The Epistemic Challenge to Longtermism

Christian J. Tarsney*

Version 3, May 2020
(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 options is mainly determined by short-term considerations. This paper aims to precisify and evaluate (a version of) this epistemic objection to longtermism. To that end, I develop two simple models for comparing "longtermist" and "short-termist" interventions, incorporating the idea that, as we look further into the future, the effects of any present intervention become progressively harder to predict. 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 (2019) and Ord (2020), let's call this view *longtermism*. The longtermist thesis represents a radical departure from conventional thinking about how to make

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

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]

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 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]

[1]Proponents of longtermism include Bostrom (2003, 2013) (who focuses on the long-term value of reducing existential risks to human civilization), Beckstead (2013a, 2019) (who gives a general 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 (2019) (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 (2019)), and are frequently raised in conversation, but have not yet been extensively explored. For expressions of epistemically-motivated skepticism toward longtermism in non-academic sources, see for instance Matthews (2015) and Johnson (2019).

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

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 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 "chaotic" 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

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.

[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 $\geq$ 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, which remaining significantly *above* chance—for discussion, see §1.7 of Muehlhauser (2019).

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. Specifically, I will identify a version of the challenge that seems especially compelling, precisify that version of the challenge by means of two simple models, and assess the strength of the challenge by examining the results of those models for various plausible combinations of parameter values.

Since my goal is to assess whether the case for longtermism is robust to a particular kind of objection, I will make some assumptions meant to screen off other objections. In particular, I assume: (i) a total welfarist consequentialist normative framework (the *prima facie* most favorable setting for longtermism), setting aside axiological and ethical challenges to longtermism that are mostly orthogonal to the epistemic challenge[7]; (ii) a precise probabilist epistemic framework (i.e., that the rational response to uncertainty involves assigning precise probabilities to the possibilities over which one is uncertain), setting aside for instance the imprecise probabilist worries discussed in Mogensen (forthcoming); and (iii) the decision-theoretic framework of expected value maximization, setting aside worries arising from risk aversion or from "anti-fanaticism" considerations of the sort discussed in chapters 6–7 of Beckstead (2013a) (though we will take up the issue of fanaticism in §6.2).

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.

The paper proceeds as follows: In §2, I attempt to state the longtermist thesis more precisely. In §3, I similarly attempt to precisify the epistemic challenge, and identify the version of that challenge on which I will focus. In §4, I describe the first model for comparing longtermist and short-termist 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, growth in the potential value of human-originating civilization is a cubic function,

[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.

[7]For discussion of these axiological and ethical challenges, see Beckstead (2013a) and Greaves and MacAskill (2019).

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 effect of higher-level uncertainties—both uncertainty about key parameter values and uncertainty between the two models. §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 Sharpening the longtermist thesis

The longtermist thesis is challenging to state precisely.[8] But for our purposes, I will adopt the following admittedly imperfect characterization:

Longtermism In most choice situations (or at least, most of the most important choice situations) faced by present-day human agents, what one ought to do all things considered is mainly determined by the possible effects of one's options on the far future.

This statement needs a few immediate notes and clarifications. First, it clearly inherits the vagueness of terms like "most", "the most important choice situations", "present-day", "mainly", and "far future". For the most part, I will not try to say how these terms should be precisified.[9]

Second, by "ought to do all things considered", I mean to indicate the "action-guiding" sense of *ought* that tells an agent what to do, full stop, in light of her epistemic and doxastic situation. What one ought to do in this sense is to choose one of the available options that an agent in the same situation with one's own capacities, but whose deliberations are practically rational and otherwise normatively in good order relative to those capacities, *might* choose.[10] It is convenient, in stating the longtermist thesis, to understand this *ought* as "evidence-relative"/"prospective" rather than "belief-relative"/"subjective", but I don't mean to take any position in the debate between subjectivists and prospectivists about *ought*. If one is inclined to understand the action-guiding *ought* as subjective, then the longtermist thesis should simply be understood as applying only to agents whose beliefs are relevantly well-aligned with their evidence.

[8] For extended discussion, see Greaves and MacAskill (2019).

[9] The exception is "far future", which I will sometimes 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.

[10] Cf. the "central ought" in Broome (2013), which I take to indicate the same concept.

