The scope of longtermism

David Thorstad

Global Priorities Institute | June 2021

GPI Working Paper No. 6-2021
The scope of longtermism

Abstract

Longtermism holds roughly that in many 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 longtermism holds? Although longtermism was initially developed to describe the situation of cause-neutral philanthropic decisionmaking, it is increasingly suggested that longtermism holds in many or most decision problems that humans face. By contrast, I suggest that the scope of longtermism may be more restricted than commonly supposed. After specifying my target, swamping axiological strong longtermism (swamping ASL), I give two arguments for the rarity thesis that the options needed to vindicate swamping ASL in a given decision problem are rare. I use the rarity thesis to pose two challenges to the scope of longtermism: the area challenge that swamping ASL often fails when we restrict our attention to specific cause areas, and the challenge from option unawareness that swamping ASL may fail when decision problems are modified to incorporate agents’ limited awareness of the options available to them.

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 experts 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).

Many 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 2019; Mogensen forthcoming; Ord 2020; MacAskill 2022). The scope question for longtermism asks: how large is the class of decision situations for which longtermism holds?

Longtermism was originally developed to describe the decisions facing present-day, cause-neutral philanthropists. Longtermists suggest that the best thing philanthropists can do today is to safeguard the long-term future, for example by mitigating risks of human extinction. But many have held that the scope of longtermism extends considerably beyond cause-neutral philanthropic decisionmaking. Hilary Greaves and Will MacAskill (Greaves and MacAskill 2019) suggest that the cause-specific choice between two anti-malaria programs should be governed, not by their direct effects of preventing present-day
malaria deaths, but by the potential far-future impact of each intervention.\footnote{Relatedly, Mogensen (2021) suggests that it may not be clearly better for a longtermist to donate to the Against Malaria Foundation rather than the Make-A-Wish Foundation due to uncertainty about long-run impacts.} And Owen Cotton-Barratt 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 (Cotton-Barratt 2021).

In this paper, I argue that the scope of longtermism may be narrower than often supposed. Section 2 clarifies my target: swamping axiological strong longtermism (swamping ASL). Section 3 argues that swamping ASL may be true in the case of present-day, cause-neutral philanthropy. However, Sections 4-5 give two arguments for the rarity thesis that longtermist options of the type needed to vindicate swamping ASL are rare. Section 6 argues that the rarity thesis poses no challenge to swamping ASL in the special case of present-day, cause-neutral philanthropy. However, Sections 7-8 use the rarity thesis to put two limits on the scope of swamping ASL. Section 7 poses the area challenge: swamping ASL fails in many specific cause areas, such as choosing between anti-malaria programs. Section 8 poses the challenge from option unawareness: swamping ASL often fails when decision problems are modified to incorporate our limited ex ante awareness of the options available to us. 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. 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):

\[(\text{ASL}) \text{ In a wide class of decision situations, the option that is ex ante best is contained in a fairly small subset of options whose ex ante effects on the very long-run future are best. (Greaves and MacAskill 2019).}\]

My target in this paper will be a restricted form of ASL.
2.2 Swamping axiological longtermism

There are two ways in which ASL could be true. First, there might be a swamping longtermist option whose expected long-term benefits exceed in magnitude the expected short-term effects produced by any option.\(^2\) I call these swamping longtermist options because their long-term effects begin to swamp short-term considerations in determining ex ante value. We will see in Section 3 that ASL holds in many decision problems involving swamping longtermist options, so the first way that ASL could be true would be if it held in virtue of swamping longtermist options.

Swamping axiological strong longtermism (Swamping ASL) In a wide class of decision situations, the option that is ex ante best (a) is a swamping longtermist option, and (b) is contained in a fairly small subset of options whose ex ante effects on the very long-run future are best.

My focus in this paper will be on swamping ASL.

Second, the best option may be a convergent option, an option whose long-term value is near-best, and whose short-term value is at least modestly comparable to the best-achievable short-term values.\(^3\) If there are few swamping longtermist options, we could defend ASL through the convergence thesis that the best options are often convergent options. For example, you might think that the best thing we can do 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.

