Minimal and Expansive Longtermism

Hilary Greaves (University of Oxford) and Christian Greaves (Population Wellbeing Initiative, University of Texas at Austin)

Global Priorities Institute | March 2023

GPI Working Paper No . 3-2023

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Minimal and Expansive Longtermism

Hilary Greaves and Christian Tarsney

January 2023

[Introduction](#)

[How many interventions?](#)
- [Indirect existential risk mitigation](#)
- [Patient philanthropy](#)
- [Accelerating growth](#)
- [Space settlement](#)
- [Improving moral values and institutions](#)
- [Taking stock](#)

[How many decision situations?](#)

[How many resources?](#)

[Summary and conclusions](#)

Abstract

The standard case for longtermism focuses on a small set of risks to the far future, and argues that in a small set of choice situations, the present marginal value of mitigating those risks is very great. But many longtermists are attracted to, and many critics of longtermism worried by, a farther-reaching form of longtermism. According to this farther-reaching form, there are many ways of improving the far future, which determine the value of our options in all or nearly all choice situations, and will continue to do so over the coming decades even if we make substantial investments in longtermist priorities. This chapter highlights the gap between the minimal form of longtermism established by standard arguments and this more expansive view, and considers (without reaching any firm conclusions) which form of longtermism is more plausible.

Keywords: longtermism, existential risk, minimal longtermism, expansive longtermism

Introduction

Strong longtermism (hereafter, simply *longtermism*) is, roughly, the thesis that "impact on the far future is the most important feature of our actions today" (see Beckstead 2013: 1-3; Beckstead 2019: 80; **Greaves and MacAskill 2023**:[1]). The present chapter examines the question of how expansive the most plausible form of longtermism is. Roughly for now, the

[1] The quote is from **Greaves and MacAskill (2023)**. Beckstead's thesis is that "what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, or trillions of years" (2013: 1; 2019: 80).

issue we see is that (i) the standard argument for longtermism implies only a rather minimal form of the thesis, (ii) many authors and real-world agents subscribe to a far more expansive version of longtermism, (iii) the difference between the two has not previously been in sharp focus, and (iv) for various reasons, the difference between the minimal and more expansive versions of longtermism matters. The aims of our essay are to highlight the differences between "minimal" and more "expansive" forms of longtermism, and to conduct some exploration of what arguments for the more expansive forms might look like (without either endorsing or rejecting those arguments).

Before explaining this distinction between minimal and expansive longtermism, let us rehearse the "standard argument" for longtermism alluded to above.[2] In this standard argument, the motivation for longtermism comes from considering what we will call *technological existential risks* (hereafter, *technological x-risks*). An *existential risk* (or *x-risk*) is a risk of an *existential catastrophe*, that is, either premature human extinction or another irreversible outcome that is similarly bad to extinction.[3] A *technological x-risk* is an existential risk resulting from advanced (present or future) technology.

The standard argument for longtermism then runs as follows.

First, technological x-risks are at worryingly high levels. Areas of concern include, for example, nuclear weapons, artificial intelligence and biotechnology. For the first time in history, we are creating technologies that could destroy humanity's entire future (see Rees 2003, Posner 2004, Häggström 2016, Ord 2020, Russell 2019).

Second, there are things that society could do to significantly reduce these risks. For example, we could scale up efforts to avoid great power conflict, development of technical safety tools for artificial intelligence, pandemic preparedness, and legal regulation of dangerous technologies. These opportunities correspond to extremely high-value options that are available to various agents in certain decision situations. For example:

- Philanthropists with no constraints on what causes they support could provide funding to scale up these efforts.
- Talented individuals could devote their careers to these efforts.
- Policymakers could direct more public funding towards these efforts, or implement the appropriate regulations.

---

[2] For reasons of space, our presentation of the argument is only rough. For more careful presentations, see Beckstead (2013) and (2019), and Greaves and MacAskill (2023).
Our aim is to consider what form of longtermism is most plausible *given that* this standard argument is correct. Thus, we will take the correctness of its premises for granted throughout the chapter. This is not to claim that we ourselves are certain of their correctness, or to deny that they face reasonable objections.

[3] Existential catastrophe is often defined as "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development." For alternative formulations, see Ord (2020: 37), Cotton-Barratt and Ord (2015: 1-3), Greaves (ms).
In line with the standard argument, we assume that human extinction would be very bad (at least in expectation). For a contrary view, see for instance Althaus and Gloor (2018), Arrhenius and Bykvist (1995: ch. 3), and Benatar (2006: 183-200).

Third, the stakes are enormous. The risks in question threaten not just the present generation, but the *whole of the future of humanity* - a future that, if things go well, could continue for millions or billions of years. So anything we can do to reduce x-risk by even a tiny amount will carry enormous expected value – which is the metric by which orthodox decision theory evaluates options under uncertainty.[4]

Fourth, these enormous stakes lie mainly in the far future (more than 100 or even 1000 years hence). A typical lower-end estimate for the expected number of future lives is 10^{16}. Of these, only around 10^{8} occur in the next 100 years. So the next century accounts for only around 0.000001% (or less) of the value that would be destroyed by an existential catastrophe.

Finally: Crunching the numbers, we find that the far-future expected benefits of mitigating technological x-risks are vastly greater than any plausible expected benefit one could deliver within (say) the next hundred years (what we will call the "neartermist benchmark").[5] Some other type of action might be even better than mitigating technological x-risk, but if so, it would have to share the feature that most of its (expected) benefits are in the far future. Thus, impact on the far future is the most important feature of the most important actions that agents today can take. The far future is, in this sense, the most important feature of "our actions today" - as per longtermism.[6]

Various premises and moves in this argument might be disputed, but that is not our concern here. The motivation for the present essay arises rather from the fact that the final step in the argument, even if correct, at least carries some risk of being importantly misleading.

The point is that what is most directly supported by the above line of argument is what we will call *minimal longtermism*:

> *Minimal longtermism:* It is extremely important, for reasons concerning the far future, that key decision-makers (namely: philanthropists, talented individuals choosing careers, and some policymakers) take actions to direct significantly more resources towards addressing the technological x-risks mentioned above.

Minimal longtermism is "minimal" in at least three respects.

---

[4] In line with the standard argument for longtermism, we assume the correctness of expected value maximisation for concreteness and simplicity. Thus, when we talk about the *value*, *impact*, *benefits*, etc of an option, we always mean *expected* value/impact/benefits, unless otherwise specified. Other approaches to ex ante evaluation might or might not yield similar conclusions. Investigating the extent to which the present discussion is affected by choice points within decision theory lies beyond the scope of this essay.

[5] This "neartermist benchmark" might correspond, for instance, to the effects of bednet distribution on saving the lives of already existing people. For further discussion, see Greaves and MacAskill (2023).

[6] All the longtermist theses we consider in this chapter ("minimal longtermism", "expansive longtermism", and "(strong) longtermism" simpliciter) should be understood as ex ante axiological theses, i.e., theses about the value of possible actions relative to the agent's evidence. Thus, the preceding argument (i) does not claim to establish conclusions about what agents *ought* to do, or about moral permissibility, or to deny that various non-consequentialist considerations may bear on these further questions; and (ii) does not depend on the claim that we *can in fact* make the difference between existential catastrophe occurring vs. not.

