L) is linear in p and nearly linear in v_s, it is nearly (inverse) quartic in some ranges of r, so that an order-of-magnitude decrease in r corresponds to a *four*-order-of-magnitude increase in EV(L).[31]

In closing this section, note that the values of some parameters in the cubic growth model (specifically, p, v_e, v_s, r) depend on the specific longtermist intervention we are evaluating, and so are specific to our working example, but the values of other parameters (t_f, d_g, d_s, t_l, s) do not. To apply the model to longtermist interventions other than existential risk mitigation, therefore, would require only repeating the preceding exercise with different values (or ranges of plausible values) for the first set of parameters.

[30]This particular number should not be taken seriously, since when $r = 0$, some of the simplifications in the model become extremely significant—in particular, ignoring cosmic expansion and overestimating star density outside the Virgo Supercluster. The point is simply that even small values of r do a lot to limit EV(L).

[31]More precisely, there are different 'regimes' in the model corresponding to different intervals in the value of r. When r is large, EV(L) is driven primarily by the stream of value of Earth, and so EV(L) grows inversely to r (with an order-of-magnitude decrease in r generating an order-of-magnitude increase in EV(L)). Once r is small enough for the polynomially-increasing value of interstellar settlement to become significant, the relationship becomes inverse quartic. This relationship is interrupted by the transition from the resource-rich Milky Way to the sparse environment of the wider Virgo Supercluster, but resumes once r is small enough that extra-galactic settlement becomes the dominant contributor to EV(L). Finally, for still smaller values of r, the eschatological bound t_f begins to impinge on EV(L), and its growth rate in r slows again (asymptotically to zero, as r goes to zero).

5 The steady state model

The cubic growth model crucially assumes that human-originating civilization will eventually embark on a program of interstellar expansion, and so the potential scale of the future comes not only from its duration but from the astronomical quantity of resources to which our descendants may have access. The supposition that, if we survive long enough, we will have both the capability and the motivation to settle the stars looks like a good bet at the moment.[32] But there are of course formidable barriers to such a project, and any guesses about the motivations and choices of far future agents are speculative at best. Suppose we assume, then, either that interstellar settlement will remain permanently infeasible, or that we will never be motivated to undertake it.

Adopting this hypothesis changes the analysis from the last section in at least three ways. First, of course, we must remove the cubic growth term $v_s n(s(t-t_l))$ from our model. This leaves us with what I will call the *steady state model*, where the value of human-originating civilization at a time is constant as long as we remain in the target state S. Formally, the model is now:

Steady state model

$$\mathrm{EV}(L) = pv_e \int_{t=0}^{t_f} e^{-rt}\, dt$$

Apart from changing the form of the model, the assumption of confinement to our Solar System should lead us to reassess the values of some model parameters. In particular, t_f must be reduced, since if we never leave the Solar System then presumably the lifespan of our civilization will be bounded by the habitability of the Solar System. This suggests a value of t_f between 5×10^8 years (roughly the earliest point when Earth might become uninhabitable due to increasing solar radiation) and 5×10^9 years (when the Sun exits the main sequence). It seems quite implausible that our civilization could survive for 500 million years, but go extinct by neglecting to settle any of the then-more-hospitable environments in the Solar System like Mars or the moons of Jupiter and Saturn. Nevertheless, in the name of conservatism, I will adopt the smaller figure of $t_f = 5 \times 10^8$ years.[33]

Finally, the steady state model presumably supports larger values of r than the cubic growth model, if a civilization confined to a single star system is more vulnerable to existential catastrophes (i.e., in our working example, negative ENEs) than an interstellar civilization. But since I have not tried to estimate r, I will leave this as a qualitative observation rather than trying to quantify its significance.

Table 2 gives the results of the steady state model for a range of values of r, $t_f = 5 \times 10^8$ years, and otherwise the same parameter values as in the last section. On face, these results look very unfavorable for longtermism: $\mathrm{EV}(L)$ exceeds $\mathrm{EV}(N)$

[32] With respect to capability, see for instance Armstrong and Sandberg (2013). With respect to motivation, see for instance Bostrom (2012) on resource acquisition as a convergent instrumental goal of intelligent agents.

[33] Adopting the larger figure would have almost no effect on the values of $\mathrm{EV}(L)$ reported in Table 2 except for the smallest values of r (below 10^{-8}), where it would increase $\mathrm{EV}(L)$ by up to one order of magnitude.

r	EV(L)	99th percentile	Remaining EV = 1
10^{-1}	$\sim 1.2 \times 10^{-7}$	$\sim$ 46 years	–
10^{-2}	$\sim 1.2 \times 10^{-6}$	$\sim$ 461 years	–
10^{-3}	$\sim 1.2 \times 10^{-5}$	$\sim$ 4605 years	–
10^{-4}	$\sim 1.2 \times 10^{-4}$	$\sim$ 46,052 years	–
10^{-5}	$\sim 1.2 \times 10^{-3}$	$\sim$ 460,517 million years	–
10^{-6}	$\sim 1.2 \times 10^{-2}$	$\sim$ 4.6 million years	–
10^{-7}	$\sim 1.2 \times 10^{-1}$	$\sim$ 46 million years	–
1.2×10^{-8}	~ 1	$\sim$ 365 million years	0 years
10^{-8}	$\sim$1.19	$\sim$ 409 million years	$\sim$ 17 million years
10^{-9}	$\sim$4.72	$\sim$ 494 million years	$\sim$ 371 million years
10^{-10}	$\sim$5.85	$\sim$ 495 million years	$\sim$ 412 million years
0	6	495 million years	$\sim$ 417 million years