Third, what does it mean to say that what one ought to do is "mainly determined by" effects on the far future? The best I can do here is to illustrate with an example: If one always ought to do what maximizes expected welfare, then to say that what one ought to do is mainly determined by effects on the far future means something like: For most pairs of options O_j and O_k, the difference between O_j and O_k in expected near-future welfare is less than the difference between O_j and O_k in expected far-future welfare. Longtermism, however, cannot be identified with the thesis that *differences in expected welfare* lie mainly in the far future, since it might be that what we ought to do is not always the thing that maximizes expected welfare. How the expression "mainly determined by" should be interpreted in the context of non-consequentialist normative theories and non-expectational decision theories, I will leave for the reader to infer by analogy.

3 Sharpening the epistemic challenge

The epistemic challenge to longtermism emphasizes the difficulty of predicting the far future. But to understand the challenge, we must specify more precisely the kind of predictions we're interested in. After all, *some* predictions about the far future are relatively easy. For instance, I can confidently predict that, a billion years from now, the observable universe will contain more than 100 and fewer than 10^{100} stars. (And this prediction is quite precise, since $(100, 10^{100})$ comprises only an infinitesimal fraction of the natural numbers!)

But our ability to make predictions like these doesn't have much bearing on the case for longtermism. For roughly the same reason that it is relatively easy to predict, the number of stars in the observable universe is very difficult to *affect*. And what we need, for practical purposes, is the ability to *predictably affect* the world by doing one thing rather than another. That is, we need the ability to make *practical predictions*—predictions that, if I choose O_j, the world will be different in some particular way than it would have been if I had chosen O_k.

Even long-term practical predictions are sometimes easy. For instance, if I shine a laser pointer into the sky, I can predict with reasonable confidence that a billion years from now, some photons will be whizzing in a certain direction through a certain region of very distant space, that would not have been there if I had pointed the laser pointer in a different direction. I can even predict what the wavelength of those photos will be, and that it would have been different if I had used my green instead of my red laser pointer.

But our ability to make predictions like these isn't terribly heartening either,

since photons whizzing through one region or another of empty space is not (presumably) a feature of the world that *matters*. What we really want is the ability to make *long-term evaluative practical predictions*: predictions about the effects of our present choices on evaluatively significant features of the far future. The epistemic challenge to longtermism claims that our ability to make this sort of prediction is so limited that, even if we concede the astronomical *importance* of the far future, the longtermist thesis still comes out false.

A bit more precisely:

The Pessimistic Thesis Let S be an "evaluatively significant" state of the world—a state such that the expected value of being in state S at a given time is significantly greater (or less) than the expected value of being in state $\neg S$. Let S_t be the proposition that the world is in state S at time t. And let O_1 and O_2 be any pair of alternative options faced by a present-day human agent. For any such S and O_1/O_2, $|Pr(S_t|O_1) - Pr(S_t|O_2)|$ decreases as t increases, and decreases quickly enough that most of the difference in expected value between O_1 and O_2 lies in the near future.

The Pessimistic Thesis could be true in several importantly distinct ways, which we will briefly survey in the concluding section. But our main focus will be in one particular line of reasoning that leads to the Pessimistic Thesis.

Weak Attractors Let S be any of the states that seem like promising objectives for longtermist interventions: e.g., the continued existence of the human species, the existence of just or rational political institutions, an enlargement of our "moral circle", a high growth rate in per capita consumption, etc. Conceding for the sake of argument that S has significantly greater expected value than its complement, and that we can substantially increase its probability over the medium term, interventions that have S as their objective nevertheless generate very little expected value in the far future, because our influence "fades out" over time: Even if we can substantially increase the probability that the world is in state S, say, 100 years from now, this will not substantially increase the probability that the world is in state S (or any other desirable state) a million or a billion years from now. That is, for any alternatives O_1 and O_2 facing present-day agents, $|Pr(S_t|O_1) - Pr(S_t|O_2)|$ may be significant for small values of t, but thereafter decreases quickly enough that attempts to influence S generate very little far future expected value.

Weak Attractors suggests that the same astronomical timescales that drive the *importance* of longtermist aims also drive their *intractability*: As we look further

into the future, the importance of our civilization being in one state or another may increase (and may become extremely large), but our capacity to predictably influence its state will decrease. The case for longtermism may then depend on which of these effects is dominant over particular timescales. The next three sections will investigate this question.[11]

4 The cubic growth model

In this section and the next, I set out two models for the expected value of longtermist interventions. Both models incorporate the idea that we can influence the probabilities of alternative states of the world less at more remote times, as suggested by Weak Attractors. They thereby allow us to evaluate Weak Attractors as a skeptical challenge to longtermism by quantifying this long-term "fade-out" of our capacity for predictable influence and seeing what it does to 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