First, it may recommend only one narrow type of action (or "intervention"): interventions that aim quite directly at reducing technological x-risks (such as funding to scale up technical AI safety work). The examples to which the standard argument appeals are exclusively of this type. It is consistent with that argument, and with minimal longtermism, that the only way of significantly improving the far future (in expectation) is to mitigate technological x-risks.

Second, minimal longtermism may concern only a narrow class of choice situations. The standard argument considers only the decision situations of (i) philanthropists or policymakers deciding how to allocate resources between different areas (and specifically, between x-risk mitigation vs. other, potentially more "near-termist" objectives), (ii) individuals choosing a career path, and (iii) policymakers enacting regulation in a few very specific areas (e.g., AI and biotechnology). While these decision situations may be especially important, they do not, by any natural count, encompass a majority of the decisions that present-day agents face. It is consistent with the standard argument, and with minimal longtermism, that in most decisions the far future is not a relevant issue.

Third, minimal longtermism may call for only a relatively modest reallocation of resources: for redirecting, say, less than 2% of world GDP towards addressing technological x-risks. The standard argument estimates the cost-effectiveness of opportunities to further mitigate technological x-risks at or near the current margin, but says nothing about how fast this marginal cost-effectiveness would decline if far more resources were allocated in that direction. It is consistent with the standard argument, and with minimal longtermism, that if (say) 2% of world GDP were spent on carefully chosen initiatives to mitigate technological x-risks, it would no longer be the case that attainable far-future benefits exceed the near-future benefits that are attainable for the same cost.

We said that the standard argument *establishes* only minimal longtermism. But some authors and organisations either explicitly endorse, or appear from their actions to believe, what we will call *Expansive Longtermism*. Expansive longtermism goes beyond minimal longtermism on all three of the dimensions just highlighted:[7]

Expansive longtermism:

(1) There is a very wide variety of ways to greatly improve the expected value of the far future. The full range of interventions with this property is many and varied, united by little or nothing except that they all significantly improve the far future in expectation. Actions that mitigate technological x-risks are just particularly clear-cut examples.
(2) Relatedly, possible effects on the far future are the main determinant of expected value comparisons in *nearly all* decision situations faced by *nearly all* present-day human agents.
(3) All this is not only true at the current margin, but is likely to remain true for at least the next several decades (barring an existential catastrophe), even if massively more resources (say, more than 50% of world GDP) come to be directed in ways that are optimised for the far future. The truth of longtermism arises from quite fundamental features of our situation in the early 21st century, rather than merely from an easily

[7] One could, of course, side with the "expansive" over the "minimal" view on some but not all of these three dimensions.

remediable failure to adequately fund and implement a few key safety measures related to new technologies.

Our view is that, while expansive longtermism *may* be true, the arguments for expansive longtermism are significantly less robust and significantly more speculative than the arguments for minimal longtermism. This is the thought that motivates the present essay. The essay's intended contributions are twofold. First, we call attention to the important differences between minimal and expansive longtermism, and to the fact that the standard argument establishes only the former. Second, we explore potential arguments for versions of longtermism that are more expansive than the minimal version, in any or all of the three ways just outlined. However, for the most part, we do not defend firm conclusions regarding which of those arguments ultimately hold water (a matter on which the two authors of this essay often incline in somewhat different directions).

The gap between minimal and expansive longtermism is important for at least three reasons. First, the question of *how wide a range of interventions* can significantly improve the far future in expectation relates to the question of whether longtermism is a helpful, or instead a misleading, organising concept. If the only ways of significantly improving the far future in expectation are targeted actions to mitigate technological x-risks, then even if it is *true* that impact on the far future is the most important feature of our actions today, it would be equally true and more informative to say that impact on technological x-risks is the most important feature of our actions today.[8] Currently, however, at least some philanthropic organisations explicitly adopt the more expansive "longtermist" framing.[9] This risks over-rating a large collection of hopelessly intractable would-be "longtermist" projects, by lumping them together with the extremely valuable project of technological x-risk mitigation.

Second, the remaining two dimensions on which expansive longtermism differs from its minimal cousin speak to an important source of unease about the longtermist worldview: the thought that longtermism threatens to be radically revisionary of both public and private morality on a large scale, to an extent that makes it both epistemically *prima facie* implausible and practically unappealing.

The second dimension (how many decision situations longtermism is true of) affects whether longtermism is best seen as a general moral outlook, or only a truth of much more localised relevance. The former view would be highly revisionary of the way that most people think of their everyday choices. Most of us do not ordinarily think that the far future of humanity bears

[8] The idea that "existential risk mitigation" is a more useful focus or organizing concept than "longtermism" (even if longtermism is true) comes up frequently in non-academic discussions of longtermism. See for instance Nanda (2022); Alexander (2022); and Yglesias (2022). These discussions often implicitly assume that mitigating technological x-risks is the only or most important way of effectively improving the long-run future.

[9] At the time of writing, for example, the website of Longview Philanthropy states:

> "Everything we do is guided by our core values:
>
> 1. **Longtermism:** Our collective future could be extraordinarily good or inordinately bad. We believe this generation has the power to influence which path humanity takes, and it is essential that we act responsibly.
>
> …" (Longview Philanthropy 2021).

on our decisions about, for example, what to have for breakfast. Minimal longtermism does not call this assumption into doubt. But on the expansive view, the far future is the primary determinant of the relative values of actions in nearly all decision situations.

The third dimension (*how many resources* would have to be redirected to longtermist causes before longtermism ceases to be true) relates to a concern that longtermism inappropriately deprioritizes the interests of the present generation. If taking longtermist reasoning to its logical conclusion implies that it would be better if (say) 80% of world GDP were directed in ways that are near-optimised for the far future, at massive expense to the present generation, some will regard this as a *reductio ad absurdum* of that longtermist reasoning.

Finally, of course, whether the most plausible version of longtermism is minimal or expansive on the above dimensions is practically important for those who find the standard argument for longtermism compelling. The right strategy for longtermism as an intellectual and social movement might look very different depending on whether its goal is (1) a wholesale moral transformation of society, or merely (2) a modest redirection of government and philanthropic budgets towards technological x-risk mitigation.

In the next three sections, we explore the three dimensions on which expansive longtermism goes beyond minimal longtermism. With respect to each dimension, we consider what case can be made for the more expansive longtermist thesis, and offer some (tentative, highly uncertain) assessment of that case.

How many interventions?

The standard argument for longtermism appeals to back of the envelope calculations of the cost-effectiveness of efforts to mitigate a few key existential risks, primarily in the category of technological x-risks (see Greaves and Macaskill 2023).[10] The case for *these* interventions being extremely cost-effective - and for generating far-future expected benefits that are many times higher than the highest available expected benefits for the near future - seems strong. But if one asks what *else* we can do to predictably improve the far future, *beyond* mitigating these few key risks, it is certainly possible to be unimpressed by what is on offer.

