The scope of longtermism

David Thorstad

Forthcoming in AJP, please cite published version

Abstract

Longtermism is the thesis that in a large class of decision situations, the best thing we can do is what is best for the long-term future. The *scope question* for longtermism asks: how large is the class of decision situations for which this is true? In this paper, I suggest that the scope of longtermism may be narrower than many longtermists suppose. I identify a restricted version of longtermism: *swamping axiological strong longtermism* (swamping ASL). I identify three *scope-limiting factors* — probabilistic and decision-theoretic phenomena which, when present, tend to reduce the prospects for swamping ASL. I argue that these scope-limiting factors are often present in human decision problems, then use two case studies from recent discussions of longtermism to show how the scope-limiting factors lead to a restricted, if perhaps nonempty, scope for swamping ASL.

1 Introduction

If we play our cards right, the future of humanity will be vast and flourishing. The Earth will be habitable for at least another billion years. During that time, we may travel well beyond the Earth to settle distant planets. And increases in technology may allow us to live richer, longer and fuller lives than many of us enjoy today.

If we play our cards wrong, the future may be short or brutal. Already as a species we have acquired the capacity to make ourselves extinct, and many authors put forward alarmingly high estimates of our probability of doing so (Bostrom 2002; Leslie 1996; Ord 2020). Even if we survive long into the future, technological advances may be used to breed suffering and oppression on an unimaginable scale (Sotala and Gloor 2017; Torres 2018).

Some authors have taken these considerations to motivate *longtermism*: roughly, the thesis that in a large class of decision situations, the best thing we can do is what is best for the long-term future (Beckstead 2013; Greaves and MacAskill 2021; Greaves et al. forthcoming; MacAskill 2022; Ord 2020). The *scope question* for longtermism asks: how large is the class of decision situations for which this is true?

Longtermism was originally developed to describe the decisions facing present-day philanthropists. Longtermists suggest that the best thing philanthropists can do today is to safeguard the long-term future. But many have held that the scope of longtermism extends considerably further. Hilary Greaves and Will MacAskill (2021) suggest that longtermism extends into all of the most important decisions facing humanity today. Nick Beckstead (2013) suggests that longtermism extends into global health decisionmaking. And Owen Cotton-Barratt (2021) suggests that even most mundane decisions, such as selecting topics for dinner-table conversation, should be made to promote proxy goals which track far-future value.

In this paper, I argue that the scope of longtermism is far narrower than many longtermists suppose. Section 2 clarifies my target: *ex ante*, swamping axiological strong longtermism (swamping ASL). Section 3 illustrates a historical decision problem in which swamping ASL may have been true. However, Sections 4-6 develop three scope-limiting factors: probabilistic and decision-theoretic phenomena which, when present, substantially reduce the prospects for swamping ASL. I argue that these scope-limiting factors are often present in human decision problems. Sections 7-8 use a pair of case studies to show how the presence of these scope-limiting factors leads to a substantially restricted scope for ASL. Section 9 concludes.

2 Preliminaries

2.1 Longtermism: axiological and ex ante

Longtermism comes in both axiological and deontic varieties. Roughly speaking, *axiological longtermism* says that the best options available to us are often near-best for the long-term future, and *deontic longtermism* says that we often should take some such option. Longtermists standardly begin by arguing for axiological longtermism, then arguing that axiological longtermism implies deontic longtermism across a wide range of deontic assumptions. In order to avoid complications associated with the passage between axiological and deontic claims, I focus on axiological rather than deontic longtermism.

Axiological longtermism can be construed as an *ex ante* claim about the values which options have from an *ex ante* perspective, or as an *ex post* claim about the value that options will in fact produce. It is generally thought that *ex post* longtermism is more plausible than *ex ante* longtermism, since many of our actions may in fact make a strong difference to the course of human history, even if we are not able to foresee what that difference will be.[1] For this reason, most scholarly attention has focused on *ex ante* versions of longtermism, and I follow this trend here.

The best-known view in this area is what has been called axiological strong longtermism (ASL), which holds in any decision problem for which:

> **(ASL)** The option that is *ex ante* best is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.[2,3]

My target in this paper will be a restricted form of ASL.

2.2 Swamping axiological strong longtermism

Let a *longtermist option* be an option whose *ex ante* effects on the very long-run future are near-best.[4] ASL holds whenever the *ex ante* best option is a longtermist option. This can happen in two ways.

[1] However, Section 4 and on some views also Section 6 will place limits on the scope of *ex post* longtermism.

[2] This is the form of longtermism considered in Greaves and MacAskill (2019). Greaves and MacAskill (2021) defend a scope-restricted version of ASL, focusing only on the most important decision situations facing humanity today. I use the older, more general formulation of ASL in order to avoid ruling out wider scopes for ASL, and indeed Greaves and MacAskill are sympathetic to the idea that ASL has fairly wide scope.

[3] How should an action's impact on the value of the long-term future be understood? One candidate is through causal modeling: treat actions as interventions on the world today, propagate forwards to assess their causal consequences, then apply a normative theory to assess the change in the value of the future.

[4] More formally, suppose that value is temporally separable, so that $V_o = S_o + L_o$ where V_o, S_o, L_o are the overall, short-term and long-term values of option o. Assess changes in value $\Delta V_o, \Delta S_o, \Delta L_o$ relative to a baseline, such as the effects of inaction. And take an expectational construal of *ex ante* value. Then a *longtermist option* is such that $E[\Delta L_o] \geq T * max_{o' \in O} E[\Delta L_{o'}]$ where O are the options available to the actor and T is a context-independent threshold for effects that count as 'near-best'. Perhaps we might take $T = 0.9$.

First, let a *swamping option* be an option whose expected long-term benefits exceed in magnitude the expected short-term effects produced by any option.[5][6] I call these swamping options because their long-term effects begin to swamp short-term considerations in determining *ex ante* value. The first way for ASL to be true in a decision problem is if the best option is both a longtermist option and a swamping option.

> **Swamping axiological strong longtermism (Swamping ASL)** The option that is *ex ante* best is a swamping longtermist option.

My focus in this paper will be on swamping ASL.

Second, the best option may be a *non-swamping longtermist option*, an option whose expected long-term effects are near-best, but do not exceed in magnitude the expected short-term effects of all other options. One way to defend the value of non-swamping longtermist options would be through the *convergence thesis* that what is best for the short-term is often near-best for the long-term as well.[7] The convergence thesis suggests that even when long-term effects do not swamp short-term effects in magnitude, the best option may nonetheless be a longtermist option, since the best short-term option will often be near-best for the long-term.

I focus on swamping ASL for three reasons. First, swamping ASL figures in leading philosophical arguments for ASL and in most nonphilosophical treatments of longtermism. Second, swamping ASL is the most distinct and revisionary form of ASL, because it tells us that the short-termist options we might have assumed to be best are in fact often not best.[8] Third, swamping ASL underlies many of the most persuasive arguments

[5] Using the notation and assumptions of the previous footnote, a *swamping longtermist* option is such that $E[\Delta L_o] > max_{o' \in O} |E[\Delta S_{o'}]|$ where O are the options available to the actor. This is a simplification of the model from Greaves and MacAskill (2019).

[6] In this definition, 'any option' should be read in an awareness-unrestricted sense, so that swamping options cannot become non-swamping options when an agent expands her awareness of new options. Thanks to a referee for pressing me to clarify this point.

[7] For example, you might think that the best thing we can do to ensure a good future is to promote economic growth (Cowen 2018), and that is also among the best things we can do for the short-term. Note, however, that this may be an example of a swamping longtermist option.

[8] Strictly speaking, this does not follow from swamping ASL since swamping ASL is compatible with

from axiological to deontic longtermism, which rely on the claim that sufficiently strong duties to promote impartial value may trump competing nonconsequentialist duties. As we move away from swamping longtermism, obligations to promote long-term value will diminish in strength, putting pressure against the inference from axiological to deontic longtermism.