[11]It is not obvious that what I am calling the "epistemic challenge" to longtermism (including Weak Attractors and other variants of the challenge to be discussed in §7) is genuinely *epistemic*. The challenge asserts, in essence, that we cannot identify interventions that reliably (with significant net probability) influence the far future in high-value ways. This could be true in two different ways: (1) There simply are no such interventions: Even with perfect information about the present state of the world and unlimited reasoning abilities, there would be nothing we could do to influence the far future in important ways. (2) Although such interventions exist, we lack the epistemic capacity to identify them. Perhaps with perfect information about the present state of the world and unlimited reasoning abilities, for instance, I could have a large positive influence on the far future by writing an optimific book containing (i) a list of precise instructions for key individuals and institutions that, when implemented, would put human civilization on a very good long-term trajectory, along with (ii) several revolutionary mathematical theorems and surprising-but-accurate empirical predictions, impressive enough to convince those key actors to take my advice. But given my epistemic limitations, I am unable to identify the sequence of words that would compose such a book.

In situation (1), the problem for longtermism is a "control deficit". In situation (2), the problem is an "epistemic deficit". It seems somewhat more plausible to me that we are in situation (2). But if we are in either situation, it seems to matter very little (at least for our purposes in this paper) *which* situation we are in. So I will henceforth ignore the distinction and continue to describe the challenge to longtermism as "epistemic".

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, B and L. B is a short-termist "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. In the 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. B would spend the $1 million on direct cash transfers to people in extreme poverty, through an organization like GiveDirectly. L would spend the $1 million on mitigating existential risks to human civilization, say by supporting research on pandemic risks from novel pathogens.[12]

The short-termist benchmark B, I will assume, has an expected value that is specified exogenously to the model. In the working example, where B is $1 million in direct cash transfers, EV(B) might be estimated by a standard cost-effectiveness evaluation of the sort produced by charity evaluators like GiveWell. For purposes of the example, I will assume that EV(B) is 3000 QALYs. This number is chosen partly for convenience, but is in line with recent research.[13]

[12]Existential risk mitigation is just one of several categories of longtermist intervention. But it has received the most attention to date, and its potential payoffs are especially easy to quantify.

[13]GiveWell (2019) estimates that the cost of creating a unit of value equivalent to averting the death of an individual under age 5 by means of direct cash transfers is $12,298. If we assume that averting the death of someone younger than 5 has an expected value of 40 QALYs, then this works out to a cost of $307.45 per QALY-equivalent, which implies that $1 million of direct cash transfers has an expected value of approximately 3253 QALY-equivalents. (GiveWell's estimates seem to represent average cost-effectiveness rather than marginal cost-effectiveness, but given the minuscule total volume of cash transfers relative to the number of people in extreme poverty, I assume that these are not very different, and that an additional $1 million in cash transfers would not significantly change their marginal cost-effectiveness.)

It should be noted that GiveWell do not consider direct cash transfers the *most* cost-effective of the range of "short-termist" interventions they evaluate: They rate a range of public health interventions as more cost-effective, and deworming in particularly as 19.3 times more cost-effective. But I take cash transfers as the short-termist benchmark for two reasons: (i) My goal here is not address questions of prioritization among short-termist causes/interventions or among longtermist

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 EV(B) 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 3000 QALYs. Thus, L has greater expected value than B in the working example if and only if EV(L) > 1 V.

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

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 3020). Time in the far future is measured from this boundary, so for instance $t = 6$ years corresponds to the year 3026.

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 is something like "The accessible region of the Universe contains an intelligent

causes/interventions. So I choose to focus on "representative" or "benchmark" interventions of each kind rather than trying to identify the best interventions in each category. In this spirit, I take existential risk mitigation as the prototypical or representative longtermist intervention, though it is far from obvious that it is the most cost-effective longtermist intervention (Beckstead, 2013b). (ii) Public health interventions like deworming plausibly represent "low-hanging fruit" whose marginal cost-effectiveness will steeply decline once the diseases they aim to control are well in check, or when national governments assume responsibility for the necessary interventions. Cash transfers, on the other hand, can be expected to maintain their cost-effectiveness as long as there are still people in extreme poverty to whom cash can be transferred without too much leakage. Thus, cash transfers may be more representative of the typical marginal cost-effectiveness of short-termist interventions over the coming decades.

In any case, even the most optimistic estimates of EV(B) would raise the 3000 QALY figure I have adopted by at most a couple orders of magnitude, which as we will see does relatively little to change our qualitative conclusions about the case for longtermism.

civilization", S_0 means something like "The accessible region of the Universe contains an intelligent civilization in the year 3020", which is *roughly* equivalent to "Humanity survives the next thousand years."