The problem is that, in general, the project of trying to influence the course of the far future has a strong air of intractability. The further into the future we look, the harder it becomes to predict either what the world will look like in the absence of any intervention on our part, or the effects of any particular present action. Risks of human extinction and other "existential catastrophe" create an exception to these worries about intractability, since each such risk comes with a strong and clear "lock-in" mechanism. But most other ways in which we might

[10] The arguments also sometimes appeal to *natural* x-risks such as those from asteroids and supervolcanoes (see Beckstead 2013: 68-69; **Greaves and MacAskill 2023**: **[ADD PAGE NUMBERS WHEN VOLUME PAGINATION IS FINALIZED]**). The cost-effectiveness of mitigating natural x-risks is usually judged to be significantly lower than that of mitigating technological x-risks, as a result of the fact that the absolute amount of technological x-risk we currently face is significantly higher than the amount of natural x-risk we currently face (see Ord 2020: 167). But if the space of tractable "longtermist" interventions is restricted to mitigation of *natural and* technological x-risks, rather than only technological x-risks, the general thrust of our discussion is unaffected.

hope to improve the far future of humanity can be motivated only via significantly more speculative reasoning concerning very long-term causal chains.

That said, we do think that the menu of plausible ways to improve the far future is somewhat longer than *just* highly targeted efforts to mitigate technological x-risks. The remainder of this section explores some possibilities, and gives our own tentative assessments.

Indirect existential risk mitigation

Even if an impartial concern for the far future recommends no changes to the status quo other than an effective global response to technological x-risks, this effective response might itself comprise a wide variety of interventions. It might include not just "direct work" (like AI safety research or advocating for nuclear arms reduction) but also various activities that *indirectly* mitigate either particular technological x-risks or technological x-risk in general. Several such indirect strategies have been suggested.

Aschenbrenner (2020) argues that any intervention that speeds up economic growth thereby contributes to mitigating technological x-risk, on the grounds that a growth speedup would shorten the duration of a "time of perils" during which technological x-risk per unit time is high, and after which it falls to near zero.[11] Aschenbrenner's argument seems inconclusive, however: his analysis considers only one factor, and on balance it seems unclear to us whether faster economic growth would reduce or increase existential risk.[12]

There are other broad social objectives, though, where the directionality (if not the magnitude) of the indirect impact on x-risk seems more clear. For instance, improving the quality of education would produce more informed electorates that might support and demand better policies on issues like climate change, nuclear non-proliferation, and pandemic preparedness. It might also improve the talent pool of future scientists, engineers, bureaucrats, etc, some of whom will work to manage these risks. Likewise, improving collective foresight could help us see existential risks further in advance and in greater resolution – e.g., giving us better predictions of when AI will be achieved, what kind of actor and what approach to AI will get there first, and how quickly AI capabilities will "take off" thereafter. We could improve foresight on these and other questions by training more professional forecasters, creating better systems for directing the efforts of these forecasters, and putting the results to practical use. Finally, general improvements to human moral values and character, or to the decision-making processes of powerful institutions would presumably reduce technological x-risks by causing future individuals and institutions to act against them more vigorously and effectively.

[11] The concept of a time of perils characterised by temporarily elevated levels of existential risk is originally due to Carl Sagan (1994: 173). The claim that we are living in a time of perils is not uncontroversial – for discussion, see MacAskill (2020); Thorstad (2022); and **Häggström (2023)**.
[12] For example, if the degree of existential risk from AI depends on the values of globally dominant agents at the time when AGI is first developed (e.g., the governments and large corporations that might participate in its development), those values tend to improve over time, and speeding up economic growth would hasten the advent of AGI, then speeding up economic growth could increase existential risk.

There are other broad social objectives we might pursue that would indirectly reduce some particular category of technological x-risk. For instance, working to reduce the risk of great power war, while also valuable for many other reasons, would notably reduce existential risks from nuclear weapons. Agents pursuing this goal might aim to reduce regional conflicts among minor powers that might spiral into a great power war; or try to promote and protect democracy in the great powers, since democracies are less likely to go to war with one another (see Chan 1997; Mello 2017); or promote cultural exchange between great powers, to increase public aversion to war.

For another example, Millet and Snyder-Beattie (2017) consider the project of bringing all human and animal health systems up to the minimal standards required by the International Health Regulations, starting from the status quo in which more than half of the world's countries do not have health systems of this standard. One effect of this would be to mitigate existential risks from pandemics, since it would allow rapid detection of and response to a pandemic wherever in the world it originates or spreads. Millet and Snyder-Beattie argue, to our minds convincingly, that even if one counts only the effects of this project on mitigating extinction risks from pandemics, it would save an expected number of lives per dollar that is (on most of the estimates of extinction risk magnitude they survey) significantly better than neartermist benchmarks like the nearterm benefits of anti-malarial bednet distributions or direct cash transfers.

A common feature of these indirect x-risk mitigation strategies is that they have substantial near-term cobenefits. Partly for this reason, many of the objectives discussed above already receive a great deal of attention from governments, institutional philanthropists, or other agents. This suggests that it is hard for most agents to make additional progress toward these objectives at the current margin, *perhaps* so hard that – even accounting for their indirect effects on technological x-risks – additional efforts toward these objectives have less marginal benefit than near-term benchmarks. But not all the strategies mentioned above have this character – for instance, improving long-term forecasting and collective foresight is quite neglected, and many health systems in the developing world are dramatically under-resourced. Our view, on the whole, is that there is a highly plausible longtermist case for prioritizing more highly at least some of the objectives discussed above, based on their contribution to technological x-risk mitigation.

Patient philanthropy

Let us now consider some strategies for improving the far future that are not primarily about reducing existential risk.

One very natural way of improving the far future is to *put more resources into the hands of far-future agents*. This can take two forms. The first strategy, *patient philanthropy*, aims to grow resources over time under compound interest, subsequently putting the proceeds into the hands of future altruistic agents (e.g., philanthropic institutions) for philanthropic projects at that future time. The second strategy, *speeding up growth*, has the less discriminate aim of increasing the wealth of far-future agents generally, most of whom will presumably use

those resources for their own benefit (or to benefit family and others for whom they have partial concern). We will discuss these two strategies in turn.

It has been frequently pointed out that altruistic individuals and philanthropic institutions may be able to do more good, from a temporally impartial point of view, by investing and growing their resources over time, rather than spending any philanthropic resources as soon as they become available (Landesman 1995; Moller 2006; Christiano 2013; Cotton-Barratt 2020). Importantly, this can be the case *even if* the cost-effectiveness of marginal philanthropic projects declines over time as society becomes richer, as the funds in question might grow faster than this marginal cost-effectiveness declines.[13] As emphasised by Trammell (2021a), this should in fact be expected if the majority of other philanthropists operating in the same space have positive rates of time preference. For in that case, the saving rate within the cause area in question will by default be lower than the impartial optimum, so that additional saving at the margin (within the cause area in question) is an improvement from the temporally impartial point of view.