Table 2: The steady state model for $p = 2\times10^{-14}$, $t_f = 5\times10^8$ years, and $v_e = 6\times10^5$ valons per year. Again, '99th percentile' gives the time at which 99% of EV(L) has been realized, and 'Remaining EV = 1' gives the time after which the remainder of EV(L) is equal to 1 valon (the EV of the neartermist option in the working example).

only when $r <\sim 0.000000012$ per year, which looks like quite a demanding threshold for a single-system civilization at relatively high risk of negative ENEs. It is worth remembering, however, that we have made *very* conservative assumptions about p and, to a lesser extent, v_e. EV(L) scales linearly with both these parameters in the steady state model, so it is easy to see how they affect our conclusions. If we suppose that $p = 10^{-10}$ (meaning, in the working example, that \$1 million spent on existential safety can buy a one-in-ten-billion reduction in the probability of near-future existential catastrophe) and $v_e = 6\times10^7$ valons per year (meaning that, in the far future, human-originating civilization will support 100 times as much value in the Solar System as it does today, through some combination of greater population and greater average welfare), then EV(L) exceeds EV(N) as long as $r <\sim 0.006$ per year. And keeping r below *that* threshold seems entirely realistic, even for a purely planetary civilization.

6 Uncertainty and fanaticism

6.1 Incorporating parameter and model uncertainty

Our conclusions so far look like a mixed bag for longtermism. First, in the cubic growth model, the longtermist intervention is preferred when the long-run rate of ENEs is less than approximately 1.35-in-ten-thousand (0.000135) per year. It is *prima facie* plausible that the true value of r lies below this threshold, but it is hardly obvious. Second, in the steady state model, the required threshold is much smaller: the rate of ENEs must be less than approximately 1.2-in-one-hundred-million per year. And this threshold is *extremely* demanding: the annual probability that another intelligent species evolves on Earth (one source of positive ENEs) plausibly

exceeds this threshold on its own. And on the assumption that humanity remains permanently Earthbound, it requires a lot of optimism to assume that the long-term rate of exogenous existential catastrophes (negative ENEs) will not exceed this threshold as well. So the case for longtermism looks plausible-but-uncertain in the cubic growth model, and extremely precarious in the steady state model.

But in fact, these would be the wrong conclusions to draw. First, of course, we have made very conservative assumptions about the other model parameters, and so the true threshold values of r below which $\mathrm{EV}(L)$ exceeds $\mathrm{EV}(N)$ in each model may be much more generous than the results in the last two sections suggest. But more fundamentally, it is a mistake in the last analysis to think in terms of point estimates for model parameters at all, conservative or otherwise. We are substantially *uncertain* about the values of several key parameters, and that uncertainty is very consequential for the expected value of L. We are also uncertain which model to adopt, and this uncertainty should also be incorporated into our estimate of $\mathrm{EV}(L)$. Once we account for these uncertainties, the picture resolves itself considerably.

The ideal Bayesian approach would be to treat all the model parameters as random variables rather than point estimates, choose a probability distribution that represents our uncertainty about each parameter, and compute $\mathrm{EV}(L)$ on that basis. But for our purposes, this approach has significant drawbacks: $\mathrm{EV}(L)$ would be extremely sensitive to the tails of the distributions for parameters like r, p, and v_s. And specifying full distributions for these parameters—in particular, specifying the size and shape of the tails—would require a great deal of subjective and questionable guesswork, especially since we have nothing like observed, empirical distributions to rely on. Even if we aim to adopt distributions that are conservative (i.e., unfavorable to longtermism), it would be hard to be confident that the tails of our chosen distributions are genuinely as conservative as we intended.

A simpler and more informative approach, rather than inventing full distributions for each parameter, is simply to place conservative constraints on one *point* in the distribution, and see what this tells us. Specifically, we can place constraints on our *confidence levels*: for the parameters about which our uncertainties are most consequential, we can identify values for which we can say: 'Any distribution that didn't assign at least $X\%$ credence to values at least this favorable to longtermism would be overconfident.' This amounts to merely placing an upper bound on one point in the cumulative distribution function for that parameter—a far safer enterprise, epistemically, than specifying a whole distribution. But as we will see, this modest approach is enough to deliver unambiguous qualitative conclusions.

Table 3 describes the results of this exercise. Specifically, I assume that we should assign at least one-in-a-thousand probability to the cubic growth model (i.e., to the hypothesis that our civilization will eventually embark on a long-term program of space settlement, conditional on surviving the next thousand years); that we should assign at least one-in-a-thousand probability to $r \leq 10^{-6}$ ENEs/yr (i.e., to the hypothesis that our civilization will eventually be stable enough that the expected number of extinction or replacement events per year is no more than 10^{-6}, conditional on surviving the next thousand years and on the cubic growth model); that we should assign at least one-in-a-hundred probability to $s \geq 0.8c$ (conditional on surviving the next thousand years and on the cubic growth model); and that we

Uncertainty wrt	**Value(s)**	**Min. Confidence**	**Min. EV(L)**
Cubic Growth (CG)	—	10^{-3}	$\sim 1.23 \times 10^{-5}$ V
r (+ CG)	10^{-6} ENEs/yr	10^{-6} ($10^{-3} \times 10^{-3}$)	~ 47.4 V
s (+ CG)	$0.8c$	10^{-5} ($10^{-2} \times 10^{-3}$)	$\sim 1.38 \times 10^{-5}$ V
v_s (+ CG)	10^{20} (V/yr)/star	10^{-9} ($10^{-6} \times 10^{-3}$)	$\sim 1.11 \times 10^{3}$ V
r, s (+ CG)	(see above)	10^{-8}	~ 48.2 V
r, v_s (+ CG)	(see above)	10^{-12}	$\sim 1.58 \times 10^{11}$ V
s, v_s (+ CG)	(see above)	10^{-11}	$\sim 6.76 \times 10^{3}$ V
r, s, v_s (+ CG)	(see above)	10^{-14}	$\sim 1.62 \times 10^{11}$ V