We will represent the version of the epistemic challenge on which we are focused by the presence 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 context of 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 species on Earth or somewhere else in the accessible universe (i.e., somewhere in our future light cone), 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 be in state $\neg S$, regardless of what state it was in at $t = 0$. If a positive ENE has occurred, the world will be in state S, regardless of what state it was in at $t = 0$. Thus, if the longtermist intervention L succeeds in making a difference by putting the world in 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 B, from the agent's perspective.

Along with the presence of ENEs, the second key assumption of the cubic growth model 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 at a constant rate. 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.[14]

The model is characterized by the following parameters:

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 B. 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 short-termist benchmark. Formally, $p = Pr(S_0|L) - Pr(S_0|B)$.

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 for large values of the parameter r introduced below.)

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

[14] In assuming a constant rate of spatial expansion, 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 short-termist 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 EV(L), then L has already handily defeated B on the basis of nearer-term considerations.

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. (I use the star density of the Virgo Supercluster rather than the accessible universe as a whole because whether L or B 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.

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 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 can be understood as follows: The goal of L is to put the world into the target state S in the far future. It does this not by continuously acting on the world at every future period, but by increasing the probability that the world "starts off" in state S at the near future/far future boundary, and hoping that it will remain in that state. L is able to increase 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 time integral of (i) the expected value of being state S rather than $\neg S$ at time t (given by $(v_e + v_s n((t - t_l)s))$) multiplied by (ii) the probability that the state of the world at time t still depends on its state at $t = 0$ (given by e^{-rt}).

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 will eventually "win": After some point, the integrand of EV(L) will go monotonically to zero, and will do so quickly enough that EV(L) is guaranteed to be finite even without the presence of the eschatological bound. Thus, for any ε, there is some time t such that the total expected value of L at all times after t is less than ε.[15,16]

The cubic growth 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 (viz., comparing a short-termist and a longtermist intervention, rather than making comparisons among longtermist

[15]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.

[16]The assumption of merely-polynomial growth may seem revisionary relative to the assumption of exponential growth that is standard in economic models. But we are here concerned with growth in *value* (understood as total welfare), whereas exponential growth in standard economic models is growth in *consumption*. Given the standard assumption of an isoelastic function from consumption to welfare/utility with $\eta > 1$, individual per-period welfare is bounded above, and so in the long run—even if we assume endless exponential growth in per capita consumption—growth in total per-period welfare converges to the rate of population growth. This is complicated by the fact that η can reflect several different things: diminishing marginal contributions of consumption to welfare, individual risk-aversion with respect to welfare, social aversion to inequality, or any combination of these factors. Nevertheless, the assumption that individual welfare is bounded above—at least as a function of consumption—is not particularly revisionary.

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 early 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, viz., whether we are in state S or state $\neg S$. Of course, in the real world, actions can 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 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, it is widely believed that we live in a "time of perils" (Sagan, 1994) and that the likelihood 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 (like asteroid impacts or climate change). In §6, we will consider the effects of uncertainty about r, which introduces a kind of effective time-dependence: If we know that ENEs occur with a fixed chance in any given period but are unsure what that chance is, then the overall probability that the *first* ENE occurs in period t decreases as a function of t, since the assumption that no ENE has occurred before t favors hypotheses on which the per-period chance is small. But uncertainty over time-independent values of r is of course distinct in principle from genuine time-dependency, and cannot substitute for or approximate all of the ways in which r could plausibly be time-dependent (e.g., it cannot replicate a sharp

drop in the value of r in the years immediately after we start settling the stars).

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. And while it is harder to imagine existential catastrophes that might wipe out an interstellar civilization than those that might wipe out a planetary civilization, it is worth bearing in mind that *none* of the existential risks that most concern us today (climate change, nuclear war, engineered pathogens, artificial superintelligence...) were imaginable to anyone living even 200 years ago. So the difficulty of imagining existential risks to far-future civilization is only weak evidence that such risks will be minimal (and the apparent pattern of increasing existential risk in the 20th and 21st centuries gives at least some *prima facie* evidence that the future will be *more* dangerous than the present). 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 3020 (assuming we survive that long in the first place) and that, at least from our present epistemic vantage point, the annual probability of existential catastrophe is largely time-independent thereafter.

4.2 Filling in parameter values

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, viz., 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 B. That is, bounding the integral anywhere after $t = 10^8$ years will almost never affect whether $\mathrm{EV}(L) > \mathrm{EV}(B)$.

p is more consequential, and harder to estimate. I will make a lower-bound estimate based on the details of our 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 $p \approx 2 \times 10^{-14}$.[17]

This is an *extremely* 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 earlier time should have greater impact, all else being equal. (If nothing else, this is true

[17]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 $p = 2 \times 10^{-14}$.