2.3 Scope-limiting phenomena

In this paper, I illustrate three *scope-limiting phenomena*. These are probabilistic and decision-theoretic phenomena which, when present in a decision problem, tend to reduce the prospects for swamping ASL to correctly describe that problem. Sections 4-6 introduce the scope-limiting phenomena that will concern me: rapid diminution (Section 4); washing out (Section 5); and option unawareness (Section 6). I argue that each scope-limiting phenomenon is often present in the decisions that we face, then show how the presence of each phenomenon reduces the prospects for swamping ASL.

To say that these scope-limiting phenomena reduce the prospects for swamping ASL is not to say that swamping ASL has empty scope. Section 3 illustrates a case in which swamping ASL may perhaps have been true, and Section 7 argues that this case may not be significantly afflicted by any of the scope-limiting phenomena. Moreover, it is not impossible for swamping ASL to correctly describe some cases where all of the scope-limiting phenomena obtain. However, the presence of these scope-limiting phenomena does put pressure on many cases in which swamping ASL has been claimed to obtain. Section 8 illustrates one case of this type. More generally, the scope-limiting phenomena combine to put the burden back on the longtermist to show, in any given case, that the scope-limiting phenomena are absent, or else how swamping ASL may continue to hold despite the scope-limiting phenomena. Section 9 uses this thought to reflect on the scope of longtermism and the argumentative burdens facing longtermists.

the convergence thesis. However, in practice most of the examples used to support swamping ASL are not near-best in their short-term effects. Some longtermists may disagree (Shulman and Thornley forthcoming), though this is controversial.

Summing up, my target in this paper is *ex ante*, swamping axiological strong longtermism. I illustrate three scope-limiting phenomena to suggest that swamping ASL has more limited scope than many longtermists suppose. But first, let us consider where swamping ASL may be plausible.

3 Swamping ASL and the Space Guard Survey

A popular way to motivate swamping ASL is to think about risks of human extinction (Bostrom 2013; Greaves and MacAskill 2021; Ord 2020). Now on some views, the continued survival of humanity may have indifferent or even negative value (Benatar 2006). Given our potential to spread death and suffering, the universe may be better off once it is rid of humanity. On these views, risks of human extinction will not motivate swamping ASL.[9] But many philosophers are cautiously optimistic that the survival of humanity would be a good thing (Beckstead 2013; Ord 2020; Parfit 2011). On these views, it may be very important to protect humanity from premature extinction. And in some cases, decisions to mitigate extinction risk may motivate swamping ASL.

One way that humans might go extinct is through the impact of a large asteroid on Earth.[10] NASA classifies asteroids with diameter greater than 1 kilometer as catastrophic, capable of causing a global calamity or even mass extinction.[11] Our best estimates suggest that such impacts occur on Earth about once in every 6,000 centuries (Stokes et al. 2017). It may be worth our while to detect and prepare for such events.[12]

[9] Perhaps these views might motivate a version of swamping ASL on which efforts to bring about human extinction are swamping options. I will not be concerned with this version of swamping ASL in this paper.

[10] Indeed, there is mounting evidence that an asteroid impact during the Cretaceous period killed every land-dwelling vertebrate with mass over five kilograms (Alvarez et al. 1980; Schulte et al. 2010). As recently as 2019, an asteroid 100 meters in diameter passed five times closer to the Earth than the average orbital distance of the moon and was detected only a day before it arrived (Zambrano-Marin et al. 2021).

[11] In addition to localized blast damages (Chesley et al. 2002) and extreme regional tsunamis (Paine 1999), an impact of this size would lead to global climactic changes reminiscent of nuclear winter, radically affecting the viability of agriculture and posing a strong threat of ecosystem collapse (Toon et al. 1997).

[12] However, the inference from Cretacean extinction risk to human extinction risk may pose some difficulties. After all, technologically sophisticated human civilizations are more capable than dinosaurs of adapting to swift changes in global climate.

As evidence mounted of the threat posed by asteroid impacts, the United States Congress funded the Space Guard Survey, a collection of projects aimed at tracking potentially dangerous asteroids, comets and other near-Earth objects. Since the 1990s, the Space Guard Survey has mapped approximately 95% of the near-Earth asteroids with diameters exceeding 1 kilometer, at a cost of $70 million. From an *ex ante* perspective, how valuable was the Space Guard Survey?

A longtermist might argue in the following way. Assume conservatively that the Space Guard Survey can only accurately predict impacts during the next century. Next, suppose that if an undetected asteroid with diameter greater than 1 kilometer were to strike Earth during the next century, the chance of extinction would be one in a million. Now, consider that estimates of the expected number of future humanlike lives range from about 10^{13} to 10^{55} (Bostrom 2014; Newberry 2021). This puts the Space Guard Survey's expected cost of saving a life at about $7 per expected future life, and fractions of a penny using anything but the most conservative estimate of future lives.[13] For comparison, our best estimates put the cost of saving a life through short-termist interventions at several thousand dollars (GiveWell 2021), far exceeding the cost of the Space Guard Survey if we have any confidence at all in our ability to prepare for and survive an otherwise-catastrophic impact with sufficient warning.[14][15]

There are some ways that we might contest the longtermist's valuation. For example, estimates of the number of future humanlike lives might be substantially reduced once

[13] This estimate is arrived at by multiplying the lower-bound expected number of future lives by the per-century probability of a catastrophic asteroid impacting Earth, as well as by the probability that an undetected catastrophic asteroid impact would lead to extinction, then dividing the result by the program cost.

[14] For example, even a one-in-a-thousand level of confidence in our ability to prepare puts cost-effectiveness at $7,000 per life on the upper bound, and $7*10^{-39}$ at the lower bound.

[15] What might be done to prepare? In addition to chancy efforts at asteroid deflection (Sanchez et al. 2009) we might increase the resilience of global food supplies (Baum et al. 2015), create and protect key infrastructure (Garshnek et al. 2000), or create policy institutions charged with planning for the challenges posed by asteroids and other global catastrophic risks (Posner 2004), such as the continuation of national and international governance as well as the maintenance of law and order. Although these methods might fail to prevent millions of deaths, they could well ensure the survival of humanity and perhaps even the survival of most people and nations.

we incorporate other existential risks (Thorstad 2023). Or we might adopt a non-fanatical axiology on which small chances of preventing large harms are weighted far below their expected values.[16] But let us suppose, for the sake of argument, that the longtermist's valuation is correct.

Now consider the decision facing Congress in the early 1990s: whether to fund the Space Guard Survey or to redirect the money towards alternative programs. Suppose, plausibly, that the expected long-term effects of the Space Guard Survey were near-best out of all programs available for Congress to fund. Or, if this is not plausible, replace the Space Guard Survey with any program that had near-best expected long-term effects and repeat the argument. Then suppose we also grant that the expected long-term effects of the Space Guard Survey exceeded in magnitude the best-achievable short-term effects of any competing program. For example, we might estimate the long-term effects of the Space Guard Survey at several dollars per life saved, and the best-achievable short-term effects of competing programs at several thousand dollars per life saved. If this is right, then swamping ASL was true of Congress's decision problem. Funding the Space Guard Survey was the best thing that Congress could have done, its long-term effects were near-best, and they swamped in magnitude the expected short-term effects of all options. Indeed, it may be precisely on these grounds that Congress decided to fund the Space Guard Survey.

Some readers might disagree with the claim that swamping ASL holds of Congress's decision problem. Perhaps you hold a person-affecting axiology on which it is neither good nor bad to ensure that future humans come into existence. Or perhaps you think that the likely outcome of asteroid detection research is research into dangerous technologies for asteroid deflection, and that the dangers posed by these technologies are greater than the dangers they eliminate (Ord 2020). In this paper, I want to emphasize a different line of resistance: cases such as the Space Guard Survey are quite special (Section 7), in that

[16] See for example Monton (2019), Smith (2014) and Pettigrew (2022). For pushback see Isaacs (2016) and Wilkinson (2022). Note that some care is needed in interpreting these debates, since fanaticism can be a deontic thesis as well as an axiological thesis.

they avoid a number of scope-limiting phenomena (Sections 4-6) that serve to reduce the prospects for swamping ASL. This means that we can, and perhaps should, acknowledge some cases in which swamping ASL holds, while resisting swamping ASL as a description of many other decision problems.