There are various reasonable objections to this idea. For instance, perhaps there will simply be much less acute need in the future than in the present. Long-term philanthropic funds could also be expropriated or otherwise destroyed before they can be put to use. And one must trust that the future agents who manage the disbursement of resources, perhaps centuries from now, will do so wisely and with appropriately altruistic motivations (for a discussion of related worries, see Aird 2020). While these risks are real, however, they are far from certain to occur; while they decrease the amount of long-term philanthropic saving that is optimal, they do not reduce it to zero. It seems very plausible to us that at least at the current margin (with relatively few philanthropic resources invested on very long time horizons), additional long-term philanthropic saving would indeed be an improvement, even when the funds will subsequently be used for (what are then) "neartermist" projects.[14] If so, this could well be another type of intervention whose far-future expected benefits significantly exceed neartermist benchmarks.[15]

Accelerating growth

The less discriminate way of transferring resources to the far future is to speed up economic growth.[16] Cowen (2007,2018) argues that doing this could bring it about that people at every future time are better off than they otherwise would have been, and that since the Earth

[13] There is historical precedent for this. See, e.g., **Greaves and MacAskill (2023: [ADD PAGE NUMBERS WHEN VOLUME PAGINATION IS FINALIZED; pages 15-16, especially footnote 16, in this version])**.

[14] It might turn out, in the end, that the *most* effective use of a patient philanthropist's funds is to mitigate future existential risks, rather than to provide immediate material benefits to those alive at the time of disbursement (Trammell 2021c, section 2.5). In this way, patient philanthropy is *in part* another indirect strategy for existential risk mitigation. But its plausible applications are not limited to existential risk mitigation.

[15] One example of this approach being put into practice is the Patient Philanthropy Fund (see Hoeijmakers 2021).

[16] This could be via any combination of 'level effects' (one-time growth events that increase the *baseline* for future growth, without increasing the future growth *rate*) and 'growth effects' (increases in the growth rate itself over an extended period).

might support human life and civilization for millions of years to come, this permanent improvement would carry enormous value. Actions that we might take to this end include advocating for pro-growth policies and increases in funding for scientific research and technological development.

For unoriginal reasons, we are unconvinced by this argument. We will grant the assumption that historically and currently, welfare has generally increased over time as a result of economic growth (with, of course, many local deviations from this general trend). However, it seems far less clear whether we should expect this trend to continue indefinitely into the future. Both common sense and happiness research suggest that beyond a certain point, further economic progress makes little difference to individual welfare (see, e.g., Kahneman and Deaton (2010); Myers (2000)). And while *total* welfare at a time can nonetheless increase as long as population increases, there are also limits to the number of people who can sustainably live on Earth at any one time. In that sense, it seems that speeding up economic growth today does not yield exponentially increasing gains over an indefinite future, but only gets us to the inevitable saturation point faster. Even when we aggregate across a long future, the gains to be made from speeding up progress towards a saturation point are limited, and might well be relatively modest, though of course one must ultimately crunch the numbers to be sure (see Beckstead 2013: 67-73;MacAskill 2022: 136-139).[17]

Space settlement

There may be fairly modest limits to the value of economic growth, and on what we can achieve by speeding up growth, *as long as human civilization remains Earth-bound*. But if humanity eventually begins to settle space, and in particular if it embarks on an indefinite program of interstellar expansion, that is a different matter. First, it makes the upper bounds on accessible resources and sustainable population size astronomically larger. But second, because cosmic expansion means that the number of galaxies we could in principle reach is decreasing every year, the value of those upper bounds depends on how soon we begin to settle space and how fast we expand once we do (see Bostrom 2003). Thus, speeding up space settlement can raise the level at which humanity eventually plateaus, rather than merely getting us to the plateau faster.

Perhaps the strongest longtermist argument for speeding up economic growth is that it speeds up space settlement. But considering the speed of space settlement also suggests other interventions - for example, funding the development of new propulsion technologies or test projects like long-term bases on the Moon and Mars, and lobbying governments to increase funding for their own space agencies.

Bostrom claims (though with only minimal argument) that it is more important to increase the *probability* that space settlement eventually happens than to increase its *speed*. If this is right, then it primarily supports the conclusion that longtermists should prioritize mitigating

[17] In diagrammatic terms, the point is that if total welfare on Earth per unit time will reach a plateau, then speeding up progress is best thought of in terms of shifting an S-curve slightly to the left, rather than shifting the curve of progress slightly upwards for all time. See, e.g., MacAskill (2022: 140). A contrary perspective is suggested by Trammell (2021b).

risks of premature human extinction and similarly irreversible catastrophes that would prevent humanity from ever settling space. But one could also work to avoid futures in which humanity survives but *chooses* not to settle space, e.g. by supporting positive cultural depictions of space settlement and building legal frameworks that make space settlement positive-sum and attractive to all relevant parties.[18]

However these considerations turn out, the objective of space settlement seems highly likely to expand the menu of interventions whose far-future expected benefits plausibly exceed any benefits attainable in the near future.[19]

Improving moral values and institutions

Finally (in our incomplete exploration), some longtermists have argued that we can greatly improve the far future in expectation by working to permanently improve either the moral values of human civilization (e.g., via "moral circle expansion"; see Anthis and Paez 2021) or the quality of our political institutions (see MacAskill 2022: 70-96). Perhaps, for instance, we will eventually reach an equilibrium where one set of moral values is permanently ascendant, but multiple such equilibria are currently possible, and by careful moral reasoning and persuasion, we can positively influence which equilibrium is realised. Or perhaps we are in a similar situation with respect to forms of government. It might be that a world of liberal democracies, a world of totalitarian surveillance states, and a world of extractive quasi-feudal oligarchies are all stable and all currently possible long-term outcomes. In this case, interventions like promoting democracy might have enormous expected long-term benefits, by nudging humanity towards a better long-term equilibrium.

A difficulty for these strategies is that these areas are very crowded: enormous numbers of motivated and talented people have been trying for thousands of years to influence human values and institutions, in hundreds of different, competing directions. There are also no clear mechanisms for long-term persistence of cultural or institutional improvements, as there is for an outcome like extinction. In addition, perhaps more than anything else we have considered so far, it seems nearly impossible to give any remotely objective estimate of the

[18] Even these interventions may count as forms of existential risk mitigation, in the broad sense of "existential risk" common in the literature. If humanity never settles space, even if it survives happily for hundreds of millions of years on Earth, it will have achieved only a tiny fraction of its potential – an existential catastrophe. (Among other things, if humanity remains Earth-bound, it will survive for only a small fraction of its potential civilizational lifespan, i.e., will suffer "premature" extinction.)

