

# The Case for Strong Longtermism

*Hilary Greaves and William MacAskill*

## 1 Introduction

A striking fact about the history of civilisation is just how early we are in it. There are 5,000 years of recorded history behind us, but how many years are still to come? If we merely last as long as the typical mammalian species, we still have over 200,000 years to go (Barnosky et al. 2011); there could be a further 1 billion years until the Earth is no longer habitable for humans (Wolf and Toon 2015); and trillions of years until the last conventional star formations (Adams and Laughlin 1999:34). Even on the most conservative of these timelines, we have progressed through a tiny fraction of history. If humanity's saga were a novel, we would be on the very first page.

Normally, we pay scant attention to this fact. Political discussions are normally centred around the here and now, focused on the latest scandal or the next election. When a pundit takes a 'long-term' view, they talk about the next five or ten years. With the exceptions of climate change and nuclear waste, we essentially never think about how our actions today might influence civilisation hundreds or thousands of years hence.

We believe that this neglect of the very long-run future is a serious mistake. An alternative perspective is given by *longtermism*, according to which we should be particularly concerned with ensuring that the far future goes well (MacAskill, 2022). In this essay we go further, arguing for *strong longtermism*: the view that impact on the far future is the *most* important feature of our actions today. We will defend both axiological and deontic versions of this thesis.

Humanity, today, is like an imprudent teenager. The most important feature of the most important decisions that a teenager makes, like what subject to study at university and how diligently to study, is not the enjoyment they will get in the short term, but how those decisions will affect the rest of their life.

The structure of the essay is as follows. Section 2 sets out more precisely the thesis we will primarily defend: axiological strong longtermism (ASL). This thesis states that, in the most important decision situations facing agents today, (i) every option that is near-best overall is near-best for the far future, and (ii) every option that is near-best overall delivers much larger benefits in the far future than in the near future.

We primarily focus on the decision situation of a society deciding how to spend its resources. We use the cost-effectiveness of anti-malarial bednet distribution as an approximate upper bound on attainable near-future benefits per unit of spending. Towards establishing a lower bound on the highest attainable far-future expected benefits, section 3 argues that there is, in expectation, a vast number of sentient beings in the future of human-originating civilisation. Section 4 then argues, by way of examples involving existential risk, that the project of trying to beneficially influence the course of the far future is sufficiently

tractable for ASL(i) and ASL(ii) to be true of the above decision situation. Section 5 argues that the same claims and arguments apply equally to an individual deciding how to spend resources, and an individual choosing a career. We claim these collectively constitute the most important decision situations facing agents today, so that ASL follows.

The remainder of the essay explores objections and extensions to our argument.

Section 6 argues that the case for ASL is robust to several plausible variations in axiology, concerning risk aversion, priority to the worse off, and population ethics. Section 7 addresses the concern that we are clueless about the very long-run effects of our actions. Section 8 addresses the concern that our argument turns problematically on tiny probabilities of enormous payoffs.

Section 9 turns to deontic strong longtermism. We outline an argument to the effect that, according to any plausible non-consequentialist moral theory, our discussion of ASL also suffices to establish an analogous deontic thesis. Section 10 summarises.

The argument in this essay has some precedent in the literature. Nick Bostrom (2003) has argued that total utilitarianism implies we should maximise the chance that humanity ultimately settles space. Nick Beckstead (2013) argues, from a somewhat broader set of assumptions, that ‘what matters most’ is that we do what’s best for humanity’s long-term trajectory. In this essay, we make the argument for strong longtermism more rigorous, and we show that it follows from a much broader set of empirical, moral, and decision-theoretic views. In addition, our argument in favour of *deontic* strong longtermism is novel.

We believe that strong longtermism is of the utmost importance: that if society came to adopt the views we defend in this essay, much of what we prioritise in the world today would change.

## 2 Precisifying strong longtermism

### 2.1 Axiological strong longtermism

Strong longtermism could be made precise in a variety of ways. First, since we do not assume consequentialism, we must distinguish between axiological and deontic claims. Let axiological (resp., deontic) strong longtermism be the thesis that far-future effects are the most important determinant of the value of our options (resp., of what we ought to do).

It remains imprecise what ‘most important determinant’ means. Taking the axiological case first, in this essay we consider the following more precise thesis:<sup>1</sup>

**Axiological strong longtermism (ASL):** In the most important decision situations facing agents today,

- (i) Every option that is near-best overall is near-best for the far future.
- (ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future.

<sup>1</sup> We discuss deontic strong longtermism in section 9.

Where condition (i) holds, one can identify the near-best options by focusing in the first instance only on far-future effects. If (as we believe, but will not argue here) the analogous statement regarding near-future effects is not also true, that supplies one sense in which far-future effects are ‘the most important’. Where condition (ii) holds, the evaluation of near-best options is primarily driven by far-future effects. That supplies another such sense.

In sections 3–5, we will argue that clauses (i) and (ii) of ASL hold of particular decision situations: those of a society deciding how to spend money with no restrictions as to ‘cause area’, an individual making the analogous decision, and individual career choice. Because these decision situations have particularly great significance for the well-being of both present and future sentient beings, we claim, they are the most important situations faced by agents today. Therefore, strong longtermism follows, even if ASL(i) and (ii) do not hold of any other decision situations.

Throughout, ‘the far future’ means everything from some time  $t$  onwards, where  $t$  is a surprisingly long time from the point of decision (say, 100 years). ‘The near future’ means the time from the point of decision until time  $t$ . We will interpret both ‘near-best overall’ and ‘near-best for the far future’ in terms of proportional distance from zero benefit to the maximum available benefit, and ‘much larger’ in terms of a multiplicative factor.