4 Rapid diminution

In the next three sections, I illustrate a series of scope-limiting factors.[17] I argue that these factors are often present in the decisions that we face and that, when present, these factors substantially reduce the prospects for swamping ASL.

The first scope-limiting factor is *rapid diminution* (Figure 1). Fix an option *o* and consider the probability distribution over long-term impacts of *o*.[18] In most cases, the probabilities of long-term impacts decrease as those impacts increase in magnitude. If probabilities of impacts decrease more slowly than the magnitudes of those impacts increase, then the expected long-term consequences of *o* may be astronomically high. But if the probabilities of large impacts decrease quickly, the expected long-term impacts of *o* may be quite modest.

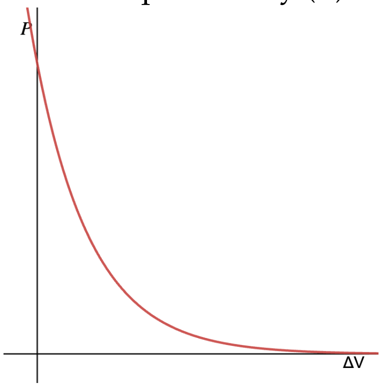
Rapid diminution is a familiar feature of many of the best-known probability distributions. For example, suppose that we model the expected long-term impact of *o* using a normal distribution, centered around the origin, with a standard deviation equivalent to the value of ten lives saved.[19] On this model, the probability of long-term impacts exceeding five times this value is less than one in a million. And the probabilities of astronomical long-term impacts, while nonzero, will be so negligible as to have no significant impact

[17] A referee notes that there may be important complementarities between the arguments for these scope-limiting factors. For example, evidential paucity may be a reason for forecasting skepticism, and persistence skepticism may make our evidence sparser than needed to ground high value estimates. This illustrates a key feature of my approach, which seeks to build an overlapping network of complementary challenges to longtermism.

[18] I.e. consider the probability distribution over the partition $\{[\Delta L = k] : k \in \mathbb{R}\}$.

[19] To be clear, I am not suggesting that a normal distribution is always the correct model of far-future impacts. Thanks to an anonymous referee for pressing me to address this point.

Figure 1: Rapid diminution in the probability (P) of long-term impacts (ΔV)



on the expected long-term impact of o.[20]

The argument from rapid diminution claims that most options exhibit rapid diminution in the probability of long-term impacts, limiting the contribution that long-term impacts can make to the expected value of those options.[21] Why think that most options exhibit rapid diminution? The case for rapid diminution is supported by *persistence skepticism*: the view that most of our actions do not make a large persisting impact on the long-term future.

We can assess the case for persistence skepticism by looking at the burgeoning academic field of persistence studies, which studies examples of persistent long-term changes (Alesina and Giuliano 2015; Nunn 2020).[22] The argument for rapid diminution claims that

[20] One thing that must be conceded is that even mundane actions may have widespread effects on the *identities* of future people (Lenman 2000; Greaves 2016). This may not imply that our actions have widespread effects on the value of the future, for example because mundane identity-affecting actions may not have significant distorting effects on the values of futures left open to us or the probabilities of those futures being realized (Shiller 2021), a point largely complementary to the argument in this section. One way in which identity-affecting actions could matter is on person-affecting views and other views in population ethics which treat long-term benefits to persons who would otherwise not have existed less favorably than longtermists often do. This would probably not tend to widen the scope of longtermism. Thanks to a referee for pressing me to discuss these points.

[21] On some views, the causal consequences of even the most mundane actions may be quite extreme (Schwitzgebel forthcoming). This might be interpreted as an extreme form of *ex ante* washing out on which the long-term future effects of actions are in fact likely to be large, but in expectation are quite small due to uncertainty-driven symmetry in the probabilities of future effects (Schwitzgebel ms). While that is not my view, this view would also tend to restrict the scope of swamping ASL.

[22] A referee notes that methodological limitations of persistence studies may prevent the field from fully examining all possible types of persistent effects. Concerns about methodological conservatism in the social sciences have been raised before (Noy and Noy forthcoming), and those moved by these concerns may wish to bring forward alternative sources of evidence in future research.

most options exhibit rapid diminution in the probabilities of significant long-term effects, which predicts that persistence studies should struggle to identify options with significant and persistent long-term effects. And indeed, that is what we find.

One of the most surprising results in persistence studies is that even many actions which we might have expected to have significant effects on the long-term future in fact make no persistent impact after years or decades.[23] For example, given the scale of American bombing in Japan and Vietnam, one might expect persistent economic effects in the heaviest-hit areas. Perhaps bombed areas might be poorer and less populous. Given the number of people affected and the magnitude of potential effects, this is exactly the type of persistent effect that would interest a longtermist. But a half-century later, there are no statistically significant differences between the most- and least-affected areas on standard economic indicators such as population size, poverty rates and consumption patterns (Davis and Weinstein 2008; Miguel and Roland 2011). For a striking example, the cities of Hiroshima and Nagasaki returned to their pre-war population levels by the mid-1950s.[24] More generally, the findings of persistence studies are largely negative: it is surprisingly difficult to identify historical actions which had large effects on the long-term future (Albouy 2012; Colella et al. 2019; Voth 2021).

Now it is true that persistence studies has identified a few effects which might be more persistent. For example, the introduction of the plough may have affected fertility norms and increased the gendered division of labor (Alesina et al. 2011, 2013); the African slave trade may have stably reduced social trust and economic indicators in the hardest-hit regions (Nunn 2008; Nunn and Wantchekon 2011); and the Catholic Church may be responsible for the spread of so-called WEIRD personality traits identified by comparative psychologists (Schulz et al. 2019). However, these findings need to be taken with three grains of salt.

[23] This provides evidence for persistence skepticism by suggesting that other effects may also be less likely to persist than they appeared.

[24] To be clear, the claim is not that the bombing of Hiroshima and Nagasaki had *no* significant long-term effects. After all, this bombing ended a major war. The claim is rather that many demographic and economic effects of wartime bombing which we might have expected to persist in fact failed to persist.

First, many of these findings are controversial, and alternative explanations have been proposed (Kelly 2019; Sevilla 2021). Second, these findings are few and far between, so together with other negative findings they may not challenge the underlying rarity of strong long-term effects. And finally, most of the examples in this literature also involve short-term effects of comparable importance to their claimed long-term effects.[25] Hence the persistence literature suggests that, although it is possible for actions to have persistent long-term effects which far outstrip their short-term effects, this is usually highly unlikely to happen.

At the same time, there is no doubt that some actions have a nontrivial probability of making persistent changes to the value of the future far greater than any of their short-term effects. As a result, we cannot get by with the argument from rapid diminution alone. We need to supplement rapid diminution with a second scope-limiting factor: washing out.

5 Washing out

A second scope-limiting factor is *washing out*.[26] Although many options have nontrivial probabilities of making positive impacts on the future, they also have nontrivial probabilities of making negative impacts. For example, by driving down the road I might crash into the otherwise-founder of a world government, but I might also crash into her chief opponent. As a result, the argument from washing out holds that there will often be significant cancellation between possible positive and negative effects in determining the expected values of options.