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 evidential probability of extinction. Given our uncertainties, the evidential 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.

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 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 $p = 2 \times 10^{-14}$.

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 3000 QALYs, this means that $v_s = 10^5$ 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 around each settled star, which are then used to power simulated 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 $v_s = 3.\bar{3} \times 10^{20}$ valons per star per year.[18]

[18]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 (which, arguably, is only slightly positive). 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.

For now, I will adopt the more conservative figure, $v_s = 10^5$ 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 $v_e = 2 \times 10^6$ valons per year.[19]

The parameters d_g and d_s, which define 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), 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 $s = 0.1c$.[20]

[19]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.

[20]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 intergalactic 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-

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 3020). Other reasonable guesses would not qualitatively change our results.[21]

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 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

maximally-conservative assumption of the cubic growth model itself that interstellar settlement will eventually be feasible.

[21] For instance, if we instead used $t_l = 500$ years, the crucial value of r below which L overtakes B in expected value would only decrease from ~ 0.000182 to ~ 0.000178.

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 is plausible to suppose 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.[22]

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 The results of the model

We have now specified provisional values for all the model parameters except r. So we can see what the model tells us for various values of r. The results are summarized in Table 1. 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 $\mathrm{EV}(L) > \mathrm{EV}(B)$ iff r is less than ~ 0.000182 (roughly two-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

[22]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)	Horizon*
0.1	$\sim 4 \times 10^{-7}$	~ 214 years
0.01	$\sim 4.11 \times 10^{-6}$	~ 3232 years
0.001	$\sim 1.15 \times 10^{-3}$	$\sim 42,642$ years
0.000182	~ 1	$\sim 275,287$ years
0.0001	~ 11.1	$\sim 526,998$ years
0.00001	$\sim 1.11 \times 10^{5}$	~ 6 million years
0.000001	$\sim 1.58 \times 10^{8}$	~ 72 million years
0.0000001	$\sim 5.13 \times 10^{9}$	~ 821 million years
0.00000001	$\sim 1.46 \times 10^{13}$	~ 9.18 billion years

Table 1: The cubic growth model for $p = 2 \times 10^{-14}$, $t_f = 10^{14}$ years, $v_e = 2 \times 10^6$ valons per year, $v_s = 10^5$ 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 $t_l = 0$. [$*$] The "horizon" is the time after which the discount rate "wins" and the integrand of EV(L) is no longer significant—specifically, the point at which $t^3 e^{-rt}$ falls (permanently) below $1/t$.

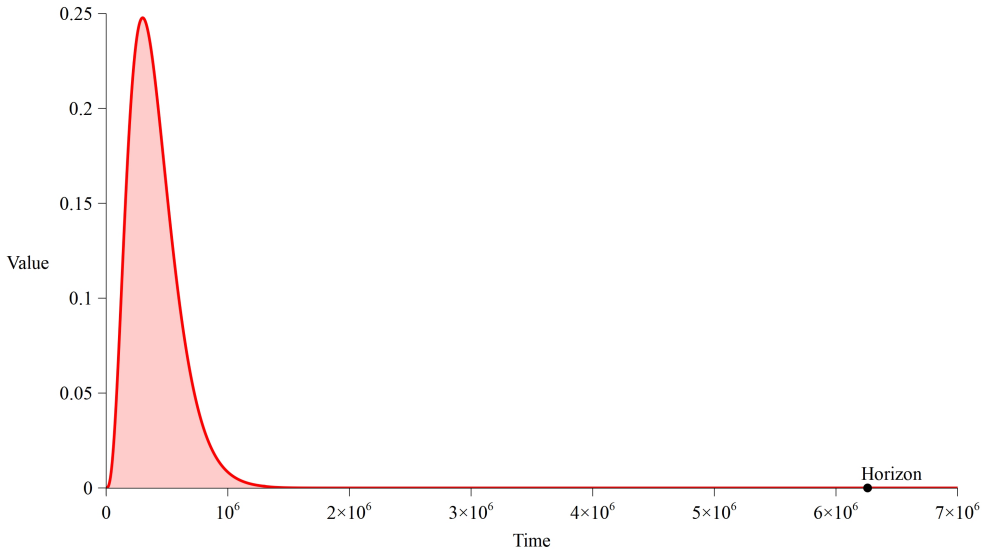

Figure 1: Integrand of EV(L) for $r = 0.00001$ (10^{-5}). Note that the large majority of expected value comes well before the "horizon" as we have defined it.

and an exponential discount rate does not automatically sink the case for longtermism, but does leave it open to question.