Table 3: Implications of proposed minimum confidence levels in cubic growth model and in values of r, s, v_s at least as favorable to longtermism as those specified. Each row gives the minimum value of EV(L) implied by minimum confidence levels in the cubic growth model plus specified parameter values, assuming that probabilities are independent and that remaining probability goes to the steady state model, $r = 10^{-3}$ ENEs/yr, and values for other parameters specified in §4.2.

should assign at least one-in-a-million probability to values of v_s at least as great as those suggested by the 'Dyson Spheres' scenario in §4.2 (conditional on surviving the next thousand years and on the cubic growth model). When combined with our point estimates for other parameters, each of these bounds implies a lower bound on EV(L).[34]

Bounding our confidence levels in this way is an unavoidably subjective exercise. Nevertheless, it seems to me that these bounds quite conservative. Given our enormous uncertainty about all aspects of the far future, we should distribute our credence liberally over a wide range of scenarios, and we have no basis for extreme skepticism of scenarios that require only apparently-feasible technologies and intelligible motivations.[35]

Nor can we be extremely confident that future civilization will not enjoy a higher level of existential security than we do today ($r \leq 10^{-6}$).

Taking each source of uncertainty in isolation yields mixed results, as we see in Table 3. Small credences in the cubic growth model and in more optimistic values of s do not by themselves guarantee that EV(L) > EV(N) (given the very conservative assumptions we have made about other parameter values). But small credences in small values of r or in 'Dyson Spheres' values of v_s do have that effect, even when combined with small credence in the cubic growth model itself.

But when we consider uncertainties in combination, the picture is clearer: com-

[34]In the case of v_s, which can take negative values, we must also assume that its expected value conditional on its being less than the 'Dyson sphere' value of 10^{20} (V/yr)/star is non-negative.

[35]See Armstrong and Sandberg (2013) for arguments for the feasibility of interstellar travel at speeds greater than $0.8c$ and of Dyson swarms (vast collections of satellites orbiting a star that capture most or all of its energy output while avoiding the principal engineering challenges of the classic Dyson sphere). Again, if it is technologically feasible for our future civilization to settle the universe at high speed and harness the full energy resources of stars, it seems plausible (though far from certain) that we will chose to do so, since resource acquisition is a 'convergent instrumental goal' for intelligent agents that can serve a vast array of final goals (Bostrom, 2012).

bining the proposed confidence bounds with respect to the cubic growth model plus any two of r, s, and v_s guarantees that $\mathrm{EV}(L) > \mathrm{EV}(N)$ (by at least an order of magnitude). And combining all four confidence bounds guarantees that $\mathrm{EV}(L) >\sim 1.62 \times 10^{11}$ V.[36]

Accounting for uncertainty in our estimates of parameter values (even in the very limited way we have attempted here) will tend to strengthen rather than weaken the case for longtermism, because the *potential* upside of longtermist interventions is so enormous. Hypotheses that tap into that potential can generate astronomical expected value for longtermist interventions, even if the credence we assign those hypotheses is very small.

Uncertainty about r is particularly consequential both because, in general, an order-of-magnitude decrease in r implies a four order-of-magnitude increase in $\mathrm{EV}(L)$ (with the complications described in fn. 31) and because the range of uncertainty with respect to r is very large. For instance, $r = 10^{-8}$ implies $\mathrm{EV}(L) \approx 4.39 \times 10^{12}$ V, so even very small credence in the combination of the cubic growth model with values of r at least this small can suffice to ensure that $\mathrm{EV}(L) > \mathrm{EV}(N)$. And if we think that both the emergence of intelligent civilizations and catastrophes that could destroy an advanced, spacefaring civilization are sufficiently rare, we might assign substantial credence to even smaller values of r.[37]

A final point about the effects of uncertainty: so far, I have simply assumed a total welfarist consequentialist ethical framework. But if we take expectational reasoning to be the correct response to *all* forms of uncertainty, normative as well as empirical, this may be another hypothesis for which a little credence goes a long way. Specifically, if we respond to normative uncertainty by maximizing expected value, and make intertheoretic comparisons (i.e., normalize the value scales of rival normative theories) in any way that looks intuitively plausible in small-scale choice situations, the astronomical quantities of value that aggregative consequentialist theories take to be at stake in the far future are likely to 'swamp' other normative theories in determining the overall expected value of our options. (For a careful exposition of this point in the context of population axiology, see Greaves and Ord

[36]These calculations assume that r, s and v_s are either independent conditional on the cubic growth model, or correlated in such a way that values of one parameter more favorable to longtermism (smaller values of r, larger values of s and v_s) predict more favorable values for the other parameters. It seems natural that there should be at least some of this correlation between 'optimistic' parameter values, which would further increase the expected value of L.

[37]It is worth noting that uncertainty about r makes r effectively time-dependent in the cubic growth and steady state models. What matters in these models is when the *first* ENE occurs, after which the state of the world no longer depends on its state at $t = 0$. This means we are interested, not in the unconditional probability of an ENE occurring at time t, but in the probability that an ENE occurs at t *conditional on no ENE having occurred sooner*. If we know that ENEs come along at a fixed rate, but don't know what that rate is, then this conditional probability decreases with time: conditioning on no ENE having occurred before time t favors hypotheses on which the rate of ENEs is low, more strongly for larger values of t. This is just another way of understanding the fact that, when we are unsure what discount rate to apply to a stream of value, the discount *factor* at later times will converge with that implied by the lowest possible discount rate.