Expressed probabilistically, washing out occurs when the probability distribution of long-term impacts is highly symmetric about the origin (Figure 2).[27] There are two reasons

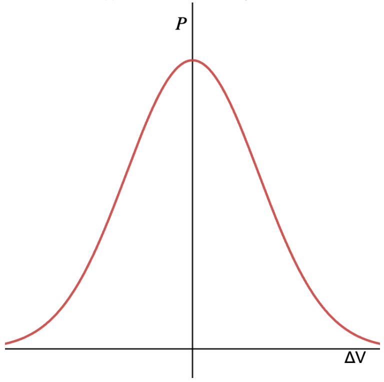
[25] For example, the African slave trade killed and enslaved millions of people, the Catholic Church exerted substantial influence on many aspects of civil and religious life, and the plough greatly increased agricultural productivity.

[26] For related discussions see Askell (forthcoming), Cowen (2006), and Friedrich (ms). For pushback see Dorsey (2012).

[27] Note that the argument from washing out does not rest on the claim that distributions will be perfectly

to expect that the probabilities of long-term impacts will often be highly symmetric about the origin. The first begins with the popular Bayesian idea that complete ignorance about the long-term value of an option should be represented by a prior distribution over possible long-term values that is fully symmetric about the origin (Jaynes 1957; Pettigrew 2016; White 2010). The motivation for this assumption is that if we know nothing about the long-term value of an option, then we should not assume anything about the directionality of its value. Next, the argument notes that we are often in a situation of *evidential paucity*: although we have some new evidence bearing on long-term values, often our evidence is quite weak and undiagnostic. As a result, the prior distribution will exert a significant influence on the shape of our current credences, so if the prior is symmetric then our current credences should be fairly symmetric as well. And a near-symmetric probability distribution about the origin gives significant cancellation when we take expected values.

Figure 2: Washing out in the probability (P) of long-term impacts (ΔV)



We can make a similar point by arguing for *forecasting pessimism*, the view that it is often very difficult to predict the impact of our actions on far-future value. For example, there is no doubt that the Roman sacking of Carthage had a major impact on our lives today, by cementing the Roman empire and changing the course of Western civilization. But even today, let alone with evidence available at the time, it is very difficult to say whether that impact was for good or for ill.

Forecasting pessimism acts like evidential paucity, limiting the degree to which even

symmetrical about the origin. Imperfect symmetry may generate significant cancellation on its own.

sharply directional forecasts can move us away from previously symmetric prior views.[28][29] When we make forecasts based on sparse data, we need to take account of the fact that the models we construct and the information we base them on are a noisy reflection of the underlying reality. As phenomena become more unpredictable, we should grow more willing to chalk up any apparent directionality in our forecasts to noisiness in the data we were dealt and the models we chose to construct. This means that we should take ourselves to learn less from any given forecast, and hence be less willing to revise our views away from their initial symmetry.

Why should we be pessimistic about our ability to reliably forecast long-run value? Intuitions about the sacking of Carthage are well and good, but it would be nice to have some concrete theoretical considerations on the table. Here are three reasons to think that we are often in a poor position to forecast long-run value.

First, we have limited and mixed *track records* of making long-term value forecasts. We do not often make forecasts even on a modest timeline of 20-30 years, and as a result there are only a few studies assessing our track record at this timescale.[30] These studies give a mixed picture of our track record at predicting the moderately-far future: in some areas our predictions are reasonably accurate, whereas in others they are not. But the longtermist is interested in predictions at a timescale of centuries or millennia. We have made and tested so few predictions at these time scales that I am aware of no studies

[28] Among the many ways to give formal expression to this idea, we might draw on Gabaix and Laibson's (2022) as-if discounting model. Assume we want to forecast the utility U_X produced by act X. As in many forecasting models, assume a forecast F_X of X is a noisy reflection of U_X, distributed as $F_X \sim U_X + \epsilon_X$ for $\epsilon_X \sim \mathcal{N}(0, \sigma_\epsilon^2)$ a normally distributed error term. Assume we begin with a symmetric prior $U_X \sim \mathcal{N}(0, \sigma_U^2)$. Upon observing signal $F_X = s$, Bayesian updating produces a posterior utility for X distributed as $\mathcal{N}(Ds, (1-D)\sigma_u^2)$, for $D = \frac{1}{1+(\sigma_\epsilon^2/\sigma_U^2)}$. Here D behaves like a discount rate, restricting the influence which the signal s has on moving the agent away from her prior. As forecasting becomes increasingly difficult, the noisiness σ_ϵ^2 of the error term increases, driving D towards zero and hence pushing the posterior increasingly back towards the symmetric prior.

[29] Note that forecasting pessimism remains distinct from evidential paucity. Forecasting pessimism can be motivated in some cases without evidential paucity, for example if the phenomenon at issue is difficult to forecast.

[30] For domain-specific track records see Albright (2002); Kott and Perconti (2018); Parente and Anderson-Parente (2011); Risi et al. (2019); Tetlock et al. (ms) and Yusuf (2009). For discussion see Fye et al. (2013) and Mullins (2018).

which assess our track record at this timescale outside of highly circumscribed scientific domains, and if our moderate-future track record is any indication, our accuracy may decline quite rapidly this far into the future.

Second, there is an enormous amount of *practitioner skepticism* on behalf of prominent academic and non-academic forecasters about the possibility of making forecasts on a timescale of centuries, particularly when we are interested in forecasting rare events, as longtermists often are.[31] Very few economists, risk analysts, and other experts are willing to make such predictions, citing the unavailability of data, a lack of relevant theoretical models, and the inherent unpredictability of underlying systems (Freedman 1981; Goodwin and Wright 2010; Makridakis and Taleb 2009).[32] And when risk analysts are asked to consult on the management of very long-term risks, they increasingly apply a variety of non-forecasting methods which enumerate and manage possible risks without any attempt to forecast their likelihood (Marchau et al. 2019; Ranger et al. 2013). If leading practitioners are unwilling to make forecasts on this timescale and increasingly suggest that we should act without forecasting, this is some evidence that the underlying phenomena may be too unforeseeable to effectively forecast.

Third, *value is multidimensional*. The value of a time-slice in human history is determined by many factors such as the number of people living, their health, longevity, education, and social inclusion. It is often relatively tractable to predict a single quantity, such as the number of malaria deaths that will be directly prevented by a program of distributing bed nets. And when we assess the track records of past predictions, we often assess predictions of this form. But the longtermist is interested in predicting value itself,

[31] A referee notes that it is important to specify the basis on which practitioner opinions are taken as evidence about the reliability of forecasting methods. Practitioner opinions may be reliable due to practitioners' theoretical and empirical knowledge about the practice of forecasting and the limits of existing forecasting methodologies. Practitioner opinion may also incorporate existing, albeit limited, track records of long-term forecasts. Some readers may place relatively less trust in the judgment of forecasting experts, in which case this will be an important point of disagreement.

[32] A referee notes that while the predictions that forecasters are willing to make provide important evidence about their beliefs about the reliability of forecasts, there may be a gap between belief and practice in this domain, due to features such as the methodological conservatism of the social sciences (Noy and Noy forthcoming). Readers concerned about the possibility of a gap may place relatively more weight on expressed opinions of forecasters in meta-analyses of long-term forecasts cited in this section.

which turns on many different quantities. This is harder to predict: distributing bed nets also affects factors such as population size, economic growth, and government provision of social services (Deaton 2015). So even if we think that the long-term effects of a program along a single dimension of value are fairly predictable, we may think that the ultimate value of the intervention is much less predictable.

Summing up, the argument from washing out claims that we often get significant cancellation between possible positive and negative effects of an intervention when taking expected values. One window into washing out comes from evidential paucity: because we have little evidence about long-term impacts, we should adopt a fairly-symmetric probability distribution over possible long-term impacts. The same phenomenon occurs in thinking about forecasting. Because our evidence about far-future value is sparse, we should think that our forecasts could easily have been different if we had received different evidence about the future, and as a result we get significant cancellation between possible positive and negative forecasts of far-future value.[33]

Together, rapid diminution and washing out put pressure on the scope of swamping ASL. They do this by suggesting that the expected far-future benefits of most options may be relatively modest, and may be significantly cancelled by the expected far-future costs of these options. In the next section, I illustrate a third and final scope-limiting factor: option unawareness.