Second, note the "horizon" values in the last column of Table 1. This measures, in an admittedly crude way, the point at which the discount rate "overwhelms" the growth rate, so that further contributions to EV(L) are no longer significant. As Figure 1 illustrates, this estimate is quite generous, marking a point far in the tail of the integrand of EV(L). It is notable that, even for values of r on which the longtermist option is expectationally superior, the horizon is not always astronomically far in the future. The results in Table 1 therefore suggest the possibility that, although we should be longtermists, we should be longtermists 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 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 note 17) could still raise EV(L) by six or seven orders of magnitude, enough to make the case for L over B 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![23] 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). 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 transi-

[23]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).

tion 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.[24] 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 future civilization will not 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

[24] 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.

r	EV(L)	Horizon*
0.1	$\sim 4 \times 10^{-7}$	$\sim$ 36 years
0.01	$\sim 4 \times 10^{-6}$	$\sim$ 647 years
0.001	$\sim 4 \times 10^{-5}$	$\sim$ 9118 years
0.0001	$\sim 4 \times 10^{-4}$	$\sim$ 116,671 years
0.00001	$\sim 4 \times 10^{-3}$	$\sim$ 1.42 million years
0.000001	$\sim 4 \times 10^{-2}$	$\sim$ 17 million years
0.0000001	$\sim 4 \times 10^{-1}$	$\sim$ 191 million years
0.00000004	~ 1	$\sim$ 501 million years
0.00000001	$\sim$3.97	$\sim$ 2.15 billion years
0	20	n/a

Table 2: The steady state model for $p = 2 \times 10^{-14}$, $t_f = 5 \times 10^8$ years, and $v_e = 2 \times 10^6$ valons per year. The "horizon" is now simply the point at which e^{-rt} falls permanently below $1/t$.

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.[25]

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: EV(L) exceeds EV(B) only when $r <\sim 0.00000004$ 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 researching existential risks can buy a one-in-ten-billion reduction in the probability of near-future existential catastrophe) and $v_e = 2 \times 10^8$ 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

[25]Adopting the larger figure would have almost no effect on the values of EV(L) reported in Table 2 below, except for $r = 0$, where it would increase EV(L) by one order of magnitude.

greater population and greater average welfare), then EV(L) exceeds EV(B) as long as $r <\sim 0.02$ per year. And keeping r below *that* threshold seems eminently feasible, even for a purely planetary civilization.

6 Parameter uncertainty and model uncertainty

6.1 Incorporating uncertainties

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 two-in-ten-thousand (0.000182) 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 four-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 extinction events (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 EV(L) exceeds EV(B) 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 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 EV(L) on that basis. But for our purposes, this approach has significant drawbacks: EV(L) would be extremely sensitive to the tails of the distributions for parameters like r, s, 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

Parameter	**Confidence level**	**Value**	**Min. EV(L)**
(Cubic growth)	10^{-3} (0.1%)	—	$\sim 4.12 \times 10^{-5}$ V
r	10^{-3} (0.1%)	10^{-6} ENEs/yr	$\sim 1.58 \times 10^{2}$ V
s	10^{-2} (1%)	$0.8c$	$\sim 4.47 \times 10^{-5}$
v_s	10^{-6} (0.0001%)	$3.\bar{3} \times 10^{20}$ (V/yr)/star	$\sim 3.83 \times 10^{6}$ V

Table 3: Effects of uncertainty re the cubic growth model and the values of key parameters in that model, considered in isolation. Each row indicates the minimum value of EV(L) that results from assigning at least the specified level of confidence to values of the specified parameter at least as favorable to longtermism as the specified value. Except where otherwise specified, calculations assume 10^{-3} credence in the cubic growth model (with remaining credence in the steady state model), $r = 10^{-3}$ ENEs/yr, and the values for other model parameters specified in §4.2.

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, assuming we survive 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 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 $\mathrm{EV}(L)$.[26]

Bounding our confidence levels in this way is an unavoidably subjective exercise. Nevertheless, it seems to me that the bounds I've identified are quite conservative and hard to reasonably dispute. Given our enormous uncertainty about all aspects of the far future, we should distribute our credence liberally over a wide range of scenarios, and it is unreasonable for instance to be extremely confident that we will not choose to settle the stars, or that we will not achieve a significantly higher level of security and stability than we enjoy today. It seems clear, therefore, that any credence distribution that didn't satisfy the confidence bounds described in Table 3 would be overconfident.