I focus on swamping ASL over the convergence thesis 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. Third, swamping ASL underlies many of the most persuasive arguments from axiological to deontic longtermism, which rely on the claim that sufficiently strong duties to promote impartial value may trump competing non-consequentialist 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 The rarity thesis

In this paper, I argue for the rarity thesis that swamping longtermist options are rare, in the sense that the vast majority of the options that we face as decisionmakers are not swamping longtermist options. The rarity thesis does not imply that swamping ASL fails

---

\(^2\)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 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).

\(^3\)For the purposes of this paper, the best option will be a convergent option if it is not a swamping longtermist option. We may wish to expand this taxonomy to capture gradations between swamping and convergent ASL, but that will not be my focus here.
in all decision situations. Indeed, I argue in Section 3 that swamping ASL is plausible in the case of present-day cause-neutral philanthropy.

However, the rarity thesis does suggest that the scope of swamping ASL may be more limited than is often supposed. I use the rarity thesis to pose two challenges to the scope of swamping ASL: the area challenge that swamping ASL fails in many specific cause areas of interest, and the challenge from option unawareness that swamping ASL fails in many decision problems when those problems are modified to account for decisionmakers’ unawareness of relevant options.

Summing up, my target in this paper is ex ante, axiological, swamping strong longtermism. I use the rarity thesis to suggest that swamping ASL has more limited scope than we might otherwise suppose. But first, let us consider where ASL may be plausible.

3 The good case: Present-day cause-neutral philanthropy

Many effective altruists have claimed that swamping ASL is true in the case of present-day, cause-neutral philanthropic decisionmaking. In this section, I argue that they may well be right.

Let a strong swamping option be one whose expected far-future impact exceeds the best-achievable expected short-term impact by a large margin — say, for concreteness, by a factor of ten.\(^4\) We can argue that swamping ASL holds for present-day philanthropy using the argument from strong swamping.\(^5\) This argument claims first, that there exists at least one strong swamping option for present-day philanthropists, and second that the existence of a strong swamping option in a decision problem implies swamping ASL for that problem.

Here is an argument for the second premise. Assume there exists a strong swamping option. Note first that the best option must be a swamping longtermist option, since a strong swamping option is better than any non-swamping option. Note next that the best option must have at least 9/10 of the best-achievable long-term value, since the existence of a strong swamping option implies that short-term value differences cannot make up any larger gap than this. Hence the best option is a swamping longtermist option with near-best long-term effects, and swamping ASL follows.

The weight of the argument from strong swamping is carried by the first premise, which asserts the existence of a strong swamping option. In the case of present-day philanthropy, we can argue for the existence of strong swamping options using the argument from single-track dominance. For a given option, the argument from single-track dominance makes three claims.

First, it claims that the option exhibits:

\textbf{(Single-track dominance)} The lion’s share of the option’s expected impact on the long-term future is driven by a single causal pathway or effect.

For example, efforts to reduce extinction risk plausibly exhibit single-track dominance along the pathway of preventing human extinction. Single-track dominance allows us

\(^4\)That is, \(E[\Delta L_o] > 10 \times \max_{o' \in O} |E[\Delta S_{o'}]|\).

\(^5\)This argument is largely drawn from Greaves and MacAskill (2019).
to simplify the overwhelming array of future effects by considering harms and benefits along a single track.

Second, the option may produce overwhelmingly positive far-future benefits along this track:

(Nontrivial probability of significant benefit) For some extremely large value of $N$, the probability that this option produces long-term benefits of value exceeding $N$ is nontrivial, considering only effects along the specified track.\(^6\)

For example, we might require that $N$ is the value of saving a billion future lives and that the probability of producing long-term benefits in excess of $N$ is at least one in a million.

Third, the option is much less likely to produce overwhelmingly large future harms along this track:

(Relatively trivial probability of significant harm) With $N$ as before, the probability that this option produces long-term harms of value beneath $-N$ is significantly lower than the probability of producing benefits with value exceeding $N$, considering only effects along the specified track.

Most options with these three features will be strong swamping options.\(^7\)

To show that ASL holds for present-day, cause-neutral philanthropy, it remains to argue that some options available to present-day philanthropists are swamping longtermist options in virtue of meeting the conditions set out in the argument from single-track dominance. There are many detailed arguments alleging that these conditions are met by specific philanthropic options (Bostrom 2013; Tarsney ms; Ord 2020). Here is one such argument (Greaves and MacAskill 2019; Newberry 2021a; Ord 2020).