[19] Note that there are two distinct questions we can ask about proposed longtermist interventions: whether they beat neartermist benchmarks like the near-term benefits of bednet distributions, and whether they are (at least in some decision situations) *optimal* from a longtermist perspective, producing greater far-future expected value than any other available longtermist interventions. We focus primarily on the first question for two reasons. First, it seems plausible that any longtermist intervention that beats the neartermist benchmarks will be optimal by temporally impartial lights in at least some circumstances, where higher-priority longtermist interventions are not available or have already been carried out. (Among other considerations, the value of information favours testing out many promising strategies for improving the far future even if one, like x-risk mitigation, initially seems to have much greater expected benefits than the rest.) Second, the question of what longtermist interventions beat the neartermist benchmarks is especially relevant when our interest (as here) includes the robustness of the case for longtermism: if the list of such interventions is long and diverse, then even someone who is skeptical of interventions that are in fact more cost-effective might agree with longtermism on the basis of considering interventions that are (in fact, or by another's lights) "lower down the list".

expected value of pursuing these projects, so the case for prioritising them will depend very much on `squishy', subjective probabilities, with plenty of space for reasonable disagreement.[20] We find ourselves correspondingly very uncertain about the longtermist case for trying to permanently influence values or institutions.

Taking stock

The standard argument for longtermism aims to show most directly that *there exist some* interventions whose far-future expected benefits are many times larger than the highest available near-future expected benefits. Our focus in this section has been on *how large* is the class of interventions with that feature, beyond the handful of examples that appear in the existence proof.

In general, the thrust of our discussion is that (i) the majority of interventions for which there is a *fairly robust* such "longtermist case" are in the category of mitigating technological x-risks, though (ii) that category is itself quite broad. We have found a couple of scattered interventions that are exceptions to this general rule (most plausibly, patient philanthropy and space settlement). Beyond technological x-risk mitigation and those scattered additional possibilities, the argument that any other interventions generate greater far-future benefits than the neartermist benchmark seems to us to depend much more heavily on subjective probabilities. This is of course not to say that no additional interventions with that "longtermist" property exist, but it is to say that there is plenty of room for reasonable disagreement.

How many decision situations?

The decision situations that make longtermism most compelling are those in which some resource (e.g. money or work hours) can be allocated to work that directly mitigates technological x-risks. One is deciding, perhaps, whether to spend philanthropic funding on technological x-risk mitigation or instead on bednets, or whether to pursue a career in AI safety or instead in medicine.

But most token decision situations don't seem to have this feature. When a transport minister makes decisions about an urban transport system, the possibility of taking the funding away from urban transport and donating it to AI safety research is not in the relevant

[20] As MacAskill (2022) notes, it is also true that a case for *not* prioritising the projects in question would rest on equally "squishy" subjective probabilities - probabilities that give projects of this nature higher far-future expected value than plausible near-term benchmarks don't seem actively outlandish. But there may be reason to err on the side of conservatism about the long-term impact of our actions when probabilities are squishy. In particular, perhaps we should start from a relatively confident prior that any given choice has only an extremely small effect on the probabilities of different long-term outcomes for humanity. Then, when our subjective estimates of the probabilities in a *particular* choice situation are based on ambiguous, non-robust evidence and arguments (and therefore more error-prone than our prior), we should update only very little from that prior. For a formal illustration of this idea, see Russell (ms).

option set: if they tried, the action would be blocked by higher authorities in the government or by the courts. When one is deciding what to have for breakfast, the available options include things like “eat Cheerios” and “eat peanut butter toast”, not things like “negotiate a nuclear arms reduction treaty”: the latter aren’t alternatives to the former.

It is *possible* that many of these other decision situations, too, are such that far-future considerations are the most important determinant of which options are best. But the usual argument for longtermism, focussing as it does entirely on cases of unconstrained resource allocation, does not establish that.

Beyond mere possibility: Are there any positive reasons to think that longtermism is true more broadly, encompassing many or even most real-world decision situations?

Here is one reason: Once it is agreed that philanthropic funding of technological x-risk mitigation has very high marginal benefits, any situation in which an agent’s choice has financial implications for agents who are willing to fund such work inherits some of its longtermist significance. Suppose, for example, that someone on a monthly salary has committed to donate half the money left in her bank account at the end of each month to an organisation working to prevent engineered pandemics. Then every purchasing decision she makes (where the options differ in price) has a fairly direct, though typically small, impact on total funding for x-risk mitigation. Similarly, every choice she makes at work and every choice in her personal life that affects her capabilities at work has an impact on her job performance, and thereby her expected future earnings, and thereby expected future funding for x-risk mitigation. Likewise, a policymaker facing any decision that affects the overall health of the economy thereby affects the financial wherewithal of many other agents, some of whom contribute to x-risk mitigation efforts.

The scope of this argument is somewhat unclear. For instance, most agents do not in fact donate any of their disposable income to x-risk mitigation, and so choices that affect their personal finances do not (at least in this way) implicate the total stream of longtermist funding. And perhaps most choices have no (or virtually no) financial implications for anyone – for instance, deciding where to sit at the family dinner table, or which park to visit with one’s children at the weekend. Even if most choices have *some* indirect impact on x-risk mitigation funding, this impact might be so attenuated that it is trumped by other factors, the standard argument for longtermism notwithstanding.

The proponent of expansive longtermism could respond here with what we might call the “appeal to astronomical stakes” – the idea that the sheer scale of the long-term future is so astronomically great that *anything* that affects the probability of existential catastrophe by *just about any amount* is likely to outweigh any near-term considerations, even without crunching the numbers (and perhaps the same goes for other shifts in the probability of different long-term outcomes or trajectories, apart from existential catastrophe).[21] If this is so,

[21] In this spirit, Bostrom (2013: 19) argues for the conclusion that “the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole” *without even considering any particular numbers* for the actual cost-effectiveness of available risk mitigation options.

then even highly indirect and heavily attenuated effects on this pool of funding may be overwhelmingly important.

On the other hand, one could easily accept the standard case for longtermism while believing that such estimates of the impact of tiny changes in funding for existential risk mitigation are wildly inflated. If the marginal returns to existential risk mitigation are merely large and not astronomical (say, beating the relevant neartermist benchmarks by only 1-2 orders of magnitude, rather than 10-15), then these attenuated effects may be trumped by neartermist considerations.

If there are more channels through which our everyday choices can influence the far future, however, those effects may be less attenuated and more significant. A more general version of the preceding argument from indirect effects on funding is (what we will call) the *many levers argument*.

> **Many Levers argument:** There are many social, political, and economic variables whose present-day values make some not-completely-negligible difference to the expected value of the far future. For instance, there are reasons to think that a higher present rate of economic growth would be good in the long run and reasons to think it would be bad, and these reasons probably do not net out to exact neutrality. So in one direction or another, the present growth rate matters for the long-run future. The same might be true, for instance, of population size, average educational attainment, various cultural norms (e.g., norms of tolerance or civility), the relative power of particular states (that have a positive or negative influence on international affairs), and so on. Nearly every choice we make, however mundane, has *some* influence on *at least one* of these factors. For instance, if you eat a more expensive breakfast, you'll have less money to save, slightly reducing the expected economic growth rate; or if you eat a nutritious breakfast, you'll be more productive at work, slightly increasing the expected economic growth rate. And since the future is vast, even a small effect on a variable that has a small effect on the long-term trajectory of civilization can carry substantial expected value, positive or negative — enough to swamp the immediate stakes of, say, having a more or less delicious breakfast.[22]

As with the narrower argument concerning effects on existential risk mitigation funding, this argument relies to a significant extent on an appeal to astronomical stakes. Even if variables like GDP and population make some difference to the probabilities of different long-term futures, their effects may be indirect, and in many mundane decision situations, the strongest effect of the agent's choice on any of those variables may itself be indirect or simply very weak. But if the far-future stakes are astronomically large (say, equivalent to 10^{30} or 10^{52} lives), then even these weak effects may be overwhelmingly important in expectation.