6 Option unawareness

Rational *ex ante* choice involves taking the *ex ante* best option from the options available to you. But which options are these? We might take a highly unconstrained reading on which any option that is physically possible to perform belongs to your choice set. But in

[33] A referee notes that one theme lending support to forecasting skepticism is the complexity of the underlying phenomena being predicted. The referee notes that this may bear useful relationships to discussions of complex cluelessness (Greaves 2016). In particular, as we have seen, it may be inappropriate to take washing out to involve complete symmetry about the origin under conditions of complex cluelessness. For a different argument that cluelessness may complicate the case for longtermism, see Mogensen (2021).

practice, this reading seems to betray the *ex ante* perspective (Hedden 2012).

Suppose you are being chased down an alleyway by masked assailants. A dead end approaches. Should you turn right, turn left, or stop and fight? Trick question! I forgot to mention that you see a weak ventilation pipe which, if opened, would spray your attackers with hot steam. That's better than running or fighting. Let us suppose that, in theory, all of this could be inferred with high probability from your knowledge of physics together with your present perceptual evidence, but you haven't considered it. Does this mean that you would act wrongly by doing anything except breaking the pipe?

Many decision theorists have thought you would not act wrongly here. Just as *ex ante* choosers have limited information about the values of options, so too they have limited awareness of the many different options in principle available to them. Theories of *option unawareness* incorporate this element of *ex ante* choice by restricting choice sets to options which an agent is, in some sense, relevantly aware of (Bradley 2017; Karni and Vierø 2013; Steele and Stefánsson 2021).[34] In the present case, this means that your options are as first described: turning right, turning left, or stopping to fight. Unless, perhaps, you happen to be James Bond.

How is option awareness relevant to swamping ASL? To see the relevance, note that rapid diminution and washing out are features of options, not decision problems. Together, rapid diminution and washing out imply that many of the options we face will not be swamping longtermist options, because their expected far-future benefits may be relatively modest and may be significantly cancelled by expected far-future costs. However, swamping ASL is a thesis about decision problems, which present us with a set of options rather than a single option. Swamping ASL holds in any decision problem for which the *ex ante* best option is a swamping longtermist option. The presence of a single swamping longtermist option in a decision problem may be enough to vindicate swamping ASL.

[34] Although many philosophers incorporate an awareness constraint on options, Koon (2020) reminds us of the importance of avoiding overly subjectivist construals of an agent's option set. Though I do not follow Koon in taking an externalist view of options, I do think that many of Koon's criticisms of subjectivist views are important and correct. Thanks to a referee for pressing me to address this point.

This means that the number of options present in a decision problem bears strongly on the likelihood that swamping ASL will be true in that problem. If the vast majority of options are not swamping longtermist options, then swamping ASL will be unlikely to hold in decision problems containing a dozen options, since it is unlikely that any of these will be swamping longtermist options. But swamping ASL may be more likely to hold in decision problems containing millions or billions of options, simply because one of those options is likely to be a swamping longtermist option, and because swamping longtermist options are often, when present, the best options we can take.[35] Hence swamping ASL may be somewhat plausible before we restrict agents' option sets to incorporate their limited awareness of available options, but less plausible once option unawareness is incorporated.

To see the point in context, consider interventions aimed at combatting childhood blindness. Nick Beckstead (2013) has suggested that the short-term benefits of these interventions, namely preventing children from going blind, may be swamped by the long-term benefits of preventing blindness, such as speeding up a nation's economic development or changing the world's trajectory by changing the role that children will play in the national and global economy. Our discussion of rapid diminution and washing out suggests that, for most particular children, Beckstead's claim will be false. Because it is hard for a single individual to make a lasting impact on the long-term future, and because individuals may also make negative impacts on the long-term future, for most children, the expected benefit of preventing them from going blind will be driven primarily by short-term considerations, such as the value of not being blind.

However, perhaps it is not implausible that somewhere in the world, there is a collection of seventeen children and a sequence of days such that, if each child were given

[35] As always, there is a problem of option individuation, since it is often possible to chop a single option into millions or billions of nearly-identical options, but that is unlikely to improve the prospects of swamping ASL. Readers are invited to approach this discussion in a way that treats awareness of *relevantly different* options as raising the prospects for swamping ASL to be true. Like most philosophers, I do not pretend to be in possession of a formal criterion for relevant difference, or another fully formal solution to the problems induced by option individuation.

preventative treatment on the requisite day, the long-term trajectory of the world would be significantly improved. Let O^* be the option of giving just this course of treatment to each of the children in question. And perhaps it is not unreasonable to suppose that, in principle, the high value of O^* could be worked out *ex ante* on the basis of available information, even if the calculations required to see this would be astronomically complex.[36]

Now suppose that you have five thousand dollars to spend, and you want to use that money to combat childhood blindness. We might take an awareness-restricted view of your decision problem, on which you are deciding among donating to the half-dozen most prominent international efforts to combat childhood blindness. In this problem, swamping ASL may be relatively implausible. On the other hand, we might take an awareness-unrestricted view of your decision problem, on which you are deciding among any physically possible use of five thousand dollars to combat childhood blindness, including options such as O^*. In this awareness-unrestricted decision problem, swamping ASL may be more plausible. In this way, the prospects for swamping ASL may be substantially reduced once reasonable levels of option unawareness are incorporated into *ex ante* decisionmaking.[37]

So far, we have met three scope-limiting factors: rapid diminution, washing out, and option unawareness. We saw that these scope-limiting factors are very often present in decisionmaking, and that, when present, they substantially diminish the prospects for swamping ASL. But this may not imply that swamping ASL has empty scope. To see the point, let us return to our discussion of the Space Guard Survey.

[36] In this example, option unawareness determines whether the agent ought to take this specific course of action, but may not change the fact that she should not engage in a demanding calculation of the *ex ante* value of this action. Thanks to an anonymous referee for pushing me to address this point.

[37] Some longtermists have suggested that a potential solution to option unawareness is to fund research aimed at identifying and evaluating new options (Greaves and MacAskill 2021). The viability of this solution depends on the likelihood of identifying valuable options and the values that those options will have. Readers are welcome to qualify the impact of option unawareness using their own views about each quantity. Importantly, as a referee notes, research may be a good solution to 'shallow' forms of option unawareness that can be solved with a few hours of reading through the existing literature. For precisely this reason, my focus in this section is on deeper forms of option unawareness that cannot be resolved so easily.

7 The good case revisited

In Section 3, I showed how a longtermist might take swamping ASL to characterize a decision problem facing Congress in the 1990s: whether to fund the Space Guard Survey, or to redirect the money elsewhere. In support of that suggestion, note that all three of the scope-limiting factors introduced above may be largely absent from this example.

Begin with the problem of rapid diminution: the probabilities of large long-term impacts diminish rapidly. The argument for rapid diminution drew on skepticism about the persistence of short-term effects into the long-term future. It is often hard to make a persisting impact on the long-term future. But it may not be so hard to see how the proposed effects of asteroid detection, namely preventing human extinction, could persist into the long-term future.[38] Not being extinct is a status that can last for a very long time if we play our cards right.

Turn next to the problem of washing out: possible long-term benefits are often significantly cancelled by possible long-term harms. The first argument for washing out drew on evidential paucity: we don't have much evidence about the long-term effects of our actions. But asteroid detection is an area in which we do have significant evidence about possible long-term effects. This includes evidence from past asteroid impacts together with a good scientific understanding of the determinants of asteroid impact force. While there is a good deal that we do not know, existing knowledge is sufficient to build reasonably compelling computational models of impact damages (Stokes et al. 2017).