One way that humans might go extinct is through the impact of a large asteroid on earth. 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).

NASA classifies asteroids with diameter greater than 1 kilometer as catastrophic, capable of causing a global calamity or even mass-extinction. Our best estimates suggest that such impacts occur on earth about once in every 6,000 centuries (Stokes et al. 2017). Plausibly, it is worth our while to detect and prepare for such events.

Since the 1990s, the SpaceGuard survey has mapped approximately 95% of the near-earth asteroids with diameters exceeding 1 km at a cost of $70 million. Assuming a conservative chance of one in a million that an asteroid impact of this magnitude would result in human extinction, this project would, in expectation, be worth funding based on its short-term value alone if we value early warning of an extinction-causing impact within the next century at at least $74 trillion.

---

\(^6\)I.e. for specified $N$ large and $k$ nontrivial, $Pr(\Delta L_o > N) > k$, ignoring off-track effects.

\(^7\)Strictly speaking, it is possible for an option with these features to fail to be a swamping longtermist option, for example if for some $M > N$ we have $Pr(\Delta L_o > M) < Pr(\Delta L_o < -M)$. I hope it is clear how the requirements might be tightened to avoid this worry, but also that this tightening would be a distraction in most cases of interest, where this is not a live worry.
That is a close call, and you might well be skeptical of the short-term value of such a project. But now, consider that estimates of the expected number of future humanlike lives range from about $10^{13}$ to $10^{55}$ (Bostrom 2014; Newberry 2021b). This puts the project’s expected cost of detecting an extinction-causing asteroid impact, counting only impacts within the next century, at no more than about $7 per expected future life, and at fractions of a penny using anything but the most conservative estimate. For comparison, our best estimates put the cost of saving a life through short-termist interventions at several thousand dollars (GiveWell 2021), making the SpaceGuard survey a prime example of a strong swamping option, with expected long-term value well in excess of the best expected short-term impacts if we have any confidence at all in our ability to prepare for and survive an otherwise-catastrophic impact with sufficient warning.

Now the SpaceGuard survey has already been funded, but there is doubtless more that could be done to reduce extinction risks from asteroid impacts, such as mapping the remaining 5% of large near-earth asteroids or developing contingency plans for deflecting asteroids and surviving asteroid impacts. And many other interventions aimed at reducing existential risks have been claimed to be yet more cost-effective than these (Ord 2020). So I do not want to deny the existence of strong swamping options, and hence I do not deny that swamping ASL is plausible in the special case of present-day cause neutral philanthropy. However, as we move away from the case of cause-neutral philanthropy, matters begin to look different.

4 Rapid diminution

In the next two sections, I defend the rarity thesis that swamping longtermist options are rare. Sections 7-9 then assess the impact of the rarity thesis on the scope of swamping ASL.

The first argument for the rarity thesis is the argument from rapid diminution. Fix an option $o$ and consider the probability distribution over long-term impacts of $o$.$^8$ 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. 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 on the expected long-term impact of $o$.

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. This argument is supported by persistence

---

$^8$I.e. consider the probability distribution over the partition $\{[\Delta L = k] : k \in \mathbb{R}\}$. 

6
skepticism: the view that many 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). Persistence studies often returns surprising negative results, where effects that we might have expected to persist for a long time evaporate after several decades. For example, given the scale of American bombing in Japan and Vietnam, one might expect persistent economic effects in the heaviest-hit areas. 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 in fact, the effects on population size, poverty rates and consumption patterns appear to have already vanished (Davis and Weinstein 2008; Miguel and Roland 2011).

Now it is true that persistence studies has identified a few-dozen effects which might be more persistent. For example, the introduction of the plough may have affected fertility norms and increased the gendered division of labour (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.

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. Hence the persistence literature may not provide strong support for the swamping longtermist’s hope that persistent long-term effects could swamp short-term effects in importance.

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 the argument from rapid diminution with a second argument: the argument from washing out.

5 Washing out

The second argument for the rarity thesis is the argument from washing out. 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 be significant cancellation between possible positive and negative effects in determining the expected values of options.

There are two related ways that the argument from washing out can be articulated. The first begins with the popular Bayesian idea that complete ignorance about the long-term value of an option should be represented by a symmetric prior distribution over possible
long-term values. 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 over long-term impacts gives significant cancellation when we take expected values.