As we intend it, ASL is not directly concerned with the *objective* value of options and their *actual* effects. Rather, terms like ‘near-best’ and ‘benefits’ relate to the *ex ante* value of those options, given the information available at the time of decision, and their *prospects* for affecting the near or far future. *Ex ante* value may be expected value, but the statement of ASL does not presuppose this.

Since it refers to ‘benefits’, ASL makes sense only relative to a status quo option: benefits are increases in value relative to the status quo. As above, our primary examples will be cases of deciding how to spend some resource (either money or time). For concreteness, we will then take the status quo to be a situation in which the resources in question are simply wasted. However, other plausible choices of status quo would be unlikely to significantly affect our argument, and the argument does not require that the status quo be special in any deep sense.

ASL makes only comparative claims. We do not claim, and nor do we believe, that options cannot deliver large benefits without being near-best for the far future, or that available near-future benefits are small in any absolute sense. Our claim is rather that available benefits for the far future are many times larger even than this.

## 2.2 Benefit ratio and ASL

Our argument for ASL proceeds via the intermediate claim that the following property holds of the decision situations in question:

**Benefit ratio (BR):** The highest far-future *ex ante* benefits that are attainable without net near-future harm are many times greater than the highest attainable near-future *ex ante* benefits.

We prove in the Appendix that if BR holds of a given decision situation, then (first) so does ASL(ii), and (second) ASL(i) holds of a certain restriction of that decision situation. (The restriction involves removing any options that do net expected near-future harm; this restriction is innocuous in the context of our argument.)

Evaluating BR, and hence ASL, requires quantitative analysis; any particular quantitative analysis requires strong evaluative assumptions. To this end, we will *temporarily* make a particular, plausible but controversial, set of evaluative assumptions. Section 6, however, shows that various plausible ways of relaxing these assumptions leave the basic argument intact. One controversial assumption that may be essential, concerning the treatment of very small probabilities, is discussed in section 8.

The inessential assumptions in question include the following. First, we will identify the *ex ante* value of an option with its expected value: the probability-weighted average of the possible values it might result in. Second, we will identify value with total welfare: that is, we will assume a total utilitarian axiology. Third, and a near-corollary of the latter, we will assume time-separability. The latter allows us to separately define near-future and far-future benefits: overall value is then simply the sum of near-future value and far-future value, where these in turn depend only on near-future (resp., far-future) effects.

For a rough upper bound on near-future expected benefits in the context of a society spending money, we consider the distribution of long-lasting insecticide-treated bednets in malarial regions, which saves a life on average for around US\$4,000. Each \$100 therefore saves on average 0.025 lives in the near future (GiveWell 2020a).<sup>2</sup>

We cannot argue that this is the action with the very largest near-future benefits. In particular, though it seems hard to beat this cost-effectiveness level via any intervention that is backed by rigorous evidence, it might be possible to achieve higher short-term *expected* benefits via some substantially more speculative route.<sup>3</sup> A full examination of the case for strong longtermism would involve investigation of this, and the corresponding sensitivity analysis. However, even quite large upward adjustments to the figure we use here would leave our argument largely unaffected.

We emphasise that we are not considering the long-run knock-on effects of bednet distribution. It is possible, for all we say in this essay, that bednet distribution is the best way of making the far future go well, though we think this unlikely.<sup>4</sup>

We will argue in section 4 that, for a society's decision about how to spend its resources, the lower bound on attainable far-future expected benefits is many times higher than this upper bound for near-future expected benefits, and therefore BR holds of this decision situation. Section 5 discusses related decision situations facing individuals.

<sup>2</sup> Following GiveWell (2018), we will assume that the short-term benefits of the interventions that do the most short-run good would scale proportionately even if very large amounts of money were spent.

<sup>3</sup> A sister organisation to GiveWell, Open Philanthropy, has tried hard to find human-centric interventions that have more short-term impact, and has struggled (Berger 2019). There might be more cost-effective interventions focused on preventing the suffering of animals living in factory farms (Bollard 2016). We leave this aside in order to avoid getting into issues of interspecies comparisons; again, there is a corresponding need for sensitivity analysis.

<sup>4</sup> It would amount to a 'surprising and suspicious convergence' between near-future and far-future optimisation (Lewis 2016).

### 3 The size of the future

There is, in expectation, a vast number of lives in the future of human civilisation.<sup>5</sup> Any estimate of just how ‘vast’ is of course approximate. Nonetheless, we will argue, existing work supports estimates that are sufficiently large for our argument to go through.

There are several techniques one can use for estimating the expected number of future beings. Let us start with the question of the expected *duration* of humanity’s future existence, temporarily setting aside questions of how large the population might be at any future time.

First, one might use evidence regarding the age of our species to ground judgements on the annual risk of extinction from natural causes, and extrapolate from there. Given that *Homo sapiens* has existed for over 200,000 years, Snyder-Beattie, Ord, and Bonsall (2019:2) thereby estimate that the expected future lifespan of humanity is at least 87,000 years, as far as natural causes of extinction are concerned.

Second, one might undertake *reference class forecasting* (Kahneman and Lovallo 1993; Flyvbjerg 2008). Here, the lifespans of other sufficiently similar species serve as benchmarks. Estimates of the average lifespan of mammalian species (resp., hominins) are between 0.5 and 6 million years (resp., around 1 million years) (Snyder-Beattie et al. 2019:6). Thus reference class forecasting, naively applied, suggests at least 1 million years for the expected future duration of humanity.