Our second argument for washing out drew on forecasting skepticism: it is hard to predict the future. First, I argued that in most areas we have no good track record of predicting the far future. But astronomy is one of the few areas in which we have a good track record of predictions on this time-scale. Second, I argued that experts are typically unwilling to make forecasts of the relevant type. But the key forecast driving the example was a prediction by NASA scientists of the probability of catastrophic asteroid

[38] However, if we are pessimistic about current levels of existential risk, as many longtermists are, this point is no longer so clear (Thorstad 2023).

impacts. Third, I argued that due to the multidimensionality of value we may only be able to estimate the probability of a catastrophic impact, but not its value. But where human extinction is concerned, this may pose less of a problem. To evaluate whether preventing human extinction would be a good thing, we must only answer a single question: whether the continued existence of humanity would be a good thing. While answering this question is not straightforward, many theorists are cautiously optimistic that the future will be good (Beckstead 2013; Ord 2020; Parfit 2011).

Turn finally to the problem of option unawareness: decisionmakers are unaware of some options which may be swamping longtermist options. But in the case of the Space Guard Survey, we were already aware of feasible options which could produce the desired results at a reasonable cost. It may well be true that other options, of which we were unaware, would have been still better, but this does not mean that the options ultimately chosen were not swamping longtermist options.

So far, we have seen that the scope-limiting factors may not scuttle the case for swamping ASL in some special cases, for example the decision to fund the Space Guard Survey. That should be unsurprising. We did not expect the scope of swamping ASL to be completely empty, and the Space Guard Survey is an example in which many decisionmakers agreed with the longtermist's evaluative claims. However, the scope-limiting factors begin to significantly threaten the case for swamping ASL in many other decision problems, including some problems touted as friendly to swamping ASL. The next section provides an illustration.

8 Existential risk

Longtermists often argue that humanity faces a great number of existential risks, including risks from engineered pandemics and rogue artificial intelligence (Bostrom 2013; MacAskill 2022; Ord 2020). Our discussion of the Space Guard Survey raises two questions about existential risk mitigation.

First, could our discussion of the Space Guard Survey generalize to show that existential risk mitigation efforts typically avoid many of the challenges raised in this paper? If so, then the scope of longtermism may expand to include many large-scale policy decisions in which resources could be directed towards addressing existential risks.

Second, even a great many mundane decisions may have some effect, however small, on levels of existential risk. For example, I could choose to eat a cheaper breakfast cereal this morning and donate the savings to an organization focused on existential risk mitigation. Hilary Greaves and Christian Tarsney (forthcoming) note one reason it may be best to do just that, namely the *appeal to astronomical stakes*. Because it would be astronomically valuable to prevent an existential catastrophe, just about anything we can do, however slight, to reduce the chance of existential catastrophe may have astronomical expected value.[39] Might the astronomical stakes of existential risk mitigation show that the scope of swamping ASL extends even into mundane decisions such as cereal choice after all?

The second question reveals the importance of challenging the astronomical value of existential risk mitigation efforts. As Greaves and Tarsney note, if direct existential risk mitigation efforts were perhaps ten or twenty orders of magnitude better than any competing option, then the appeal to astronomical stakes would seem to show that the expected value of even mundane decisions is dominated by their impact on existential risk. However, if direct existential risk mitigation efforts were at most a few orders of magnitude better than any competing option, the appeal to astronomical stakes would be less plausible. Recent work by David Thorstad (2023; forthcoming) suggests that under many circumstances, the value of existential risk mitigation may not be astronomical. If that is right, then the appeal to astronomical stakes may not go through.

[39] More generally, Greaves and Tarsney consider the *many levers argument* that there are many variables whose present-day values make non-negligible differences to the expected value of the long-term future, and even most mundane choices have enough influence on at least one such factor to swamp short-term considerations in determining expected values. The arguments from rapid diminution and washing out considered in this paper aim to show why influence on these crucial values may often not swamp expected value calculations, though they might be helpfully supplemented by other arguments, such as skepticism about estimates of the size and value of the future (Thorstad forthcoming).

Let us now turn to the first question: might existential risk mitigation efforts be broadly immune to the challenges raised in this paper? Certainly there are many things that can be said about the value of existential risk mitigation, but it is not clear that many of the existential risk mitigation efforts proposed by leading longtermists do avoid the scope-limiting factors.

Begin with rapid diminution in the probabilities of long-term impacts. There are at least two reasons why many existential risk mitigation efforts may exhibit rapid diminution. The first has the form of a dilemma.[40] If risks are high, then recent work suggests even the successful prevention of an existential catastrophe will not have astronomical value, since other catastrophes are likely to take its place (Thorstad 2023). On the other hand, if risks are low, then it will be difficult to identify and prevent pending catastrophes, and again existential risk mitigation efforts are unlikely to make a large impact.

A second reason why many existential risk mitigation efforts may exhibit rapid diminution is that longtermists take many of the most pressing risks to be driven by technological advances in areas such as artificial intelligence and biosecurity. These are large and well-funded sectors that even wealthy philanthropists may find difficult to shift or even significantly delay. For example, effective altruists invested at least thirty million dollars and many hours of effort into OpenAI, in the hopes that they could steer the industry in a more safety-conscious direction, but longtermists increasingly feel that OpenAI has not been a strong advocate for safety, and when they complained in late 2023 the longtermists found themselves ousted from the board.[41]

Turn next to washing out. Section 5 argued for washing out on the basis of evidential paucity and forecasting pessimism. By contrast to the well-studied case of asteroid risk, more speculative risks such as engineered pandemics and rogue artificial intelligence are significantly less-evidenced and more difficult to forecast, suggesting that both evidential

[40] This is closely related to what Amanda Askell and Sven Neth (forthcoming) call the optimism-pessimism dilemma for longtermism.

[41] See McMillan and Seetharaman (2023) for discussion of the role of effective altruists in the OpenAI board dispute, and the comments section of EA Forum (2023) for a sample of recent effective altruist opinions about OpenAI's safety record.

paucity and forecasting pessimism may get a take on these and similar risks. Further, our discussion of OpenAI reminds us that efforts to mitigate existential risks can backfire, perhaps even increasing the risks we aimed to mitigate. For example, it may turn out that the investment into OpenAI directly funded a company that will develop dangerous AI systems, or that the boardroom fighting at OpenAI cost longtermists status and influence within technology circles.[42] Finally, efforts to mitigate existential risks can have negative effects in other ways: for example, longtermist lobbying for export restrictions on computing chips sent from the United States to China may have exacerbated an already tense relationship between the world's greatest superpowers (Davis 2023).

Turn finally to option unawareness. Many longtermists think that option unawareness has been resolved in the case of existential risk mitigation: we are aware of risks, such as risks from unaligned artificial intelligence, and interventions, such as funding AI safety research, which are extremely valuable to fund at the present margin. However, these claims are controversial and many opponents of longtermism are skeptical both of the risks claimed by longtermists and of the effectiveness of the interventions proposed to mitigate those risks. If that is right, then many opponents will think that existential risk mitigation is a prime example of an area in which no swamping options have been identified.

Summing up, even those sympathetic to the case for existential risk mitigation need not think that the scope of swamping ASL extends even into mundane decision problems. At the same time, while I have not attempted to provide a conclusive refutation of the case for existential risk mitigation, there is good reason to think that many of the scope-limiting factors may get a take on a number of leading proposals for existential risk mitigation advanced by longtermists today.

[42] See Matsakis and Albergotti (2023) for discussion of the impact on longtermism's status within technology circles.

9 Conclusion

This paper assessed the fate of *ex ante* swamping ASL: the claim that the *ex ante* best thing we can do is a swamping longtermist option. I showed how the longtermist might take swamping ASL to hold in some cases, such as the decision to fund the Space Guard Survey. However, I also discussed three *scope-limiting factors* which, when present in a decision problem, substantially reduce the prospects for swamping ASL. These scope-limiting factors included *rapid diminution* in the probabilities of large far-future benefits; *washing out* between possible positive and negative future effects; and *unawareness* of swamping longtermist options. I gave reasons to expect the scope-limiting factors to be present in most human decision problems.