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 $\mathrm{EV}(L) > \mathrm{EV}(B)$ (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 low credence in the cubic growth model itself.

But when we consider all these uncertainties together, the picture is clearer: Combining the four confidence bounds in Table 3 guarantees that $\mathrm{EV}(L) >\sim 1.44 \times 10^{10}$ V.[27]

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 §4.3) and because the range of uncertainty with

[26] In the case of v_s, we must also assume that its distribution is not supported on negative values. The possibility that the target state S might have negative value grounds a different kind of epistemic challenge to longtermism, "Neutral Attractors", which I discuss briefly in the concluding section.

[27] 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.

respect to r is very large. For instance, $r = 10^{-8}$ implies $\mathrm{EV}(L) \approx 1.45 \times 10^{13}$, 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 $\mathrm{EV}(L) > \mathrm{EV}(B)$. 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.[28]

A final point about the effects of uncertainty: So far, I have simply assumed a total utilitarian normative 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 (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.[29]

Distinctions between different kinds of uncertainty may be relevant elsewhere in our assessment of longtermism—in particular, in deciding how to weigh the possibility of scenarios like Dyson Spheres that involve large numbers of artificial (or

[28]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).

[29]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 (forthcoming), among others. For rival views, see Nissan-Rozen (2012), Gustafsson and Torpman (2014), Weatherson (2014), and Harman (2015), among others.

otherwise non-human-like) minds.[30] If we are uncertain whether or to what degree the sort of "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 minds instantiated on computer hardware would have the sort of subjective experiences we care about. But this may also feel more akin to moral uncertainty, and we may therefore feel reluctant to simply go by expected value.

6.2 Fanaticism

By the rules of the expected value game, the case for longtermism appears to survive 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 expectational reasoning goes wrong, or at least demands too much of us, in situations involving these "Pascalian" probabilities. (See for instance Bostrom (2009), Monton (2019).) 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.[31] Designating this minimum value μ^*, we can then measure the "Pascalian-ness" of a

[30] Thanks to Hilary Greaves for this point.

[31] 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.

choice situation on the unit interval by the formula $1 - 2\mu^*$.[32]

By this measure, the preceding analysis suggests that the choice between longtermist and short-termist interventions *could* be extremely Pascalian. We have found that longtermist interventions can have much greater expected value than their short-termist 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., the prediction of large-scale interstellar settlement, astronomical values of v_s, large values of s, and—in particular—small values of r). It could 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 its short-termist competitors—depends on a conjunction of improbable assumptions with joint probability on the order of (say) 10^{-18} or less. In this case, by the measure proposed above, the choice between L and B is extremely Pascalian ($1-(2\times10^{-18})$ or greater).

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 B 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 expectational 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 a fanatical extreme.[33]

[32]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.

[33]In Tarsney (ms), 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 rationality combined with (ii) recognition of the high level of "background uncertainty" about the choiceworthiness of our options that is engendered by attaching normative weight to aggregative consequentialist considerations. Simplifying the story considerably: Under levels of background uncertainty that seem warranted at least for total utilitarians, the decision-theoretically modest requirement to reject stochastically dominated options implies that we are generally required

7 Drawing conclusions

The preceding investigation suggests several broad conclusions:

1. If we accept a total utilitarian (or other aggregative consequentialist) moral framework, respond to uncertainty by simply maximizing expected objective value, and therefore do not mind premising our choices on minuscule probabilities of astronomical payoffs, then the case for longtermism seems robust to the kind of epistemic worry we have considered. While there are plausible point estimates of the relevant model parameters that favor short-termism, once we account for uncertainty, it takes only a very small credence in combinations of parameter values more favorable to longtermism for $\mathrm{EV}(L)$ to exceed $\mathrm{EV}(B)$.

2. There are, however, *prima facie* plausible worldviews on which the utilitarian case for longtermism 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 seems to depend to a significant extent on the possibility of interstellar settlement: It is significantly harder (though far from impossible) to make that case in the steady state model.

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 stable, since it is these interventions whose effects are likely to persist for a very

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 our degree of background uncertainty and on other features of the choice situation, but for total utilitarians in ordinary choice situations, it is probably no greater than 10^{-9} (and may be considerably smaller). So, if the stochastic dominance approach is correct, the probabilities we have considered in this paper—starting with $p = 2 \times 10^{-14}$—are on the threshold, from a utilitarian point of view: It could turn out, on further analysis, that the utilitarian case for longtermism 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 longtermist interventions have greater expected value than short-termist interventions, they are nevertheless rationally optional. Resolving this question would require much more precise estimates both of the various probabilities that determine the expected value of particular longtermist interventions and of the probability distribution that describes a utilitarian's rationally warranted background uncertainty about the amount of value in the Universe.