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 generates a type of washing-out between possible positive and negative forecasts. When we make forecasts based on sparse data, we need to take account of the fact that the data we have been dealt is a noisy reflection of the underlying reality. As phenomena become more unpredictable and our data becomes increasingly sparse, we should grow more willing to chalk up any apparent directionality in our forecasts to noisiness in the hand of data that nature has dealt us. In other words, as forecasting becomes more difficult we get increasing wash-out between possible positive and negative forecasts that we could have made based on different samples of data.

Why should we be pessimistic about our ability to 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. 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 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. Very few economists, risk analysts, and other experts are willing to make such predictions, citing the unavailability of data, a lack of relevant

---

9 Among the many ways to give formal expression to this idea, Gabaix and Laibson’s (2021) as-if discounting brings out the similarity to the argument from evidential paucity by highlighting the role of priors.

10 For domain-specific track records see Albright (2002); Kott and Perconti (2018); Parente and Anderson-Parente (2011); Risi et al. (2019) and Yusuf (2009). For discussion see Fye et al. (2013) and Mullins (2018).
theoretical models, and the inherent unpredictability of underlying systems (Freedman 1981; Goodwin and Wright 2010; Makridakis and Taleb 2009). 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, 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 prior 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.

Together, the arguments from washing out and rapid diminution suggest that large long-term impacts may be less probable than otherwise thought, and may be significantly cancelled by potential negative long-term impacts. This grounds the rarity thesis that swamping longtermist options are rare. But this does not mean that there are no swamping longtermist options. Plausibly, the case of cause-neutral philanthropy remains untouched by these arguments for the rarity thesis.

6 The good case revisited

In the case of present-day cause-neutral philanthropy, the argument from single-track dominance allows us to avoid all of the arguments raised in the previous two sections. We can see this both abstractly, by considering the argument from single-track dominance itself, and concretely, by thinking about the case of asteroid detection.

Begin with the problem of rapid diminution: the probabilities of large long-term impacts diminish rapidly. The second premise of the argument from single-track dominance avoids this problem by holding that the option in question has nontrivial probability of producing a highly significant benefit. When this is true, rapid diminution elsewhere
in the probability distribution will not threaten the status of the option as a swamping longtermist option. In fact, rapid diminution may even help to secure the third premise of the argument from single-track dominance: that the option in question has comparatively negligible probability of significant future harm. The general phenomenon of rapid diminution may give us good reason to accept this third premise as a default stance, unless given special reason to doubt it.

More concretely, the argument for rapid diminution drew on skepticism about the persistence of short-term effects into the long-term future. But it is not hard to see how the proposed effects of asteroid detection, namely preventing human extinction, could persist into the long-term future. 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 may be significantly cancelled by possible long-term harms. The second and third premises of the argument from single-track dominance avoid this problem by claiming that probabilities of very large benefits significantly outweigh the probabilities of very large harms.

More concretely, 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 understanding of the determinants of asteroid impact force, which is sufficient to build compelling computational models of impact damages (Stokes et al. 2017).

The second argument for washing out drew on forecasting skepticism: it’s hard to predict the future. First, I argued that in many areas we have no good track record of predicting the far future. But astronomy is one of the few areas in which we do have a good track record of predictions on this time-scale. Second, I argued that experts are often unwilling to make forecasts of the relevant type. But the key forecast driving the example was a prediction by NASA scientists that the per-century probability of a catastrophic asteroid impact is approximately 1/6,000. 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. This may be true for off-track effects, such as catastrophic impacts leading to non-extinction events, but it is not a significant problem if the outcome in question is human extinction. 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 of us are cautiously optimistic that the future will be good (Beckstead 2013; Ord 2020; Parfit 2011).

So far, we have seen that the argument for single-track dominance allows us to accept the rarity thesis without questioning the truth of swamping ASL in the special case of present-day, cause-neutral philanthropy. However, I argue in the next two sections that the rarity thesis does pose two challenges to the scope of swamping ASL.

7 The area challenge

If we accept that swamping ASL holds in present-day philanthropy, then we should accept swamping ASL in some philanthropy-adjacent matters as well. For example, swamping
ASL might govern career choice by indicating that we have strong reason to pursue global priorities research. But how far does the reach of swamping ASL extend? The area challenge to the scope of swamping ASL holds that swamping ASL fails in many cause areas sufficiently removed from cause-neutral philanthropy.