This discussion leaves room for swamping ASL to correctly characterize some cases, particularly when the scope-limiting factors are not present. However, I suggested that the scope of swamping ASL may be far narrower than many longtermists suppose. I used a discussion of existential risk mitigation to illustrate how the scope-limiting factors gain traction even on many cases taken to motivate swamping ASL. I suggested that as the scope-limiting factors make themselves increasingly felt, the prospects for swamping ASL sharply diminish.

Taken together, the scope-limiting factors suggest that swamping ASL is false in many human decision problems. This puts the burden back on longtermists to show, in any given case, that the scope-limiting factors are absent, or else to make a strong case for the truth of swamping ASL despite the scope-limiting factors.

In some ways, a restricted scope for longtermism may be familiar and comforting news. For example, Hilary Greaves (2016) considers the cluelessness problem that we are often significantly clueless about the *ex ante* values of our actions because we are clueless about their long-term effects. Greaves suggests that although cluelessness may correctly describe some complex decisionmaking problems, we should not exaggerate the extent of *mundane cluelessness* in everyday decisionmaking.[43] A natural way of explaining this result would be to argue that in most everyday decisionmaking, it is the expected long-term effects of our actions that are swamped by their short-term effects, and not the other way around. This would mean that cluelessness about long-term effects is often compatible with substantial confidence and precision in our views about the overall values of options.

My discussion leaves room for swamping ASL to be true and important in some contemporary decision problems. It also does not directly pronounce on the fate of ex-post versions of ASL, or on the fate of non-swamping ASL. However, it does suggest that swamping versions of ASL have a substantially more limited scope than many longtermists suppose.

References

Albouy, David. 2012. "The colonial origins of comparative development: An empirical investigation: Comment." *American Economic Review* 102:3059–76.

Albright, Richard. 2002. "What can past technology forecasts tell us about the future?" *Technological Forecasting and Social Change* 69:443–464.

Alesina, Alberto and Giuliano, Paola. 2015. "Culture and institutions." *Journal of Economic Literature* 53:898–944.

Alesina, Alberto, Giuliano, Paola, and Nunn, Nathan. 2011. "Fertility and the plough." *American Economic Review* 101:499–503.

—. 2013. "On the origins of gender roles: Women and the plough." *Quarterly Journal of Economics* 128:469–530.

[43] By 'mundane cluelessness' I take Greaves to be referring to the cluelessness we experience in many cases of mundane decisionmaking. At a minimum, I take Greaves to be arguing that cases of mundane *complex* cluelessness cannot be widespread. This leaves open the possibility of ubiquitous mundane simple cluelessness, which Greaves treats as relatively unproblematic. Thanks to a referee for pressing me to clarify this distinction.

Alvarez, Luis W., Alvarez, Walter, Asaro, Frank, and Michel, Helen V. 1980. "Extraterrestrial cause for the Cretaceous-Tertiary extinction." *Science* 208:1095–1180.

Askell, Amanda. forthcoming. "Longtermist myopia." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*, forthcoming. Oxford University Press.

Baum, Seth, Denkenberger, David, Pearce, Joshua, Robock, Alan, and Winkler, Richelle. 2015. "Resilience to global food supply catatsrophes." *Environment Systems and Decisions* 35:301–13.

Beckstead, Nicholas. 2013. *On the overwhelming importance of shaping the far future*. Ph.D. thesis, Rutgers University.

Benatar, David. 2006. *Better never to have been: The harm of coming into existence*. Oxford University Press.

Bostrom, Nick. 2002. "Existential risks: Analyzing human extinction scenarios and related hazards." *Journal of Evolution and Technology* 9:1–30.

—. 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.

—. 2014. *Superintelligence*. Oxford University Press.

Bradley, Richard. 2017. *Decision theory with a human face*. Cambridge University Press.

Chesley, Steven, Chodas, Paul, Milani, Andreas, Valsecchi, Giovanni, and Yeomans, Donald. 2002. "Quantifying the risk posed by potential Earth impacts." *Icarus* 159:423–32.

Colella, Fabrizio, Lalive, Rafael, Sakalli, Seyhun Orcan, and Thoenig, Mathias. 2019. "Inference with arbitrary clustering." IZA Discussion Paper 12584, https://www.iza.org/publications/dp/12584/inference-with-arbitrary-clustering.

Cotton-Barratt, Owen. 2021. "Everyday longtermism." EA Forum. https://forum.effectivealtruism.org/posts/3PmgXxBGBFMbfg4wJ/everyday-longtermism.

Cowen, Tyler. 2006. "The epistemic problem does not refute consequentialism." *Utilitas* 18:383–99.

—. 2018. *Stubborn attachments*. Stripe Press.

Davis, Donald and Weinstein, David. 2008. "A search for multiple equilibria in urban industrial structure." *Journal of Regional Science* 48:29–62.

Davis, Jacob. 2023. "Longtermists are pushing a new cold war with China." *The Jacobin*, https://jacobin.com/2023/05/longtermism-new-cold-war-biden-administration-china-semiconductors-ai-policy.

Deaton, Angus. 2015. *The great escape: Health, wealth, and the origins of inequality*. Princeton University Press.

Dorsey, Dale. 2012. "Consequentialism, metaphysical realism and the argument from cluelessness." *Philosophical Quarterly* 62:48–70.

EA Forum. 2023. "Sam Altman fired from OpenAI." November 17, 2023, https://forum.effectivealtruism.org/posts/HjgD3Q5uWD2iJZpEN/sam-altman-fired-from-openai.

Freedman, David. 1981. "Some pitfalls in large econometric models: A case study." *Journal of Business* 54:479–500.

Friedrich, Simon. ms. "Causation, cluelessness, and the long term."

Fye, Shannon, Charbonneau, Steven, Hay, Jason, and Mullins, Carie. 2013. "An examination of factors affecting accuracy in technology forecasts." *Technological Forecasting and Social Change* 80:1222–1231.

Gabaix, Xavier and Laibson, David. 2022. "Myopia and discounting." National Bureau of Economic Research Working Paper 23254, https://www.nber.org/papers/w23254.

Garshnek, Victoria, Morrison, David, and Burkle, Frederick. 2000. "The mitigation, management and survivability of asteroid/comet impact with Earth." *Space Policy* 16:213–22.

GiveWell. 2021. “GiveWell’s Cost-Effectiveness Analyses.” https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models.

Goodwin, Paul and Wright, George. 2010. “The limits of forecasting methods in anticipating rare events.” *Technological Forecasting and Social Change* 77:355–368.

Greaves, Hilary. 2016. “Cluelessness.” *Proceedings of the Aristotelian Society* 116:311–39.

Greaves, Hilary and MacAskill, William. 2019. “The case for strong longtermism.” Global Priorities Institute Working Paper 7-2019.

—. 2021. “The case for strong longtermism.” Global Priorities Institute Working Paper 5-2021, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

Greaves, Hilary and Tarsney, Christian. forthcoming. “Minimal and expansive longtermism.” In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, forthcoming. Oxford University Press.

Greaves, Hilary, Thorstad, David, and Barrett, Jacob (eds.). forthcoming. *Essays on longtermism*. Oxford University Press.

Hedden, Brian. 2012. “Options and the subjective ought.” *Philosophical Studies* 158:343–360.

Isaacs, Yoaav. 2016. “Probabilities cannot be rationally neglected.” *Mind* 125:759–62.

Jaynes, Edwin. 1957. “Information theory and statistical mechanics.” *Physical Review* 106:620–30.

Karni, Edi and Vierø, Marie-Louise. 2013. “‘Reverse Bayesianism’: A choice-based theory of growing awareness.” *American Economic Review* 103:2790–2810.

Kelly, Morgan. 2019. “The standard errors of persistence.” CEPR Discussion Papers 13783, https://ideas.repec.org/p/cpr/ceprdp/13783.html.

Koon, Justis. 2020. "Options must be external." *Philosophical Studies* 177:1175–89.