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* low values of r (say, on the order of 10^{-10} per year or less), 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 human-originating civilization 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 "Weak Attractors" challenge to longtermism 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.

Additionally, there are other potential sources of epistemic resistance to longtermism besides Weak Attractors that this paper has not addressed. In particular, these include:

Neutral Attractors To entertain small values of r, we must assume that the state S targeted by a longtermist intervention, and its complement $\neg S$, are both at least to some extent "attractor" states: Once a system is in state S, or state $\neg S$, it is unlikely to leave that state any time soon. But to justify significant values of v_e and v_s, it must also be the case that the attractors we are able to target differ significantly in expected value. And it's not clear that we can assume this. For instance, perhaps "large interstellar civilization exists in spatial region X" is an attractor state, but "large interstellar civilization exists in region X with healthy norms and institutions that generate a high

level of value" is not. If civilizations tend to "wander" unpredictably between high-value and low-value states, it could be that despite their astronomical *potential* for value, the *expected* value of large interstellar civilizations is close to zero. In that case, we can have persistent effects on the far future, but not effects that *matter* (in expectation).

Immediate Cluelessness I have focused on the potential long-term decay of our predictive abilities due to the continuing possibility of unpredictable events (ENEs). But perhaps I have been too generous about our capacity for "medium-term" practical prediction, on the scale of hundreds or thousands rather than millions of years. In other words, perhaps my allegedly lower-bound estimate of $p = 2 \times 10^{-14}$ was over-optimistic—e.g., because the world is "chaotic" in such a way that our ability to predict the effects of our interventions even on a scale of decades is practically zero or, on the other hand, because the world is deterministic enough that we are already locked into trajectories that are difficult or impossible to change, even temporarily.

Imprecise probabilist and non-probabilist objections At many points in this paper, I have relied on the assumption that we can assign precise probabilities to decision-relevant possibilities, even when we have nothing like observed frequencies on which to base those probabilities. But many are skeptical of this assumption. There are many other formal and informal frameworks for representing uncertainty, which generally demand significant modifications to standard expectational decision theory. Whether the case for longtermism fares better or worse against epistemic challenges in these alternate frameworks is a large question which must be addressed framework by framework. 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, there is some *prima facie* reason to suspect that things may be more difficult in other epistemic and decision-theoretic frameworks that are less welcoming to this kind of reasoning.[34]

My own sense is that Weak Attractors is the most serious epistemic challenge to longtermism, but that is little more than a hunch, and these other challenges should be thought through carefully as well.

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

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

even within the most hospitable normative frameworks (like total 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 very 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 (viz., for comparing generic longtermist and short-termist 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. (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 B.

Simplification: The longtermist intervention L has only a single effect, viz., 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 EV(B) that we adopted to analyze the working example make any attempt to model long-term/"flow-through" effects from the short-termist intervention B.

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 "short-termist" 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 short-termist 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.[35]

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 assump-

[35]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.

tions of the model) they do not have a significant effect on the comparison between longtermist and short-termist 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 short-termist 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.

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

Beckstead, N. (2013b). A proposed adjustment to the astronomical waste argument. LessWrong. URL: `https://www.lesswrong.com/posts/5czcpvqZ4RH7orcAa/a-proposed-adjustment-to-the-astronomical-waste-argument`. Published 27 May 2013. Accessed 4 April 2020.

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.

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.

Broome, J. (2013). *Rationality Through Reasoning*. Wiley-Blackwell.

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 (2019). 2019 GiveWell cost-effectiveness analysis – version 2. URL: `https://docs.google.com/spreadsheets/d/1xBKlshqbu6H-uByB2INEaDNPpj_Ulz0WrrFwgrnkXfk/`. Published 25 January 2019. Accessed 18 April 2019.

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

Greaves, H. and W. MacAskill (2019). The case for strong longtermism. *Global Priorities Institute Working Paper Series*. GPI Working Paper No. 7-2019.

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.

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 (forthcoming). Why maximize expected choice-worthiness? *Noûs*.

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. (forthcoming). Maximal cluelessness. *Philosophical Quarterly*.

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.

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.

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.

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. (2018, Jul). Exceeding expectations: Stochastic dominance as a general decision theory. *arXiv e-prints*, arXiv:1807.10895.

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.

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.