Our view is that this reasoning is suggestive but inconclusive. Even granting the standard argument for longtermism, it is still very much up for debate whether, say, the expected

[22] This argument was suggested to us by Owen Cotton-Barratt (in conversation). Balfour (2021) makes an argument along similar lines for the claim that the long-term importance of x-risk has unpleasant totalizing implications, requiring that "every waking moment" be governed by the imperative to minimise existential risk.

marginal benefit of philanthropic spending on x-risk mitigation is only a couple of orders of magnitude greater than neartermist benchmarks like anti-malarial bednets, or 10+ orders of magnitude greater, or something in between. And it is similarly unclear how much the ratio of long-term to near-term stakes is reduced in decision situations where our actions only affect important far-future outcomes very indirectly, but may affect important near-future outcomes directly. It is therefore extremely unclear how to assign the numbers required to flesh out the preceding arguments. If more rigorous assessment of these matters is possible, it would be very valuable.

How many resources?

Longtermism is often understood as having radical practical implications in the sense that it recommends a *very large* reallocation of resources from "neartermist" interventions towards whatever would most effectively improve the far future in expectation. One starts to worry, perhaps, that there would be little left for dealing with the needs or preferences of the present; we today might all become slaves to the cause of optimising the far future.

It might be helpful to make the issue more precise as follows. Suppose longtermism is true at the current margin. And imagine that we put more and more resources into the hands of agents who know that longtermism is (presently) true, and will spend those resources optimally.[23] Given a plausible model of the relevant diminishing marginal returns, at what point in this process, if ever, does longtermism cease to be true?

In this connection, three questions are pressing. First, how quickly would returns to investment in "longtermist" interventions diminish? Second, insofar as the amount of resources that would in an optimal scenario be *redirected* so as to better benefit the far future is large, how radical would the redirection be – in particular, how great a cost would it impose on the present generation? Third, if the optimal reallocation of resources is both large and radical, what should we make of this axiological fact in moral and practical terms?

First, then: How quickly would returns to investment in "longtermist" interventions diminish? For tractability, let us restrict the question to investment in existential safety (this will supply a lower bound on the continued cost-effectiveness of "longtermist" interventions more generally). There has been some investigation of the functional relationship between investment in a given area and marginal output in that area. For instance, Nicholas Rescher (1978 and 1997) has proposed a "law of logarithmic returns" to describe the growth of scientific knowledge. According to Rescher, knowledge increases logarithmically with the total quantity of resources invested in the scientific enterprise. In a series of blog posts, Owen Cotton-Barratt (2014a; 2014b; and 2014c) has argued for more general versions of the same principle. Cotton-Barratt claims that (i) the *expected* returns to resources invested in solving a problem of unknown difficulty (e.g., figuring out how to prevent extinction-level pandemics or align powerful AI systems with human values) and (ii) the *actual* returns to

[23] Where are those resources coming from, i.e., at the expense of what other people or projects? We can imagine that, for instance, we simply give the longtermists newly printed bills or newly extracted precious metals, so that the resources in the hands of other people and projects shrink at approximately the same rate, retaining their relative proportions.

resources invested in a problem *area* containing many independent problems of widely varying difficulty and importance are both approximately logarithmic within some central range (i.e., for resource investments that are not very small or very large). Cotton-Barratt's arguments could be reasonably taken to apply to spending on the mitigation of particular existential risks, or existential risk in general.

The case for this model is, as it stands, far from conclusive. But in the absence of any better-defended model, let us nonetheless consider its implications. If investment in particular cause areas obeys a law of logarithmic returns, it turns out that the optimal level of spending on x-risk mitigation or other longtermist cause areas depends very heavily on whether the current expected benefits of interventions in those areas are astronomical or merely large, in comparison to neartermist benchmarks. As a stylized illustration, suppose we can spend money on just two things: present consumption and x-risk mitigation. Assume that spending in either area yields logarithmic returns. Further assume that we currently spend 0.001% of world GDP (~$1 billion/year) on x-risk mitigation, and that the present marginal value of additional spending on x-risk mitigation is 10 times greater than the marginal value of transferring consumption to people in extreme poverty (who live on ~$500/yr). From these assumptions, we can conclude that the optimal level of spending on x-risk mitigation is less than 1% of world GDP.[24] On the other hand, if the present marginal value of additional x-risk mitigation is 10,000 times greater than the marginal value of consumption for the world's poorest, the same reasoning tells us that it would be optimal to spend ~71% of world GDP on x-risk mitigation (leaving ~$3571.43 per person per year for present consumption).[25]

Our impression is that some longtermists believe that the marginal value of the best longtermist interventions exceeds the neartermist benchmark by 10:1 or less, while others believe that this ratio is 10,000:1 or more. And there is more than enough reasonable basis for this wide range of views. In particular, estimates of the expected future population that would be lost to existential catastrophe span tens of orders of magnitude.[26] Depending on how one distributes one's credence over these estimates, reasonable best guesses about the badness of existential catastrophe can easily span several orders of magnitude, at the

---

[24] Let v_c give the value of (optimally) allocating a given fraction of world GDP to present consumption, and v_m give the total value of (optimally) allocating a given fraction of world GDP to x-risk mitigation. On the assumption that each individual has the same logarithmic function from consumption to utility, the optimal allocation of resources to present consumption will always distribute consumption equally. Allocating 1/25 (0.04) of world GDP to (equalized) consumption would give everyone in the world ~$500/yr. So, from the assumptions in the main text, we know that $v_c'(0.04) \approx v_m'(0.00001)/10$. If $v_c(x)$ and $v_m(x)$ increase like $\ln(x)$, then their first derivatives decrease like $1/x$. This means that $v_c'(0.04) \approx v_m'(0.0001)$. And *this* means that the optimal allocation of world GDP between present consumption and x-risk mitigation, which must equalize the marginal values of investment in each area, must exhibit the same ~400:1 ratio of consumption to x-risk mitigation spending. Thus, optimal spending on x-risk mitigation will be slightly less than 0.25% of world GDP.

[25] Now, we know that $v_c'(0.04) \approx v_m'(0.00001)/10000$, which implies that $v_c'(0.04) \approx v_m'(0.1)$. Thus, the optimal allocation of world GDP between present consumption and x-risk mitigation will exhibit a ~2.5:1 ratio of x-risk mitigation spending to present consumption.

[26] For instance, Millett and Synder-Beattie (2017) estimate the badness of extinction at $1.6 * 10^{16}$ life-years, based on the assumption that humanity would otherwise maintain a population of 10 billion individuals for the next 1.6 million years. On the other hand, Bostrom (2013) suggests that a future spacefaring civilization could support at least 10^{54} life-years worth of subjective experience. The issue is discussed in depth by Newberry (2021).

very least. The upshot is that even the fairly strong assumption of logarithmic returns does not pin down an answer to the question of what portion of humanity's resources it is optimal to allocate toward longtermist causes.