An interesting analytical result is that, when a value stream is subject to an uncertain exponential discount rate, with a continuous probability distribution over possible rates supported at least on the interval $[0, k]$ for some $k \in (0, 1]$, the schedule of expected discount factors is asymptotically hyperbolic—that is, approximates hyperbolic discounting in the limit (Azfar, 1999).

(2017).) If we fully embrace this sort of reasoning, we might find that longtermist conclusions are 'robust' to objections from axiology, ethics, and normative theory in general, since even a very small credence in a normative theory like total utilitarianism is enough to secure the case for longtermism.[38]

6.2 Fanaticism

By the rules of the expected value game, the case for longtermism appears to survive the version of the epistemic challenge with which we confronted it. But it has prevailed in a way that should make us slightly uneasy: by appealing to potentially-minuscule probabilities of astronomical quantities of value.

Many people suspect that expected value reasoning goes wrong, or at least demands too much of us, in situations involving these 'Pascalian' probabilities. (See for instance Bostrom (2009), Monton (2019), Russell (2021).) But it has so far proven difficult to say anything precise or constructive about these worries. For that reason, I will limit myself to a few brief and imprecise observations.

'Pascalian' choice situations are those in which the choice set selected by risk-neutral expectational reasoning is determined by minuscule probabilities of extreme positive or negative outcomes. A natural way to measure the Pascalian-ness of a choice situation, then, is to ask how easily we can change the choice set of expectationally best options by *ignoring* these extreme possibilities. That is, we arrange the possible payoffs of each option from worst to best, snip the left and right tails of each prospect (removing the worst-case scenarios up to some probability $\mu \in (0, .5)$ and likewise the best-case scenarios up to probability μ), then compute the expectations of these truncated prospects. We then look for the minimum value of μ by which we would have to truncate the tails of each prospect in order to change the choice set.[39] Designating this minimum value μ^*, we can then measure the 'Pascalian-ness'

[38]It is controversial, however, whether we should reason expectationally in response to normative uncertainty, even given that this is the right response to empirical uncertainty. For defense of broadly expectational approaches to normative uncertainty, see Lockhart (2000), Sepielli (2009), and MacAskill and Ord (2020), among others. For rival views, see Nissan-Rozen (2012), Gustafsson and Torpman (2014), Weatherson (2014), and Harman (2015), among others.

This debate may also be relevant in deciding how to weigh outré possibilities like the Dyson spheres scenario that involve large numbers of non-human-like minds. (Thanks to Hilary Greaves for this point.) If we are uncertain whether or to what degree the 'artificial' or 'simulated' minds that might exist in a Matrioshka brain are morally statused, should we simply discount their putative interests by the probability that those interests carry moral weight? Arguably, our uncertainty here is a kind of 'quasi-empirical' uncertainty: we simply don't know whether these minds would have the sort of subjective experiences we care about. But it may also seem more akin to moral uncertainty, and we may therefore feel reluctant to simply go by expected value.

[39]We can make this precise in the framework of *risk-weighted* expected utility theory (Quiggin, 1982; Buchak, 2013), with a risk function of the form:

$$r(x) = \begin{cases} 0 & 0 \leq x \leq \mu \\ \frac{x-0.5}{1-2\mu} + 0.5 & \mu \leq x \leq 1-\mu \\ 1 & 1-\mu \leq x \leq 1 \end{cases}$$

We then choose the option that maximizes $u_1 + \sum_{i=2}^{n} r(Pr(u \geq u_{i+1}))(u_{i+1} - u_i)$, where the possible payoffs $u_1, ...u_n$ are ordered from worst to best. A similar sort of truncation is suggested by Buchak as a response to the St. Petersburg game (Buchak, 2013, pp. 73–74).

of a choice situation on the unit interval by the formula $1 - 2\mu^*$.[40]

By this measure, the preceding analysis suggests that choices between longtermist and neartermist interventions *could* be extremely Pascalian. We have found that longtermist interventions can have much greater expected value than their neartermist rivals even when the probability of having any impact at all on the far future is minuscule (2×10^{-14}, for a fairly large investment of resources) and when, conditional on having an impact, most of the expected value of the longtermist intervention is conditioned on further low-probability assumptions (e.g., large-scale interstellar settlement, astronomical values of v_s, large values of s, and—in particular—small values of r). It could well turn out that the vast majority of the expected value of a typical longtermist intervention—and, more importantly, the component of its expected value that gives it the advantage over neartermist alternatives—depends on a conjunction of improbable assumptions with joint probability on the order of (say) 10^{-18} or less.

On the other hand, there is tremendous room for reasonable disagreement about the relevant probabilities. If you think that, in the working example, p is on the order of (say) 10^{-7}, and that the assumptions of eventual interstellar settlement, astronomical values of v_s, large values of s, and very small values of r are each more likely than not, then the amount of tail probability we would have to ignore to prefer N might be much greater—say, 10^{-8} or more.

These numbers should not be taken too literally—they are much less robust, I think, than the expected value estimates themselves, and at any rate, it's not yet clear whether we should care that a choice situation is Pascalian in the sense defined above, or if so, at what threshold of Pascalian-ness we should begin to doubt the conclusions of expected value reasoning. So the remarks in this section are merely suggestive. But it seems to me there are reasonable grounds to worry that the case for longtermism is problematically dependent on a willingness to take expectational reasoning to fanatical extremes.[41]

[40] This measure is imperfect in that it will classify as highly Pascalian some choice situations that are not intuitively Pascalian, but where two or more options are just very nearly tied for best. But the measure is only intended as a rough heuristic, not as something that should play any role in our normative decision theory.