Kott, Alexander and Perconti, Phillip. 2018. "Long-term forecasts of military technologies for a 20-30 year horizon: An empirical assessment of accuracy." *Technological Forecasting and Social Change* 137:272–9.

Lenman, James. 2000. "Consequentialism and cluelessness." *Philosophy and Public Affairs* 29:342–70.

Leslie, John. 1996. *The end of the world: The science and ethics of human extinction*. Routledge.

MacAskill, William. 2022. *What we owe the future*. Basic books.

Makridakis, Spyros and Taleb, Nassim. 2009. "Decision making and planning under low levels of predictability." *International Journal of Forecasting* 25:716–733.

Marchau, Vincent, Walker, Warren, Bloemen, Pieter, and Popper, Steven (eds.). 2019. *Decision making under deep uncertainty*. Springer.

Matsakis, Louise and Albergotti, Reed. 2023. "The AI industry turns against its favorite philosophy." *Semafor*, November 21, 2023, https://www.semafor.com/article/11/21/2023/how-effective-altruism-led-to-a-crisis-at-openai.

McMillan, Robert and Seetharaman, Deepa. 2023. "How a fervent belief split Silicon Valley — and fueled the blowup at OpenAI." *Wall Street Journal*, November 6, 2023, https://www.wsj.com/tech/ai/openai-blowup-effective-altruism-disaster-f46a55e8.

Miguel, Edward and Roland, Gérard. 2011. "The long-run impact of bombing Vietnam." *Journal of Development Economics* 96:1–15.

Mogensen, Andreas. 2021. "Maximal cluelessness." *Philosophical Quarterly* 71:141–62.

Monton, Bradley. 2019. "How to avoid maximizing expected utility." *Philosophers' Imprint* 19:1–25.

Mullins, Carie. 2018. “Retrospective analysis of long-term forecasts.” Open Philanthropy Project Technical Report, https://www.openphilanthropy.org/files/Blog/Mullins_Retrospective_Analysis_Longterm_Forecasts_Final_Report.pdf.

Newberry, Toby. 2021. “How cost-effective are efforts to detect near-Earth-objects?” Global Priorities Institute Technical Report T1-2021, https://globalprioritiesinstitute.org/how-cost-effective-are-efforts-to-detect-near-earth-objects-toby-newberry-future-of-humanity-institute-university-of-oxford/.

Noy, Ilan and Noy, Shakked. forthcoming. “The short-termism of ‘hard’ economics.” In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, forthcoming. Oxford University Press.

Nunn, Nathan. 2008. “The long term effects of Africa’s slave trades.” *Quarterly Journal of Economics* 123:139–176.

—. 2020. “The historical roots of economic development.” *Science* 367:eaaz9986.

Nunn, Nathan and Wantchekon, Leonard. 2011. “The slave trade and the origins of mistrust in Africa.” *American Economic Review* 3221–3252.

Ord, Toby. 2020. *The precipice*. Bloomsbury.

Paine, Michael. 1999. “Asteroid impacts: The extreme hazard due to tsunami.” *Science of Tsunami Hazards* 17:155–66.

Parente, Rick and Anderson-Parente, Janet. 2011. “A case study of long-term Delphi accuracy.” *Technological Forecasting and Social Change* 78:1705–1711.

Parfit, Derek. 2011. *On what matters*, volume 1. Oxford University Press.

Pettigrew, Richard. 2016. “Accuracy, Risk, and the Principle of Indifference.” *Philosophy and Phenomenological Research* 92:35–59.

—. 2022. “Effective altruism, risk, and human extinction.” Global Priorities Institute Working Paper 2-2022, https://globalprioritiesinstitute.org/effective-altruism-risk-and-human-extinction-richard-pettigrew-university-of-bristol/.

Posner, Richard. 2004. *Catastrophe: Risk and response*. Oxford University Press.

Ranger, Nicola, Reeder, Tim, and Lowe, Jason. 2013. “Addressing ‘deep’ uncertainty over long-term climate in major infrastructure projects: Four innovations of the Thames Estuary 2100 project.” *EURO Journal on Decision Processes* 1:233–262.

Risi, Joseph, Sharma, Amit, Shah, Rohan, Connelly, Matthew, and Watts, Duncan. 2019. “Predicting history.” *Nature Human Behavior* 3:906–912.

Sanchez, Joan Pau, Colombo, Camilla, Vasile, Massimiliano, and Radice, Gianmarco. 2009. “Multicriteria comparison among several mitigation strategies for dangerous near-earth objects.” *Journal of Guidance, Control and Dynamics* 32:121–42.

Schulte, Peter et al. 2010. “The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary.” *Science* 327:1214–8.

Schulz, Jonathan F., Bahrami-Rad, Duman, Beauchamp, Jonathan, and Henrich, Joseph. 2019. “The Church, intensive kinship, and global psychological variation.” *Science* 36:eaau5141.

Schwitzgebel, Eric. forthcoming. *The weirdness of the world*. Princeton University Press.

—. ms. “The washout argument against longtermism.” Unpublished manuscript.

Sevilla, Jaime. 2021. “Persistence: A critical review.” Technical report, Forethought Foundation.

Shiller, Derek. 2021. “Chance and the dissipation of our acts’ effects.” *Australasian Journal of Philosophy* 99:334–48.

Shulman, Carl and Thornley, Elliott. forthcoming. "How much should governments pay to prevent catastrophes? Longtermism's limited role." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Essays on longtermism*. Oxford University Press.

Smith, Nicholas. 2014. "Is evaluative compositionality a requirement of rationality?" *Mind* 123:457–502.

Sotala, Kaj and Gloor, Lukas. 2017. "Superintelligence as a cause or cure for risks of astronomic suffering." *Informatica* 41:389–400.

Steele, Katie and Stefánsson, Orri. 2021. *Beyond uncertainty*. Cambridge University Press.

Stokes, Grant, Barbee, Brent, Bottke, William, et al. 2017. "Update to determine the feasibility of enhancing the search and characterization of NEOs: Report of the near-earth object science definition team." NASA Technical Report, https://cneos.jpl.nasa.gov/doc/2017_neo_sdt_final_e-version.pdf.

Tetlock, Philip, Karvetski, Christopher, Satopää, Ville, and Chen, Kevin. ms. "Long-range subjective-probability forecasts of slow-motion variables in world politics: Exploring limits on expert judgment." SSRN, https://ssrn.com/abstract=4377599.

Thorstad, David. 2023. "High risk, low reward: A challenge to the astronomical value of existential risk mitigation." *Philosophy and Public Affairs* 51:373–412.

—. forthcoming. "Mistakes in the moral mathematics of existential risk." *Ethics* forthcoming.

Toon, Owen, Zahnle, Kevin, Morrison, David, Turco, Richard, and Covey, Curt. 1997. "Environmental perturbations caused by the impacts of asteroids and comets." *Reviews of Geophysics* 35:41–78.

Torres, Phil. 2018. "Space colonization and suffering risks: Reassessing the 'maxipok rule'." *Futures* 100:74–85.

Voth, Hans-Joachim. 2021. "Persistence — myth and mystery." In Alberto Bisin and Giovanni Federico (eds.), *Handbook of historical economics*, 243–67. Academic Press.

White, Roger. 2010. "Evidential symmetry and mushy credence." *Oxford Studies in Epistemology* 3:161–86.

Wilkinson, Hayden. 2022. "In defense of fanaticism." *Ethics* 132:445–77.

Yusuf, Moeed. 2009. "Predicting proliferation: the history of the future of nuclear weapons." Brookings Institution Policy Paper 11, https://www.brookings.edu/research/predicting-proliferation-the-history-of-the-future-of-nuclear-weapons/.

Zambrano-Marin, L.F., Howell, E.S., Devogéle, M., et al. 2021. "Radar observations of near-earth asteroid 2019 OK." In *Proceedings of the 52nd Lunar and Planetary Science Conference 2021*, LPI Contribution Number 2548.