Both of these estimates, however, ignore the fact that humans today are highly atypical. Humanity today is significantly better equipped to survive extinction-level threats either than other species are, or than our own species was in the past, thanks to a combination of technological capabilities and geographical diversity. Therefore a range of substantially higher benchmarks is also relevant: for instance, the frequency of mass extinction events (1 in every 30–100 million years (Snyder-Beattie et al. 2019:7)), and the time over which the Earth remains habitable for humans (around 1 billion years (Adams 2008:34)).

The above figures concern the expected *duration* of humanity’s future. Since we are interested in the expected *number* of future beings, we also need to consider population size.

We again consider several benchmarks. First, the UN Department of Economic and Social Affairs (2019:6) projects that the global population will plateau at around 11 billion people by the year 2100. Second, the large majority of estimates of the Earth’s ‘carrying capacity’—that is, its long-run sustainable population, based on relatively conservative assumptions about future technological progress—are over 5 billion, and sometimes substantially higher (Cohen 1995:342; Bergh and Rietveld 2004:197). Third, for predicting the further future, we might extrapolate from the historical trend of human population increasing over time. Given this trend, it is at least plausible that continued technological advances will enable an even larger future population up to some *much* higher plateau point (say, 1 trillion), even if we cannot currently foresee the concrete details of how that might happen (Simon 1998).

Importantly, it is the *expected* number of future beings, not the median, that is relevant for our purposes. In addition to the possibility of numbers like the higher benchmarks indicated above, it is of course also possible that the future duration and/or population size

<sup>5</sup> We will use ‘human’ to refer both to *Homo sapiens* and to whatever descendants with at least comparable moral status we may have, even if those descendants are a different species, and even if they are non-biological.

of humanity are much *smaller*.<sup>6</sup> However, the effects of these possibilities on the expected number are highly asymmetric. Even a 50% credence that the number of future beings will be *zero* would decrease the expected number by only a factor of two. In contrast, a credence as small as 1% that the future will contain, for example, 1 trillion beings per century for 100 million years (rather than 10 billion per century for 1 million years) increases the expected number by a factor of 100.

We must also consider two more radical possibilities that, while very uncertain, could greatly increase the duration and future population sizes of humanity. The first concerns space settlement. There are currently no known obstacles to the viability of space settlement, and some scientific investigations suggesting its feasibility using known science (Armstrong and Sandberg 2013; Beckstead 2014). If humanity lives not only on Earth but also on other planets—in our own Solar System, elsewhere in the Milky Way, or in other galaxies too—then terrestrial constraints on future population size disappear, and astronomically larger populations become possible. Even if we only settle the Solar System, civilisation would have over 5 billion years until the end of the main sequence lifetime of the Sun (Sackmann, Boothroyd, and Kraemer 1993:462; Schröder and Smith 2008:162), and we would have access to over 2 billion times as much sunlight power as we do now (Stix 2002:6; Sarbu and Sebarchievici 2017:16). If we are able to widely settle the rest of the Milky Way, then we could access well over 250 million rocky habitable-zone planets (Bryson et al. 2021:22), each of which has the potential to support trillions of lives over the course of their sun's lifetime. Moreover, an interstellar civilisation could survive until the end of the stelliferous era, on the order of 10 trillion years hence (Adams and Laughlin 1999). If we consider possible settlement of the billions of other galaxies accessible to us, the numbers get dramatically larger again.

The second radical possibility is that of digital sentience: that is, conscious artificial intelligence (AI). The leading theories of philosophy of mind support the idea that consciousness is not essentially biological, and could be instantiated digitally (Lewis 1980; Chalmers 1996: ch. 9). And the dramatic progress in computing and AI over just the past 70 years should give us reason to think that if so, digital sentience could well be a reality in the future. It is also plausible that such beings would have at least comparable moral status to humans (Liao 2020), so that they count for the purposes of the arguments in this essay.<sup>7</sup>

Consideration of digital sentience should increase our estimates of the expected number of future beings considerably, in two ways. First, it makes interstellar travel much easier: it is easier to sustain digital than biological beings during very long-distance space travel (Sandberg 2014:453). Second, digital sentience could dramatically increase the number of beings who could live around one star: digital agents could live in a much wider variety of environments (Sandberg 2014:453), and could more efficiently turn energy into conscious life (Bostrom 2003:309).

One might feel sceptical about these scenarios. But given that there are no known scientific obstacles to them, it would be overconfident to be certain, or near-certain, that space

<sup>6</sup> On duration: technological progress brings not only protection against existing extinction risks, but also novel sources of extinction risk (Ord 2020: esp. chs. 4 and 5). On population size: the tendency for richer societies to have lower fertility rates has led some to conjecture that human population, after plateauing around 2100, might significantly decline into the indefinite future, a high 'carrying capacity' notwithstanding (Bricker and Ibbitson 2019).

<sup>7</sup> We return to the likelihood of artificial superintelligence in subsection 4.3.

settlement, or digital sentience, will not occur. Imagine that you could peer into the future, and thereby discovered that Earth-originating civilisation has spread across many solar systems. How surprised would you be, compared to how surprised you would be if you won the lottery?

To move towards particular numbers, we consider three specific future scenarios, taken from Newberry (2021a), where civilisation is: (i) Earthbound; (ii) limited to the Solar System; and (iii) expanded across the Milky Way. In each case, Newberry makes a conservative estimate of the carrying capacity of civilisation in that scenario, on the assumptions that digital life is and is not possible, giving six scenarios in all. He also provides a best-guess estimate of the duration of civilisation in that scenario. These scenarios are not meant to exhaust the space of possibilities, but they give an indication of the potential magnitudes of future population size (see Table 2.1).