[41] In Tarsney (2020), I set out a view that is meant (among other things) to give a principled and intuitively attractive response to the problem of 'Pascalian fanaticism' discussed in this section. The essence of the view is (i) first-order stochastic dominance as a necessary and sufficient criterion of rational choice combined with (ii) recognition of the enormous 'background uncertainty' about the choiceworthiness of our options that results from attaching normative weight to aggregative consequentialist considerations and our uncertainty about the amount of value in the world independent of our choices. Simplifying considerably: under levels of background uncertainty that seem warranted at least for total utilitarians (and in particular, assuming that the probability distribution representing the agent's background uncertainty has exponential-or-heavier tails), the decision-theoretically modest requirement to reject stochastically dominated options implies that agents are generally required to choose options whose 'local' consequences maximize expected objective value when the decision-relevant probabilities are intermediate, but are often free *not* to maximize expected objective value when the balance of expectations is determined by minuscule probabilities of astronomical positive or negative payoffs.

The line between 'intermediate' and 'minuscule' probabilities depends on the scale of the agent's background uncertainty and other features of the choice situation. But consider a simplified case where you can choose between a 'sure thing' option that yields a modest payoff s for certain, and a

7 Drawing conclusions

The preceding investigation suggests several broad conclusions:

1. If we accept expectational utilitarianism, and therefore do not mind premising our choices on minuscule probabilities of astronomical payoffs, then the case for longtermism (specifically, for the persistent-difference strategy of existential risk mitigation) seems robust to the epistemic challenge we have considered (namely, epistemic persistence skepticism). While there are plausible point estimates of the relevant model parameters that favor neartermism, once we account for uncertainty, it takes only a very small credence in combinations of parameter values more favorable to longtermism for EV(L) to exceed EV(N) in our working example.

2. There are, however, *prima facie* plausible worldviews on which this conclusion depends very heavily on minuscule probabilities of astronomical payoffs. To the extent that we are wary of simply maximizing expected value in the face of such Pascalian probabilities, we are left with a residual decision-theoretic worry about the case for longtermism.

3. More concretely, the case for longtermism may depend to a significant extent on the possibility of interstellar settlement: it is significantly harder (though not impossible) to make the case for persistent-difference interventions entirely within the steady state model.

'long shot' option that yields an astronomical positive payoff a with very small probability p, and nothing otherwise, where $pa > s$. Here, to a good approximation under reasonable assumptions, the long shot option will be stochastically dominant just in case $p > \frac{s}{\text{IQR}}$, where IQR is the interquartile range of your background uncertainty. (See §5.4 of Tarsney (2020). This ratio is a good approximation of the threshold for stochastic dominance under background uncertainty $a \gg \text{IQR}$. When $\text{IQR} \gg a$, the long shot option will usually be stochastically dominant as long as it has greater expected value.) Simplifying our working example considerably, we might represent it as a choice between a sure payoff of 3000 QALYS (1 valon) and probability p of a definite astronomical payoff from preventing existential catastrophe. Suppose we adopt our lower-bound estimate of $p = 2 \times 10^{-14}$, and assume that the value of preventing existential catastrophe is large enough to make the long shot option expectationally superior (i.e., greater than 5×10^{13} V or 1.5×10^{17} QALYs)—though of course, as we have seen, the expectational superiority of the long shot option in this case is far from a given. Under those assumptions, for the longtermist long shot option to stochastically dominate the neartermist sure thing would require background uncertainty with IQR greater than $\sim 5 \times 10^{13}$ V/1.5×10^{17} QALYs. Although we should be very uncertain about the amount of pre-existing value in the world, it is not obvious that our uncertainty should be *this* great.

So, if the stochastic dominance approach is correct, it seems that choice situations like our working example are difficult, borderline cases (at least from a utilitarian point of view). It could turn out, on further analysis, that the utilitarian case for choosing the longtermist option is on very firm decision-theoretic footing (requiring no decision-theoretic assumptions beyond first-order stochastic dominance). But it could also turn out that, even though the longtermist option is expected value-maximizing, it is nevertheless rationally optional. Resolving this question would require much more precise estimates both of the various decision-relevant probabilities (like p) and of the probability distribution that describes a utilitarian agent's rationally warranted background uncertainty about the amount of value in the universe.

4. The potentially enormous impact that the long-term rate of ENEs has on the expected value of longtermist interventions has implications for 'intra-longtermist' prioritization: we have strong *pro tanto* reason to focus on bringing about states such that both they and their complements are highly persistent, since it is these interventions whose effects are likely to persist for a very long time (and thus to affect our civilization when it is more widespread and resource-rich). This suggests, in particular, that interventions focused on reducing existential risk may have higher expected value than, say, interventions aimed at reforming institutions or changing social values: intuitively, the intended effects of these interventions are relatively easy to undo, or to achieve at some later date even if we fail to achieve them now. So the long-term rate of ENEs (i.e., value of r) may be significantly higher for these interventions than for existential risk mitigation.

5. Finally, there is some reason to think that, while the longtermist conclusion is ultimately correct, we should be 'longtermists' on the scale of thousands or millions or years, rather than billions or trillions of years. The case for this conclusion is far from conclusive: if you assign substantial probability to *very* high levels of persistence for some longtermist interventions (say, $r < 10^{-10}$ per year), then you will have substantial reason to care about the future billions of years from now. And it is certainly conceivable that far-future civilization might be so stable that these values are appropriate. But it is clearly an open question just how stable we should expect far-future civilizations to be, and the answer to this question makes a big difference to how we should distribute our concern over time.

On the whole, my sense is that the version of the epistemic challenge we have considered in this paper is serious and, in the last analysis, probably has significant practical implications for optimal utilitarian resource allocation, but is not fatal to the longtermist thesis. But the models and results in this paper are at best a first approximation, and much more work is needed to reach that last analysis.