Second: Assuming that a large reallocation of resources toward longtermist objectives is optimal, how radical would the optimal reallocation be compared to the status quo, and how great a cost would it impose on the present generation? To illustrate this question, consider the following two scenarios:

> ***Scenario 1*:** Most workers are employed in highly focussed efforts to mitigate existential threats that have few if any co-benefits for the present generation. The largest employers are the asteroid detection and deflection industry, mathematical computing labs carrying out analysis directed towards AGI safety, and counterterrorism operations working to contain advanced terrorist threats from biotechnology. A small proportion of the population works in food production, but only to the minimal extent required to ensure that the risk mitigation lab staff have enough rice and beans to survive. A small proportion works in education, but only to the extent needed to train up the next generation to continue the fight against existential threats. There is little funding for healthcare: generally it works out more cost-effective to let nature take its course, and to devote the resources that might fund healthcare instead to making the existential safety industry a little larger. Similarly for leisure, entertainment, hospitality, the arts, and games: these are neartermist luxuries that have no place in society focused on minimising existential risk.
>
> ***Scenario 2*:** Returns to *highly focussed* efforts to mitigate technological x-risks (e.g., research on AI control and alignment, or negotiating nuclear arms reduction treaties) diminish quickly, such that the optimal level of investment in these efforts is only a small fraction of world GDP. Beyond a relatively modest investment in those focussed efforts, while it remains highly cost-effective to invest in technological x-risk mitigation, the most effective ways of mitigating those x-risks are things that are also very good for the near term. It is important for x-risk mitigation, for example, that there be effective and rational governance and decision-making at the national and supranational levels, a well-functioning economic system to implement society's preferences, and a populace that is free, prosperous, well-educated, and happy, which makes it a fertile generator of new ideas that will improve the long-run future, prone to support large cooperative projects, and immune to dangerous and destabilising forms of radicalization.

From the partial point of view of the present generation, Scenario 1 is a bleak vision in which impartial concern for the far future has taken the joy out of life today. Scenario 2, on the other hand, is not so radically different from the status quo (or rather, differs largely in ways that are *beneficial* to present people). It involves some minor reordering of existing priorities and some change in the rationales for existing projects, but to a reasonable first approximation, Scenario 2 reasonably well resembles the world that someone focused exclusively on the near term might aspire to create.

Between these scenarios, we find the bleak Scenario 1 somewhat less plausible. It seems to us likely that a world like Scenario 1 could be stable only if human psychology were

dramatically different. With actual human psychology, even if longtermism once commanded universal assent, the demands that longtermist morality could realistically make before leading to burnout, backlash, and conflict would likely be severely limited.

Third: The possibility that the optimal shift of resources towards optimisation for the far future might be very radical (as in Scenario 1) naturally raises a "demandingness" concern about longtermism: does an outlook that accepts this axiological claim also hold that we are morally obliged to pursue the identified optimum? If so, is that unacceptably demanding?

Here as elsewhere, though, it is important to keep the distinction between axiological and deontic claims in sharp focus. If the implications of an axiological longtermist thesis *together with maximising consequentialism* strike one as overly demanding, then (absent some other reason for doubting the axiological longtermist claim) the natural response is to reject maximising consequentialism, not to revise one's axiology or one's empirical beliefs.

At the same time, however, it is not plausible that axiological matters are *altogether irrelevant* to the deontic and practical issues. A morally conscientious person who rejects the conclusion that we ought to donate 99% of our income to the world's poorest usually does not conclude that the axiological analysis of global poverty is morally or practically irrelevant.[27] Similarly, if as an axiological matter something like Scenario 1 would indeed be optimal, it is much more plausible that deontically *some significant movement towards* Scenario 1 is warranted than that none whatsoever is warranted.

Summary and conclusions

The standard argument for longtermism appeals to a narrow range of examples: interventions *to mitigate technological x-risks* that are available *at the current margin* in a *small (if important) class of decision situations*. Many who are initially moved by this argument, though, often subsequently endorse a more sweeping longtermist view: what we have called "expansive longtermism". There are three dimensions of this strengthening, corresponding to the three clauses just italicised.

First, the expansive longtermist holds that the class of interventions whose far-future benefits far exceed any attainable near-future benefits is large and diverse, containing many thing that are not matters of mitigating technological x-risks. On this view, the role of the appeals to technological x-risks in the standard arguments is that of example to prove an existence claim.

Second, the expansive longtermist holds that the class of decision situations in which the value of the best options is determined primarily by far-future considerations is very large, encompassing not only decision situations of unconstrained resource allocation (as in some cases of private philanthropy and top-level public budget-setting), but also just about every aspect of public, professional and private life.

[27] It is much more plausible, in light of the facts about global poverty, that one ought to donate some significant but modest portion of one's income - perhaps, one's "fair share" (see, e.g., Miller 2011) - than that because the *optimal* level donation would be too demanding, there is no requirement to support the global poor at all.

Third, the expansive longtermist holds that longtermism is not only true *now, at the current margin*, but (further) would remain true even after very significantly more resources were directed towards mitigating technological x-risks (or improving the course of the far future more generally).

In all cases, our main purpose in this essay has been to highlight that there is an important gap between longtermism as usually argued for on the one hand, and expansive longtermism on the other. But we have also considered the question of how plausible the more expansive longtermist view is, on each of the three dimensions of expansion. Regarding the question of "how many interventions", we highlighted (i) patient philanthropy and (ii) projects of space settlement as particularly promising candidates for "longtermist interventions" that are not obviously matters of mitigating technological x-risks. Regarding the question of "how many decision situations", we outlined arguments from (1) indirect effects on the funding of technological x-risk mitigation and (2) the more general idea of "many levers" that might be used to support this claim. We found both arguments suggestive, but as they stand inconclusive. Regarding the question of how large and radical a reallocation of resources longtermist reasoning would deem optimal, we highlighted the issues of (1) how fast returns to spending on improving the course of the far future would diminish and (2) how costly the recommended reallocation would be in terms of present-day welfare as particularly important open questions. In all these cases, though, our attempts at initial contributions are only that, and there is significant scope for careful research to improve the state of the debate.[28]

References

Aird, M. (2020), 'Crucial questions about optimal timing of work and donations', *Effective Altruism Forum*. Available at https://forum.effectivealtruism.org/posts/LD3mNJ367tSMna6WR/crucial-questions-about-optimal-timing-of-work-and-donations.

Alexander, S. (2022), '"Long-termism" vs. "Existential Risk"', *Effective Altruism Forum*. Available at https://forum.effectivealtruism.org/posts/KDjEogAqWNTdddF9g/long-termism-vs-existential-risk.

Althaus, D., and Gloor, G. (2019), 'Reducing Risks of Astronomical Suffering: A Neglected Priority', *Foundational Research Institute*. Available at https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected -priority/

Anthis, J., and Paez, E. (2021), 'Moral circle expansion: A promising strategy to impact the far future', in *Futures* 130.