To see how quickly the area challenge gains bite, consider a problem involving cause-specific philanthropic giving: a donor is committed to funding anti-malaria work, but wants to do the most good possible in this space. Greaves and MacAskill (2019) suggest that ASL governs this choice as well, because the lion’s share of the expected value-difference between competing malaria charities will be driven by their differential impacts on the long-term future. While I don’t want to pronounce on the prospects for convergent versions of ASL in this case, I think that the prospects for swamping ASL in this case are relatively poor.

Certainly nothing like the argument from single-track dominance can be made here. We would not want to accept the first premise: that the lion’s share of the expected long-run value of an anti-malarial program is driven by a single type of far-future effect. What the longtermist wants to stress here is precisely the opposite, that there are many different ways in which anti-malarial work can affect the far future, for example by changing the international balance of power, speeding technological growth and space settlement, or modulating population size, each of which makes an important contribution to the program’s far-future value. Removing the commitment to single-track dominance, we may or may not want to accept the second premise, that an anti-malarial program has nontrivial probability of significant far-future benefit. But most natural arguments for the second premise would also be arguments against the third premise, which asserts that anti-malarial programs have comparatively trivial probabilities of significant far-future harm.

As the premises of the argument from single-track dominance begin to fail, the arguments driving the rarity thesis gain traction against swamping ASL. Begin with the argument from washing-out. The first argument for washing-out cited evidential paucity: we have little concrete evidence about the long-term impacts of many actions. I think we should concede that this is the case with anti-malarial work. Although malaria prevention could have some impact on the very long-run balance of international power, for example, we don’t have much to go on in determining what that impact might be.

The second argument for washing-out cited forecasting pessimism: it is hard to predict the long-term impacts of our actions. Here all three of the arguments made for forecasting pessimism gain traction. As far as track records are concerned, we have no sizable track record of predicting the effects of global health programs on a scale of centuries or millennia. As far as practitioner skepticism is concerned, no existing anti-malarial organization attempts to estimate its very long-run impacts. Nor, for that matter, does the GiveWell foundation, a philanthropic organization which steers millions of dollars in grants towards anti-malarial programs, conducts extensive impact evaluations on its grantees, and is more sympathetic to longtermism than almost any other major philanthropic actor. So it looks as though practitioners take forecasting the long-run impacts of anti-malarial work to be very difficult. Finally, the multidimensionality of value is a strong obstacle in this area: in fact, anti-malarial work was the example that I used to illustrate the multidimensionality of value. Although we are relatively good at forecasting immediate health outcomes such as malaria deaths prevented, we have little experience
in, or methodologies for predicting the overall value impact of these programs at long
time scales. This matters because global health programs have many diverse sociopolitical
impacts that determine their long-run value.

The argument from rapid diminution also gains traction against swamping ASL here. The
motivation for rapid diminution was skepticism about the persistence of short-term
effects into the long-term future. If we are generally skeptical about long-term persistence,
then we should demand some special reason to think that anti-malarial work is likely to
have persistent effects on the far-future even though other actions often do not. And when
the effects that we point to are quite speculative future possibilities, such as an accelerated
pace of interstellar expansion due to regional economic growth, it is not clear what could
be said in favor of expecting such persistent effects from anti-malarial work that would not
tell against persistence skepticism more generally. Absent a detailed story, the persistence
skeptic is likely to remain unconvinced.

In this section, we have seen that the rarity thesis poses an area challenge to the scope
of swamping ASL. Although swamping ASL may be true in cause-neutral philanthropy
and some adjacent areas, in many other cause areas including many types of cause-specific
philanthropy, swamping ASL may fail. In the next section, I pose a second scope challenge
to swamping ASL: the challenge from option unawareness.

8 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
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 and you
can hardly be blamed for that. 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). In the present case, this means that your options are as they
first described them: turning right, turning left, or stopping to fight. Unless, perhaps, you
happen to be James Bond.

Now option unawareness does not pose any problem for swamping ASL in the case of
cause-neutral philanthropy. That is because the philanthropist is already aware of some
options, such as asteroid detection, which witness the truth of swamping ASL. Moreover,
if the philanthropist were not aware of any such options, then it might turn out that the
best thing she could do would be to fund option generation research (Kalis et al. 2013; Greaves and MacAskill 2019). If this research had a high probability of turning up a swamping longtermist option, then it could itself be a swamping longtermist option, and again swamping ASL would be saved.