To arrive at an *overall* estimate of the expected number of future people, one would further need to estimate probabilities for scenarios such as those in Table 2.1 (and for all other scenarios). However, since the number of lives in the future according to different possible scenarios is spread over many orders of magnitude, in any such expected-value calculation, it tends to be the ‘largest’ scenario in which one has any non-zero credence that drives the overall estimate. Even a 0.01% credence that biological humanity settles the Milky Way at carrying capacity, for example, contributes at least  $10^{32}$  to the expected number of future beings. Precisely how one’s remaining credence is spread among ‘smaller’ scenarios then makes very little difference.

Because of this, we believe that any reasonable estimate of the *expected* number of future beings is at least  $10^{24}$ . (In fact, we believe that any reasonable estimate must be substantially higher than this; since higher numbers would make little difference to the arguments of this essay, however, we will not press that case here.) However, we are also sympathetic to the concern that if this is the only estimate we consider, the case for strong longtermism would be driven purely by tiny credences in highly speculative scenarios. We will therefore also consider the extent to which the same arguments would go through on some vastly more conservative estimates, as shown in Table 2.2.

Our low estimate ( $10^{18}$ ) corresponds, for instance, to a 0.0000001% credence in the Solar System (biological life) scenario, with zero credence in either digital sentience or more wide-ranging space settlement. Our restricted estimate ( $10^{14}$ ) corresponds to the above

**Table 2.1** Estimates of the number of future lives in various scenarios.

Scenario	Duration (centuries)	Carrying capacity (lives per century)	Number of future lives
Earth (mammalian reference class)	$10^4$	$10^{10}$	$10^{14}$
Earth (digital life)	$10^4$	$10^{14}$	$10^{18}$
Solar System	$10^8$	$10^{19}$	$10^{27}$
Solar System (digital life)	$10^7$	$10^{23}$	$10^{30}$
Milky Way	$10^{11}$	$10^{25}$	$10^{36}$
Milky Way (digital life)	$10^{11}$	$10^{34}$	$10^{45}$



**Table 2.2** Estimates of the expected number of future beings.

	Expected number of future beings
Main estimate	$10^{24}$
Low estimate	$10^{18}$
Restricted estimate	$10^{14}$

estimate for Earthbound life, with zero credence in any larger-population scenario (including both digital sentience and any space settlement). In the arguments that follow, the reader is invited to substitute her own preferred estimate throughout.

We will argue that BR (and hence ASL) holds of society's decision situation even on our restricted estimate, and clearly holds by a large margin on our main estimate.

## 4 Tractability of significantly affecting the far future

The far-future effects of one's actions are usually harder to predict than their near-future effects. Might it be that the expected instantaneous value differences between available actions decay with time from the point of action, and decay sufficiently fast that in fact the near-future effects tend to be the most important contributor to expected value? If that were so, then neither BR nor ASL would hold.

This is a natural reason to doubt strong longtermism. We will call it *the washing-out hypothesis*.<sup>8</sup>

We agree that the washing-out hypothesis is true of some decision situations. However, we claim that it is false of our society's decision situation.

Given the argument of section 2, our task is to show that there exists at least one option available to society with the property that its far-future expected benefits are significantly greater than the near-future expected benefits of bednet distribution (that is, recall: 0.025 lives saved per \$100 spent). We will consider examples in two categories: mitigating extinction risk, and positively shaping the development of artificial superintelligence.

### 4.1 Influencing the choice among persistent states

Here is an abstract structure which, *insofar as* it is instantiated in the real world, offers a recipe for identifying options whose effects will not wash out.

Consider the space  $S$  of all possible fine-grained states the world could be in at a single moment of time (that is, the space of all possible instantaneous microstates). One can picture the history of the universe as a path through this space. Let a *persistent state* be a subset  $A$  of  $S$  with the property that, given the dynamics of the universe, if the instantaneous state

<sup>8</sup> It is important here to distinguish between *ex ante* and *ex post* versions of the washing-out claim. The *ex post* version is false, as is established by the literature on cluelessness; cf. subsection 7.1. However, it is the much more plausible *ex ante* washing-out claim that is relevant to the arguments of this essay.



of the world is in *A*, then the expected time for which it remains in *A* is extremely long. Now suppose that there are two or more such persistent states, differing significantly from one another in value. Suppose further that the world is not yet in any of the states in question, but might settle into one or the other of the states in question in the foreseeable future. Finally, suppose that there is something we can do now that changes the probability that the world ends up in a better rather than a worse persistent state. Then, as a result of the persistence that is built into the definition, the effects of these actions would not wash out at all quickly.

The empirical question is whether there are, in the real world, any options available that instantiate the structure just described. We claim that there are.

## 4.2 Mitigating risks of premature human extinction

The non-existence of humanity is a persistent state *par excellence*. To state the obvious: the chances of humanity re-evolving, if we go extinct, are miniscule. Only slightly more subtly, the existence of humanity is also a persistent state: while we face significant risks of premature extinction, as argued in section 3, humanity's *expected* persistence is vast.