[28] For feedback on drafts of this chapter, we are grateful to Bradford Saad, David Thorstad, Philip Trammell, and participants in work-in-progress seminars at the Global Priorities Institute and Institute for Future Studies. For research assistance, we are grateful to Toby Newberry.

Arrhenius, G., and Bykvist, K. (1995), 'Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use', in *Uppsala Prints and Preprints in Philosophy* 21.

Aschenbrenner, L. (2020), 'Existential risk and growth', *Global Priorities Institute – Working Paper.*

Balfour, D. (2021), 'Pascal's Mugger Strikes Again', in *Utilitas* 33/1: 118-124.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future* (Unpublished doctoral dissertation). Rutgers, the State University of New Brunswick.

Beckstead, N. (2019), 'A Brief Argument for the Overwhelming Importance of Shaping the Far Future', in H. Greaves and T. Pummer (eds.), *Effective Altruism: Philosophical Issues* (Oxford University Press), 80-98.

Benatar, D. (2006), *Better Never to Have Been: The Harm of Coming into Existence* (Oxford University Press).

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', in *Utilitas* 15/3: 308-314.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', in *Global Policy* 4/1: 15-31.

Chan, S. (1997), 'In Search of Democratic Peace: Problems and Promise', in *Mershon International Studies Review* 41/1: 59-91.

Christiano, P. (2013), 'Giving now vs. later', *Rational Altruist.* Available at https://rationalaltruist.com/2013/03/12/giving-now-vs-later/.

Cotton-Barratt, O. (2014a), 'How to treat problems of unknown difficulty', *Future of Humanity Institute.* Available at http://www.fhi.ox.ac.uk/how-to-treat-problems-of-unknown-difficulty/.

Cotton-Barratt, O. (2014b), 'Theory behind logarithmic returns', *Future of Humanity Institute.* Available at http://www.fhi.ox.ac.uk/theory-of-log-returns/.

Cotton-Barratt, O. (2014c), 'The law of logarithmic returns', *Future of Humanity Institute.* Available at https://www.fhi.ox.ac.uk/law-of-logarithmic-returns/.

Cotton-Barratt, O. (2020), '"Patient vs urgent longtermism" has little direct bearing on giving now vs later', *Effective Altruism Forum.* Available at https://forum.effectivealtruism.org/posts/Eh7c9NhGynF4EiX3u/patient-vs-urgent-longtermism-has-little-direct-bearing-on.

Cotton-Barratt, O., and Ord, T. (2015), 'Existential Risk and Existential Hope: Definitions', *Future of Humanity Institute – Technical Report.*

Cowan, T. (2007), 'Caring About the Distant Future: Why It Matters and What It Means', *The University of Chicago Law Review* 74/5: 5-40.

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

Greaves, H. (MS) 'Concepts of Existential Catastrophe'

Greaves, H., and MacAskill, W. (2023), 'The Case for Strong Longtermism' in *THIS VOLUME*

Häggström, O. (2016), *Here Be Dragons: Science, Technology, and the Future of Humanity* (Oxford University Press).

Häggström, O. (2023), THIS VOLUME

Hoeijmakers, S. (2021), 'Introducing the Patient Philanthropy Fund', *Founders Pledge.* Available at https://founderspledge.com/stories/introducing-the-patient-philanthropy-fund.

Kahneman, D., and Deaton A. (2010), 'High income improves evaluation of life but not emotional well-being', in *Proceedings of the National Academy of Sciences* 107/38: 16489-16493.

Karnofsky, H. (2016a), 'Sequence thinking vs. cluster thinking', *The GiveWell Blog.* Available at https://blog.givewell.org/2014/06/10/sequence-thinking-vs-cluster-thinking/.

Karnofsky, H. (2016b), 'Why we can't take expected value estimates literally (even when they're unbiased)', *The GiveWell Blog.* Available at https://blog.givewell.org/2011/08/18/why-we-cant-take-expected-value-estimates-literally-even-when-theyre-unbiased/.

Landesman, C. (1995), 'When to Terminate a Charitable Trust?', in *Analysis* 55/1: 12-13.

Longview Philanthropy (2021), 'About Longview Philanthropy'. Available at https://www.longview.org/about.

MacAskill, W. (2020), 'Are we living at the hinge of history?', *Global Priorities Institute – Working Paper.*

MacAskill, W. (2022), *What We Owe the Future* (Basic Books).

Miller, D. (2011), 'Taking up the slack? Responsibility and justice in situations of partial compliance', in C. Knight and Z. Stemplowska (eds.), *Responsibility and Distributive Justice* (Oxford University Press).

Millett, P., and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', in *Health Security* 15/4: 373-383.

Mello, P. (2017), 'Democratic Peace Theory', in P. Joseph (ed.), *The SAGE Encyclopedia of War. 4, Social Science Perspectives* (SAGE Knowledge), 472-475.

Moller, D. (2006), 'Should We Let People Starve—For Now?', in *Analysis* 66: 240-247.

Myers, D. (2000), 'The Funds, Friends, and Faith of Happy People', in *American Psychologist* 55/1: 56-67.

Nanda, N. (2022), 'Simplify EA Pitches to "Holy Shit, X-Risk"', *Effective Altruism Forum.* Available at https://forum.effectivealtruism.org/posts/rFpfW2ndHSX7ERWLH/simplify-ea-pitches-to-holy-shit-x-risk.

Newberry, T. (2021), 'How many lives does the future hold?', *Global Priorities Institute – Technical Report.*

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

Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).

Rees, M. (2003), *Our Final Century: Will the Human Race Survive the Twenty-first Century* (Basic Books).

Rescher, N. (1978), *Scientific Progress: a philosophical essay on the economics of research in natural science* (Basil Blackwell).

Rescher, N. (1997), 'The Law of Logarithmic Returns and Its Implications' in D. Ginev and R. Cohen (eds.), *Issues and Images in the Philosophy of Science* (Springer Dordrecht).

Russell, J. (ms), 'Planning for Pascal's Mugging'. Available at https://philarchive.org/rec/RUSPFP-3

Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking).

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

Thorstad, D. (2022), 'Existential risk pessimism and the time of perils', *Global Priorities Institute – Working Paper.*

Trammell, P. (2021a), 'Dynamic public good provision under time preference heterogeneity: theory and applications to philanthropy', *Global Priorities Institute – Working Paper.*

Trammell, P. (2021b), 'New Products and Long-term Welfare', *philiptrammell.com*. Available at https://philiptrammell.com/static/New_Products_and_Long_term_Welfare.pdf.

Trammell, P. (2021c), 'Patient philanthropy in an impatient world'. Available at https://docs.google.com/document/d/1NcfTgZsqT9k30ngeQbappYyn-UO4vltjkm64n4or5r4/edit#

Vollrath, D. (2020), ‘Levels and growth rates’, *Growth Study Guide.* Available at https://growthecon.com/StudyGuide/preliminaries/levels.html#basic-concepts.

Yglesias, M. (2022), ‘What’s long-term about “longtermism”?’, *Slow Boring.* Available at https://www.slowboring.com/p/whats-long-term-about-longtermism#footnote-anchor-1.