But outside the case of cause-neutral philanthropy, option unawareness combines with the rarity thesis to generate a challenge for the scope of ASL. By the rarity thesis, swamping longtermist options are rare. Many theories of option unawareness hold that, in typical choice problems, we face modestly-sized choice sets with at most dozens of options. By rarity, most modestly-sized choice sets will not contain any swamping longtermist options, so swamping ASL turns out to have fairly restricted scope.

We can see this problem in context by thinking about the cause-specific philanthropist, who is committed to funding some anti-malarial program. Naively, we might pose her decision problem as follows: of any possible anti-malarial intervention, which should I take? And it is not so implausible that swamping ASL could hold in awareness-unconstrained problem. From an ex post perspective, there is surely some option that would bring enormous far-future benefits, such as protecting a village or region that contains an important future politician or inventor. And from an ex ante perspective, it is not so implausible that there is enough evidence to identify one such region after years of calculation. But most philanthropists face an awareness-restricted problem such as the following: out of the internationally prominent existing anti-malaria programs, which should I fund? We saw in Section 7 that the arguments for rarity suggest that none of these programs may be a swamping longtermist option.

Unlike the case of cause-neutral philanthropy, we cannot save swamping ASL by appealing to meta-options such as conducting research into the availability of new options or gathering evidence about their values. That is because even if new options could in principle be discovered, in practice it might be quite expensive to identify and evaluate them. For comparison, one of the largest anti-malaria charities, the Against Malaria Foundation, has a budget in the tens of millions of dollars. But it would likely cost far more than that to scour the globe in search of rare future talent worth protecting. So the cost of expanding awareness here would approach or exceed the cost of simply funding malaria prevention across the globe. High costs of option-generation might be bearable in larger cause areas, such as existential risk prevention, which can bear a great deal of capital investment once discovered. But even a relatively high cap on capital needs may make longtermist option-generation into a wastefully suboptimal option.

Summing up, the challenge from option unawareness holds that ex ante rational choice is restricted to the options of which we are relevantly aware. Because option unawareness often considerably restricts the size of our choice sets, the rarity thesis implies that in most choice problems we face as decisionmakers, swamping ASL fails.

9 Conclusion

This paper assessed the fate of ex ante swamping ASL: the claim that the ex ante best thing we can do is often a swamping longtermist option that is near-best for the long-term future. I gave a two-part argument that swamping ASL holds in the special case of present-day cause-neutral philanthropy: the argument from strong swamping that a strong swamping
option would witness the truth of ASL, and the argument from single-track dominance for the existence of a strong swamping option.

However, I also argued for the rarity thesis that swamping longtermist options are rare. I gave two arguments for the rarity thesis: the argument from rapid diminution that probabilities of large far-future benefits often diminish faster than those benefits increase; and the argument from washing out that probabilities of far-future benefits are often significantly cancelled by probabilities of far-future harms.

I argued that the rarity thesis does not challenge the case for swamping ASL in present-day, cause-neutral philanthropy, but showed how the rarity thesis generates two challenges to the scope of swamping ASL beyond this case. First, there is the area challenge that swamping ASL often fails when we restrict our attention to specific cause areas. Second, there is the challenge from option unawareness that swamping ASL often fails when we modify decision problems to incorporate agents’ unawareness of relevant options.

In some ways, this may be familiar and comforting news. For example, 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 be correct as a description of some complex decisionmaking problems, we should not exaggerate the extent of mundane cluelessness in everyday decisionmaking. A natural way of explaining this result would be to argue for a strengthened form of the rarity thesis on which in most mundane decisionmaking, the expected long-term effects of our actions are swamped by their expected short-term effects. So in a sense, the rarity thesis is an expected and comforting result.

In addition, this discussion leaves room for swamping ASL to be true and important in the case of present-day, cause-neutral philanthropy as well as in a limited number of other contexts. It also does not directly pronounce on the fate of ex-post versions of ASL, or on the fate of non-swamping, convergent ASL. However, it does suggest that swamping versions of ASL may have a more limited scope than otherwise supposed.

References