These persistent states have unequal expected value. Assuming that on average people have lives of significantly positive welfare,<sup>9</sup> according to total utilitarianism the existence of humanity is significantly better than its non-existence, at any given time. Combining this with the fact that both states are persistent, premature human extinction would be astronomically bad. Correspondingly, even an extremely small reduction in extinction risk would have very high expected value (Bostrom 2013:18). For example, even if there are 'only'  $10^{14}$  lives to come (as on our restricted estimate), a reduction in near-term risk of extinction by 1 millionth of 1 percentage point would be equivalent in value to a million lives saved; on our main estimate of  $10^{24}$  expected future lives, this becomes 10 quadrillion ( $10^{16}$ ) lives saved.

As is increasingly recognised, as an empirical matter of fact, there are things we could do that would reduce the chance of premature human extinction by a non-negligible amount. As a result, although precise estimates of the relevant numbers are difficult, the far-future benefits of some such interventions seem to compare very favourably, by total utilitarian lights, to the highest available near-future benefits.

The detection and potential deflection of asteroids provides a relatively robust example of such an intervention. This involves scanning the skies to identify asteroids that could potentially collide with Earth and, if one were found, investing the resources to try to deflect it, and/or to prepare bunkers and food stockpiles to help us survive an impact winter. Most of the expected costs here are in detection, because the costs of deflection and preparation are only paid in the very unlikely event that one does detect an asteroid on a collision course with Earth.

In 1996, NASA commenced the Spaceguard Survey, a multi-decade plan to track near-Earth objects with the aim of identifying any on impact trajectories. At a total cost of \$71 million by 2012, the Spaceguard Survey had tracked over 90% of asteroids of diameter 1km or more in near-Earth orbit, and all asteroids of diameter 10km or more over 99% of the sky.

<sup>9</sup> We return to this assumption in section 6.

It is not certain that a large asteroid collision would cause human beings to go extinct. We assume a status quo risk of human extinction, conditional on the impact of an asteroid of diameter 10km or more, of 1%. It is also far from certain that we could deflect such a large asteroid, even if we knew it was on a collision course. However, it is far from certain that we could not, and, as above, there are other actions we could take to protect against the extinction risk. We assume here that if such an object were detected to be on a collision course, our deflection and preparation efforts would reduce extinction risk by a proportional 5%. The assumptions in this paragraph follow Newberry (2021b), and seem fairly conservative.

Putting these numbers together, we estimate that the Spaceguard Survey, on average, reduced extinction risk by at least  $5 \times 10^{-16}$  per \$100 spent. On our main estimate of the expected number of future beings, this amounts to an additional 500 million lives; this decreases to 500 or 0.05 lives on our low and restricted estimates, respectively.

Of course, we should expect *further* work on asteroids to have lower cost-effectiveness, because of diminishing marginal returns. However, the opportunity remains significant. The remaining risk of a collision with an asteroid with a diameter of 10km or more in the next 100 years has been estimated at 1 in 150 million (Ord 2020:71). It has been estimated that the cost to detect with near-certainty any remaining asteroids of greater than 10km diameter would be at most a further \$1.2 billion (Newberry 2021b). On our main (resp., low, restricted) estimate of the expected number of future beings, every \$100 of this would, on average, result in 300,000 (resp., 0.3, 0.00003) additional lives. This example therefore supports strong longtermism on our main and low estimates, though not on the restricted estimate. Organisations whose work mitigates risk of extinction from asteroid impacts, and which would benefit from substantially more funding, include the Planetary Society and the B612 Foundation.

While asteroid defence is among the more easily quantified areas of extinction risk reduction, it is far from the only one, or the most significant (Ord 2020: ch. 3). Another possibility concerns global pandemics. Such a pandemic could be natural or man-made, with the latter being particularly concerning (Posner 2004:75–84; Rees 2018: sec. 2.1; Ord 2020). In particular, progress in synthetic biology is very rapid (Meng and Ellis 2020), and it is likely that we will soon be able to design man-made viruses with very high contagiousness and lethality. If such pathogens were released (whether deliberately or by accident (Shulman 2020; Ord 2020:129–131)) in the course of military tensions, or by a terrorist group, there is a real possibility that they could kill a sufficient number of people that the human species would not recover.

In a recent paper, Millet and Snyder-Beattie (2017) use three distinct methods to generate estimates of the risk of an extinction-level pandemic in the next 100 years. The resulting estimates range from 1 in 600,000 to 1 in 50. The authors further use figures from the World Bank to generate a very conservative estimate that \$250 billion of spending on strengthening healthcare systems would reduce the chance of such extinction-level pandemics this coming century by at least a proportional 1%.<sup>10</sup>

Taking the geometric mean to average across the two methods that generate the lower estimates for extinction risk, we obtain a risk of about 1 in 22,000 of extinction from a

<sup>10</sup> Two ways in which Millet and Snyder-Beattie's estimate is particularly conservative are (i) that the \$250 billion figure is at the extreme upper end of anticipated costs for the intervention they discuss, and (ii) that the intervention in question concerns an extremely broad-based approach to biosecurity, not specifically optimising for extinction risk reduction.

pandemic over the next 100 years.<sup>11</sup> If we use the above figure of \$250 billion to reduce the risk by 1%, and assume that the risk reduction occurs throughout the next 100 years but only in that time period, then each \$100 of such spending would, in expectation, increase the number of future beings by 200 million (resp., 200, 0.02) on our main (resp., low, restricted) estimate. According to these calculations, the far-future benefits would thereby significantly exceed the near-future benefits of bednet distribution on our main and low estimates of the size of the future, though not on our restricted estimate. Organisations working on these threats include the John Hopkins Center for Health Security, the Nuclear Threat Initiative's biosecurity programme, and Gryphon Scientific.