First, of course, there is plenty of room to improve and generalize the quantitative analysis in §§4–6. This might include: (i) building out the relatively simple cubic growth and steady state models (e.g., incorporating plausible forms of time-dependence in the rate of ENEs, or incorporating cosmic expansion to improve the accuracy of the cubic growth model on very long timescales); (ii) more rigorous estimates of the values of the various model parameters; (iii) a more systematic sensitivity analysis of the case for existential risk mitigation under model and parameter uncertainty; and (iv) application of the steady state and cubic growth models, or improved versions thereof, to a wider range of persistent-difference interventions, to assess how far the scope of longtermism extends beyond existential risk mitigation.

But second, while I have focused on what I take to be the most promising category of longtermist intervention (persistent-difference interventions) and the strongest version of the epistemic challenge that can be mounted against them (epistemic persistence skepticism), there are other aspects of the epistemic challenge that deserve investigation as well. In particular, returning to the four factors identified in §2, the epistemic skeptic might cast doubt on the importance or tractability of typ-

ical longtermist objectives, as well as their persistence or complement persistence. As noted in §3, the cubic growth and steady state models are focused on modelling the effects of our assumptions about persistence and complement persistence, not on tractability or importance. Tractability corresponds to the parameter p, and in the context of the working example, I gave only a minimal argument (in fn. 23) for what I took to be a conservative lower-bound value of $p = 2 \times 10^{14}$. While this value strikes me as very safely conservative, it would not be impossible to argue for smaller values—e.g., on the grounds that the world is 'chaotic' in such a way that we have virtually no ability to predict the effects of our interventions even on a scale of decades or centuries or, alternatively, on the grounds that the world is so deterministic that we are already locked into a long-term trajectory that is nearly impossible to change, even temporarily.

With respect to importance, I think there is a more compelling worry to be raised: perhaps the characteristic point of failure for persistent-difference strategies is not that their objectives (like the survival of human-originating civilization) are not persistent, but rather than those objectives are not persistently *good* (or more precisely, persistently better than their complements). For instance, perhaps the existence of large interstellar civilizations is highly persistent (and complement-persistent), but the features that would give such a civilization positive rather than negative value are not. If civilizations tend to cycle predictably, or 'wander' unpredictably, between high-value and low-value states (e.g., between good and bad political institutions, economic systems, or moral norms), it could be that despite their astronomical *potential* for value, the *expected* value of ensuring the existence of a large interstellar civilization is close to zero. In that case, we can have persistent effects on the far future, but not effects that *matter* (in expectation). A *prima facie* reason to doubt this story is that it seems to require some implausible symmetries: for instance, for the expected value of the continued existence of our civilization to be extremely close to zero, the positive and negative components of this expectation would have to be almost exactly equal, which seems like an improbable coincidence. Nevertheless, this worry strikes me as deserving investigation.

Finally, skeptics of longtermism might reject the simple probabilist epistemic framework that I have assumed throughout this paper. In particular, epistemic worries about longtermism might be usefully formulated in frameworks involving *imprecise probabilities* (focusing on the imprecision of epistemic probabilities concerning the far future and long-term consequences)[42], *unawareness* (focusing on our inability to conceive or assign probabilities to, for instance, fundamentally new ideas or technologies that may shape the far future), or *bounded rationality* (focusing on our limited capacity to represent and reason about the vast combinatorial space of ways the far future might go). There are also entirely non-quantitative approaches to ethical decision-making under uncertainty (based on full belief, or on non-numerical gradations of uncertainty).How the case for longtermism fares in these alternative frameworks can only be assessed one framework at a time. But insofar as, in the precise probabilist framework, the case for longtermism relies to a significant extent on expectational reasoning about very small probabilities of very large payoffs,

[42]For discussion of a potential challenge to longtermism that arises in the context of imprecise probabilist epistemology and decision theory, see Mogensen (2021).

there is some *prima facie* reason to suspect that things may be more difficult in other frameworks that are less welcoming to this kind of reasoning.

Longtermism, if true, is of enormous and revolutionary practical importance. It therefore deserves careful scrutiny. I hope to have shown, on the one hand, that even within the most hospitable normative frameworks (like expectational utilitarianism) the case for longtermism is not trivial, but on the other hand, that it has reasonable prospects of surviving an important and under-explored challenge.

Appendix: Simplifications in the cubic growth model

In this appendix, I catalog some of the many simplifications involved in the cubic growth model, in the way in which I applied that model to our working example, and in the approach of hedging between the steady state and cubic growth models suggested in §6. I briefly explain why, in my view, each of these simplifications is tolerable for our purposes (namely, for comparing longtermist and neartermist interventions with enough quantitative accuracy to draw broad qualitative conclusions about the case for longtermism). But the list is also meant as a 'wish list' of ways in which more complex expected value models for longtermist interventions might improve on the relatively simple models I have developed in this paper. I have tried to list the simplifications in descending order of importance.

Simplification: The analysis in §6 makes no attempt to comprehensively account for model uncertainty—it considers only two models from an infinite set of possible models and a probably-very-large set of plausible models.

Rationale: (1) There's no good way (that I can think of) to randomly sample or average over the set of all possible/plausible models. (2) Including other models less optimistic than the cubic growth model is unlikely to change our quantitative results by more than 1–2 orders of magnitude, and so unlikely to affect our qualitative conclusions, as long we still assign at least ~ 0.01 probability to models at least as optimistic as the cubic growth model. Including *more* optimistic models (e.g., with indefinite exponential growth) is only likely to strengthen our qualitative conclusions, by making the expectational case for longtermism even more robust under uncertainty but also exacerbating worries about Pascalian fanaticism (assuming we assign these greater-than-cubic models low probability).