### 4.3 Influencing the choice among non-extinction persistent states

A second way of positively impacting the long run is by improving the value of the future conditional on the existence of a very large number of future sentient beings. For concreteness, we focus on one way of doing this: positively shaping the development of artificial superintelligence (ASI), that is, artificial systems that greatly exceed the cognitive performance of humans in virtually all domains of interest.<sup>12</sup>

The idea that the development of sufficiently advanced artificial intelligence could prove a key turning point in history goes back to the early computer pioneers Alan Turing (1951) and I. J. Good (1966). It has more recently been the subject of wider concern.<sup>13</sup> There are two classes of long-term worry.

The first is from *AI-takeover* scenarios (Bostrom 2014; Russell 2019). This worry is that, once we build a human-level artificial intelligence, it would be able to recursively self-improve, designing ever-better versions of itself, quickly becoming superintelligent. From there, in order to better achieve its aims, it will try to gain resources, and try to prevent threats to its survival. It would therefore be incentivised to take over the world and eliminate or permanently suppress human beings. Because the ASI's capability is so much greater than that of humans, it would probably succeed in these aims.

The second worry is from *entrenchment* scenarios (MacAskill 2022). If an authoritarian country were the first to develop ASI, with a sufficient lead, they could use this technological advantage to achieve world domination. The authoritarian leader could then quash any ideological competition. An AI police force could guarantee that potential rebellions are prevented; an AI army would remove any possibility of a coup. And if the leader wanted his ideology to persist indefinitely, he could pass control of society on to an ASI successor before his death. To this end, he could hard-code the goals of the ASI to match his own, have the ASI learn his goals

<sup>11</sup> We use the geometric rather than the arithmetic mean because the estimates in question are spread across several orders of magnitude; the arithmetic mean effectively defers to the highest estimate on the question of order of magnitude. Using the arithmetic mean would lead to results that are still more favourable to strong longtermism. Similarly, we disregard Millett and Snyder-Beattie's 'Model 1' because, as the authors note, this model is flawed in important respects; including this model would also strengthen the case for strong longtermism.

<sup>12</sup> Other areas one might consider here include affecting the values that the world converges on (Reese 2018), or reducing the risk of a totalitarian world government (Caplan 2008).

<sup>13</sup> Those concerned include leading machine learning researchers such as Stuart Russell (2019) and Shane Legg (2008: sec. 7.3), philosophers such as Nick Bostrom (2014), Eliezer Yudkowsky (2013), Toby Ord (2020:138–152), and Richard Ngo (2020), physicists such as Max Tegmark (2017: ch. 4) and Stephen Hawking (2018: ch. 9), and tech entrepreneurs such as Elon Musk (2014), Sam Altman (2015), and Bill Gates (Statt 2015).

from his speech and behaviour, or even ‘mind upload’, scanning his brain and having it digitally emulated (Sandberg and Bostrom 2008b; Sandberg 2013).

In either of these scenarios, once power over civilisation is in the hands of an ASI, this could persist as long as civilisation does (Finnveden, Riedel, and Shulman 2022). Different versions of the ASI-controlled futures are therefore persistent states with significantly differing expected value, so that we have another instantiation of the structure outlined in subsection 4.1. In the scenario just mentioned, the ASI’s goals were “locked in”. The ASI would be immortal and all-powerful, so those goals would determine both the trajectory and value of the future. These aims, inherited from its creator, could be pursued across a vast cosmic civilization for billions of years.

Though extinction risks involve dramatic reductions in the size of the future population, these AI scenarios need not. In the classic statement of the AI-takeover scenario, the ASI goes on to settle the stars in pursuit of its goals (Bostrom 2014:100). Similarly, if an authoritarian leader transferred power to an ASI, they too might want their civilisation to be large, populous, and long-lasting. In particular, for a wide variety of goals (such as building the grandest possible temples, doing research, or, in a toy example Bostrom (2014:123–124) gives to illustrate the general phenomenon of misaligned AI, maximising the number of paperclips), acquiring more resources helps with achievement of these goals, which motivates settling the stars. And, in order to fulfil these goals, a populous workforce would be instrumentally valuable. In expectation, the number of future beings, in these scenarios, is very large.

Now, this workforce might consist almost entirely of AIs. But, as we noted in section 3, there are reasons to think that such beings would have moral status, and therefore how well or poorly their lives went would be of moral concern, relevant to the arguments of this essay. And, at least on the authoritarian-takeover scenarios, the ruler might wish to have a very large number of human followers, too.

There are two strands of work aimed at reducing risks from ASI. First, AI safety research, which aims to ensure that AI systems do what we intend them to do (Amodei et al. 2016). Such work is conducted by organisations such as Berkeley’s Center for Human-Compatible AI, the Machine Intelligence Research Institute, and labs within Google DeepMind and OpenAI. Second, policy work, in particular to ensure a cooperative approach between countries and companies: for example, by The Partnership on AI, the Centre for the Governance of AI, and the Center for New American Security.

Despite this work, ASI safety and policy are still extremely neglected. For example, the Open Philanthropy Project is the only major foundation with these issues as a key focus area; it spends under \$30 million per year on them (Open Philanthropy 2020).<sup>14</sup> The AI safety teams at OpenAI and DeepMind are small.

There is no hard quantitative evidence to guide cost-effectiveness estimates for AI safety work. Expert judgement, however, tends to put the probability of existential catastrophe from ASI at 1–10%.<sup>15</sup> Given these survey results and the arguments we have canvassed, we

<sup>14</sup> Neglectedness is crucial to the argument of this essay. Would strong longtermism still be true if, for example, 10% of global GDP were already spent on the most valuable long-term-oriented interventions? Even if true, would it still be significantly revisionary compared to a near-termist approach, as we have claimed it is at the current margin? We aren’t sure. Our claim here is only that the world today is clearly far below this optimum.