Simplification: The rate of ENEs, r, is treated as time-independent.

Rationale: (1) The main argument against time-independence is the hypothesis that anthropogenic extinction risk will decline as we settle more of the universe, which is plausible but non-obvious (see discussion in the final paragraph of 4.1). (2) There's no clear empirical basis for modeling the time-dependence of r, so the assumption of constant r is licensed by the principle of defaulting to a simpler model when additional complexity would require subjective and poorly-motivated guesswork. (3) This assumption is justified by the practice of making assumptions that are conservative with respect to the case for longtermism, since including time-dependence is likely to favor L over N.

Simplification: The longtermist intervention L has only a single effect, namely, increasing the probability that the world is in state S rather than $\neg S$ in the far future.

Rationale: (1) Accounting for secondary objectives seems unlikely to change our quantitative results by more than 1–2 orders of magnitude (though this is far from obvious), and so unlikely to affect our qualitative conclusions. (2) There's no clear empirical basis for modeling interactions between multiple long-term effects of a longtermist intervention.

Simplification: Neither the cubic growth model itself nor the estimate of $\mathrm{EV}(N)$ that we adopted to analyze the working example make any attempt to model long-term/'flow-through' effects from the neartermist intervention N.

Rationale: (1) There's no clear empirical basis for modeling these effects. (2) This simplification is arguably justified by the aim of assessing the longtermist thesis rather than assessing particular interventions: if the long-term indirect or flow-through effects of apparently 'neartermist' interventions give them greater expected value than apparently 'longtermist' interventions, this doesn't refute longtermism but just tells us which interventions are best from a longtermist perspective.

Simplification: Welfare per person/per settled star is assumed to be constant in the far future.

Rationale: Dropping this simplification seems unlikely to change our quantitative results by more than 1–2 orders of magnitude (though this is far from obvious), and so unlikely to affect our qualitative conclusions.

Simplification: The model ignores effects on the welfare of beings other than *Homo sapiens* and our 'descendants'.

Rationale: (1) The sign and magnitude of the effects of paradigmatic longtermist interventions on the welfare of non-human animals (or their far-future counterparts) are very unclear. (2) Dropping this simplification seems unlikely to change our quantitative results by more than 1–2 orders of magnitude (though this is far from obvious), and so unlikely to affect our qualitative conclusions.

Simplification: The speed of interstellar settlement, s, is treated as constant (ignoring, for instance, the possibility of higher speeds for intergalactic rather than intragalactic settlement).

Rationale: (1) The significance of these effects is unclear. (2) This assumption is justified by the practice of making assumptions that are conservative with respect to the case for longtermism, provided we choose a value of s that is conservative for all phases of space settlement.

Simplification: The model assumes that the effect of L, if any, happens at $t = 0$ (i.e., it ignores the potential value of L putting/keeping the world in state S at times before $t = 0$).

Rationale: (1) Dropping this simplification is unlikely to change our quantitative results by more than 1–2 orders of magnitude (except when combined with very large values of r), and so unlikely to affect our qualitative conclusions. (2) This

simplification is arguably justified by the aim of assessing the longtermist thesis rather than particular interventions: if the near-term effects of apparently 'longtermist' interventions give them greater expected value than paradigmatic neartermist interventions, this is at best a limited vindication of longtermism.

Simplification: The model uses a crude star density function and, more generally, a crude approximation of the growth in our resource endowment with spatial expansion.

Rationale: Dropping this simplification is unlikely to change our quantitative results by more than 1–2 orders of magnitude, and so unlikely to affect our qualitative conclusions.

Simplification: The model does not include any 'ramp-up period' in value generation after settling new star systems—it implicitly assumes that each star system begins producing value at its 'mature' level immediately upon settlement.

Rationale: Accounting for ramp-up periods is unlikely to change our quantitative results by more than 1–2 orders of magnitude, and so unlikely to affect our qualitative conclusions.[43]

Simplification: The model ignores various physical/astrophysical considerations that are significant on very long timescales: cosmic expansion, change in the number/composition/energy output of stars, increasing entropy, proton decay...

Rationale: These considerations become (extremely) significant on very long timescales, and hence for intra-longtermist comparisons, but (given other assumptions of the model) they do not have a significant effect on the comparison between longtermist and neartermist interventions.

Simplification: The eschatological bound t_f is treated as a hard (i.e., instantaneous) cutoff.

Rationale: The details of physical eschatology become significant on very long timescales, and hence for intra-longtermist comparisons, but (given other assumptions of the model) they do not have a significant effect on the comparison between longtermist and neartermist interventions.

References

Adams, F. C. and G. Laughlin (1997). A dying universe: the long-term fate and evolution of astrophysical objects. *Reviews of Modern Physics 69*(2), 337.

Armstrong, S. and A. Sandberg (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica 89*, 1–13.

Azfar, O. (1999). Rationalizing hyperbolic discounting. *Journal of Economic Behavior & Organization 38*(2), 245–252.

[43] We can conservatively account for this consideration by simply choosing a larger value of t_l, representing the time at which we embark on interstellar settlement *plus* the time it takes to get a new settlement up and running. (Thanks to Tomi Francis for this point.) And as we saw in §4.2, modest increases in t_l make little difference to our quantitative or qualitative conclusions.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. Ph. D. thesis, Rutgers University Graduate School - New Brunswick.

Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves and T. Pummer (Eds.), *Effective Altruism: Philosophical Issues*, pp. 80–98. Oxford: Oxford University Press.

Beckstead, N. and T. Thomas (2020). A paradox for tiny probabilities and enormous values. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 10-2020.

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

Bostrom, N. (2009). Pascal's mugging. *Analysis 69*(3), 443–445.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines 22*(2), 71–85.