<sup>15</sup> Grace et al. (2018) asked 352 leading AI researchers to give a probability on the size of existential risk arising from the development of ‘human-level machine intelligence’; the median estimate was 5%. A survey among participants at a conference on global catastrophic risks similarly found the median estimate to be 5% (Sandberg and Bostrom 2008a). One would expect a selection effect to be at work in surveys of those who have chosen to work on existential risk, but not so (or not strongly) for the survey of AI researchers.

think that even a highly conservative assessment would assign at least a 0.1% chance to an AI-driven catastrophe (as bad as or worse than human extinction) over the coming century. We also estimate that \$1 billion of carefully targeted spending would suffice to avoid catastrophic outcomes in (at the very least) 1% of the scenarios where they would otherwise occur. On these estimates, \$1 billion of spending would provide at least a 0.001% absolute reduction in existential risk. That would mean that every \$100 spent had, on average, an impact as valuable as saving 1 trillion (resp., 1 million, 100) lives on our main (resp., low, restricted) estimate—far more than the near-future benefits of bednet distribution. Readers are invited to substitute their own preferred estimate here; even quite large downward adjustments of the figure that we ourselves guess will leave the argument substantially unaffected.

#### 4.4 Uncertainty and ‘meta’ options

There is a lot of uncertainty in the numbers we have given, even in the most scientifically robust case of asteroid detection. We will give this issue a more thorough treatment in the next section, arguing against various ways in which one might worry it undermines our argument.

One thing that uncertainty can support, however, is a preference for different types of strategy to improve the far future. Rather than directly trying to influence the far future, one could instead try to invest in decision-relevant research, or invest one’s resources for use at a later date.

The possibility of either of these strategies strengthens our argument considerably. To see this, let us suppose, for the sake of argument, that no ‘first-order’ intervention (such as those we discussed in subsections 4.2–4.3) delivers higher far-future expected benefits than the highest available near-future expected benefits, relative to the credences that are appropriate in the present state of information. Suppose, however, that it is highly likely that, conditional on sufficient additional information, at least one of the proposed interventions, or another such intervention (not yet considered) in a similar spirit, would have much higher far-future benefits, relative to the updated credences, than the best available near-future benefits. Then society might *fund research into* the cost-effectiveness of various possible attempts to influence the far future. Provided that subsequent governments or philanthropists would take due note of the results, this ‘meta-option’ could easily have much greater far-future expected benefits than the best available near-future expected benefits, since it could dramatically increase the expected effectiveness of future governmental and philanthropic action (all relative to *currently* appropriate credences).

A complementary possibility is that, rather than spending now, society could save its money for a later time (Christiano 2014; MacAskill 2019; Trammell 2020). That is, it could set up a sovereign wealth fund, with a longtermist mission. This fund would pay out whenever there becomes available some action that will sufficiently benefit the far future (in expectation), whether that is during the lifetimes of current citizens or later. There would be some annual risk of future governments being misaligned and using the money poorly, but this risk could be mitigated via constitutional enshrinement of the mission, and

would be compensated by the fact that the fund would benefit from compound returns of investment.<sup>16</sup>

These considerations show that the bar that ‘intractability’ objections to our argument must meet is very high. For BR to fail to hold on such grounds, *every* option available to society must have negligible effect on the far future’s expected value. Moreover, it must be near-certain that there will be no such actions in the future, and that no such actions could be discovered through further research. This constellation of conditions seems unlikely.

## 5 Strong longtermism about individual decisions

So far we have discussed what is best for a society to do, sometimes referring to what billions of dollars would be able to achieve. But what about individuals?

We believe our arguments apply to individuals in much the same way they apply to society as a whole. Suppose Shivani is an individual philanthropist, deciding where to spend her money. Naively, we might think of Shivani as making a contribution to asteroid detection, pandemic preparedness, or AI safety that is proportional to her resources. If \$1 billion can decrease the chance of an asteroid collision this century by 1 in 120,000, then \$10,000 can decrease the chance of an asteroid collision by 1 in 12 billion. Because the individual’s ability to contribute to short-term good would also decrease proportionally, perhaps the argument goes through in just the same way.

This ‘naive’ argument is, in our view, approximately correct. We foresee three ways of resisting it.

First, one could claim that private individuals are much more limited in their options, to such an extent that Shivani can do nothing to decrease risks from asteroids, pandemics, or AI. However, this is simply not true. Multiple organisations working on these risks, including most of those we mentioned above, accept funding at all scales from private individuals, and would scale up their activity in response.

Second, one could claim that there are increasing returns to scale, so that the impact of a small donation is much less than the relevant fraction of the impact of a large donation. This is an open possibility, but it seems significantly more likely that there are fairly

<sup>16</sup> Plausibly, the gains from the investment would outweigh the risk of value-drift of the fund: the historical real rate of return on risky investments (such as stocks and housing) was around 7% during the period 1870–2015 (Jordà et al. 2019:1228). It seems reasonable to expect substantially lower returns in the future; but even if so, they would still be significantly higher than the risk of future governments misusing the funds; even a 90% probability of a future government misusing the funds over the next century would amount to only 2% annual risk.