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

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

Burch-Brown, J. M. (2014). Clues for consequentialists. *Utilitas 26*(1), 105–119.

Cowen, T. (2018). *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals*. San Francisco: Stripe Press.

Fye, S. R., S. M. Charbonneau, J. W. Hay, and C. A. Mullins (2013). An examination of factors affecting accuracy in technology forecasts. *Technological Forecasting and Social Change 80*(6), 1222–1231.

GiveWell (2021, November). Our top charities. URL: `https://www.givewell.org/charities/top-charities`. Accessed 31 January 2022.

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society 116*(3), 311–339.

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

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

Gustafsson, J. E. and O. Torpman (2014). In defence of My Favourite Theory. *Pacific Philosophical Quarterly 95*(2), 159–174.

Hare, C. (2011). Obligation and regret when there is no fact of the matter about what would have happened if you had not done what you did. *Noûs 45*(1), 190–206.

Harman, E. (2015). The irrelevance of moral uncertainty. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 10. Oxford: Oxford University Press.

Hoang, T., A. Lazarian, B. Burkhart, and A. Loeb (2017). The interaction of relativistic spacecrafts with the interstellar medium. *The Astrophysical Journal 837*(5), 1–16.

Johnson, J. (2019). Good at doing good: Effective altruism ft. Robert Wiblin. The Neoliberal Podcast.

Langacker, P. (1981). Grand unified theories and proton decay. *Physics Reports 72*(4), 185–385.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy and Public Affairs 29*(4), 342–370.

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences*. New York: Oxford University Press.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences 20*(2), 130–141.

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

Makridakis, S. and M. Hibon (1979). Accuracy of forecasting: An empirical investigation. *Journal of the Royal Statistical Society: Series A (General) 142*(2), 97–145.

Martin, T., J. M. Hofman, A. Sharma, A. Anderson, and D. J. Watts (2016, Feb). Exploring limits to prediction in complex social systems. *arXiv e-prints*, arXiv:1602.01013.

Matthews, D. (2015). I spent a weekend at Google talking with nerds about charity. I came away ... worried. Vox. URL: `https://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai`. Published 10 August 2015. Accessed 28 September 2019.

Millett, P. and A. Snyder-Beattie (2017). Existential risk and cost-effective biosecurity. *Health Security 15*(4), 373–383.

Mogensen, A. L. (2021). Maximal cluelessness. *Philosophical Quarterly 71*(1), 141–162.

Monton, B. (2019). How to avoid maximizing expected utility. *Philosophers' Imprint 19*(18), 1–25.

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

Muehlhauser, L. (2019). How feasible is long-range forecasting? The Open Philanthropy Blog. URL: https://www.openphilanthropy.org/blog/how-feasible-long-range-forecasting. Published 10 October 2019. Accessed 23 October 2019.

Mullins, C. A. (2018). Retrospective analysis of long-term forecasts. Technical report, Bryce Space and Technology.

Ng, Y.-K. (2016). The importance of global extinction in climate change policy. *Global Policy 7*(3), 315–322.

Nissan-Rozen, I. (2012). Doing the best one can: A new justification for the use of lotteries. *Erasmus Journal for Philosophy and Economics 5*(1), 45–72.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury Publishing.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review 94*(2), 251–267.

Portmore, D. (2016). Uncertainty, Indeterminacy, and Agent-Centered Constraints. *Australasian Journal of Philosophy*, 1–15.

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

Rees, M. (2003). *Our Final Century: Will the Human Race Survive the Twenty-first Century?* London: William Heinemann Ltd.

Russell, J. S. (2021). On two arguments for fanaticism. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 17-2021.

Sagan, C. (1994). *Pale Blue Dot: A Vision of the Human Future in Space* (1st ed.). New York: Random House.

Sandberg, A., S. Armstrong, and M. Ćirković (2016). That is not dead which eternal lie: The aestivation hypothesis for resolving Fermi's paradox. *Journal of the British Interplanetary Society 69*, 405–415.

Sandberg, A. and N. Bostrom (2008). Global catastrophic risks survey. Technical Report 2008-1, Future of Humanity Institute, Oxford University.

Sandberg, A., E. Drexler, and T. Ord (2018, Jun). Dissolving the Fermi Paradox. *arXiv e-prints*, arXiv:1806.02404.

Schuster, H. G. and W. Just (2006). *Deterministic Chaos: An Introduction* (4th ed.). Weinheim: Wiley-VCH.

Schwitzgebel, E. (2022). Against longtermism. The Splintered Mind. URL: https://schwitzsplinters.blogspot.com/2022/01/against-longtermism.html. Published 5 January 2022. Accessed 10 January 2022.

Sepielli, A. (2009). What to do when you don't know what to do. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Volume 4, pp. 5–28. Oxford: Oxford University Press.

Sittler, T. M. The expected value of the long-term future. Unpublished manuscript, January 2018.

Tarsney, C. (2020). Exceeding expectations: Stochastic dominance as a general decision theory. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 3-2020.

Tetlock, P. and D. Gardner (2015). *Superforecasting: The Art and Science of Prediction*. New York: Crown Publishers.

Tetlock, P. E. (2005). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.

Thorstad, D. (2022). Existential risk pessimism and the time of perils. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 1-2022.

Todd, B. (2017). The case for reducing extinction risk. 80,000 Hours. URL: `https://80000hours.org/articles/extinction-risk/`. Accessed 14 December 2019.

Tonn, B. and D. Stiefel (2014). Human extinction risk and uncertainty: Assessing conditions for action. *Futures 63*, 134–144.

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

Wilkinson, H. (2022). In defense of fanaticism. *Ethics 132*(2), 445–477.