There is some precedent for successful long-lasting trusts in the charitable sector. In the US, the John Clarke Trust was founded in 1676 (Ochs, 2019); in the UK, King’s School, Canterbury was established in 597 (House of Commons Public Administration Select Committee, 2013). In 1790, Benjamin Franklin invested £1,000 for each of the cities of Boston and Philadelphia: three-quarters of the funds would be paid out after 100 years, and the remainder after 200 years. By 1990, the donation had grown to almost \$5 million for Boston and \$2.3 million for Philadelphia (Isaacson 2003:473–474). The oldest similar government funds date back to the mid-19th century: Texas’s Permanent School Fund was founded in 1845 (Texas Education Agency 2020), and its Permanent University Fund was founded in 1876 (University of Texas System 2021). If the annual chance of failure of such funds were as high as 2%, then the chance of the Texas Permanent School Fund persisting until the present day would be 1 in 30, and the chance of the King’s School persisting until the present day would be 1 in 10 trillion. This does not merely appear to be a selection effect: to our knowledge, it is not the case that there have been very large numbers of attempted long-lasting government funds that have failed. This suggests that 2% is a conservatively high estimate of the annual risk of failure.



strongly *diminishing* returns, here as elsewhere.<sup>17</sup> This is for both theoretical and empirical reasons. Theoretically: since interventions vary in their *ex ante* cost-effectiveness, a rational altruistic actor will fund the most cost-effective intervention first, before moving to the next-most cost-effective intervention, and so on. Empirically, diminishing returns have been observed across many fields (e.g. Cassman, Dobermann, and Walters 2002:134; Arnold et al. 2018; Bloom et al. 2020).

Third, one could claim that, once we consider the actions of individuals with smaller amounts of resources, the probability of success from directing those resources to long-term-oriented interventions becomes so low that expected utility theory gives the wrong recommendations. We discuss this issue in section 8.

What of individual decisions about where to direct one's *labour*, rather than one's money? We believe that much the same arguments apply here. Suppose that Adam is a young graduate choosing his career path. Adam can choose to train either as a development economist, or as an AI safety researcher. While there are differences between Adam's decision situation and Shivani's (MacAskill 2014), there are also important similarities. In particular, the considerations that make it better in expectation for Shivani to fund AI safety rather than developing world poverty reduction similarly seem to make it better in expectation for Adam to train as an AI safety researcher rather than as a development economist.

## 6 Robustness of the argument

In our initial presentation of the argument, we have at times assumed expected total utilitarianism, for simplicity. This raises an important question of how wide a class of axiologies will support axiological strong longtermism.

First, what if instead of maximising expected total welfare, the correct axiology is risk averse?<sup>18</sup> This in fact seems to strengthen the case for strong longtermism: the far-future interventions we have discussed are matters of mitigating catastrophic risks, and in general terms, risk aversion strengthens the case for risk mitigation. With only minor modifications, similar remarks apply if, instead of replacing risk neutrality with risk aversion, we replace appeals to utilitarianism in our argument with (*ex post*) prioritarianism.

Second, if the only means of positively influencing the far future were via reducing the risk of extinction, the case for strong longtermism might rely on controversial views in population ethics, such as totalism, on which the absence of a large number of happy future beings makes things much worse. But many axiologies will not agree that premature extinction is extremely bad. In particular, person-affecting approaches to population ethics tend to resist that claim. According to the spirit of a person-affecting approach, premature

<sup>17</sup> Relatedly, it seems that insofar as scale does make a difference, ASL(i) and (ii) are *more* likely to be true of decision situations involving smaller sums of money, not less likely.

Increasing-returns phenomena are discussed by Pierson (2000).

<sup>18</sup> On the standard account, to be risk averse is to have utility be a concave function of total welfare (Pratt 1964:127; O'Donoghue and Somerville 2018:93). Some have argued that the standard account is inadequate (Rabin 2000; Buchak 2013:30). On risk-weighted expected utility theory, risk aversion is represented by a risk function that transforms the expected utility function (Quiggin 1982; Quiggin and Wakker 1994; Buchak 2013). The differences between these accounts are unimportant for our purposes.



extinction is in itself at worst neutral: if humanity goes prematurely extinct, then there does not exist any person who is worse off as a result of that extinction, and, according to a person-affecting principle, it follows that the resulting state of affairs is not worse. The far-future benefits of extinction risk mitigation may therefore beat the best near-future benefits only conditional on controversial population axiologies.<sup>19</sup>

However, risks from ASI are unlike extinction in this respect: there will be a large population in the future either way, and we are simply affecting how good or bad those future lives are. The idea that it's good to improve expected future well-being *conditional on the existence of a large and roughly fixed-size future population* is robust to plausible variations in population-ethical assumptions.<sup>20</sup>

Third, the example of ASI risk also ensures that our argument goes through even if, in expectation, the continuation of civilisation into the future would be bad (Arrhenius and Bykvist 1995: ch. 3; Benatar 2006; Althaus and Gloor 2018). If this were true, then reducing the risk of human extinction would no longer be a good thing, in expectation. But in the AI lock-in scenarios we have considered, there will be a long-lasting civilisation either way. By working on AI safety and policy, we aim to make the trajectory of that civilisation better, whether or not it starts out already 'better than nothing'.

One feature of expected utilitarianism that is near-essential to our argument is a zero rate of pure time preference. With even a modest positive rate of pure time preference (as e.g. on 'discounted utilitarian' axiologies), the argument would not go through. Our assumption of a zero rate, however, matches a consensus that is almost universal among moral philosophers, and also reasonably widespread among economists.<sup>21</sup>

This is of course nowhere near an exhaustive list of possible deviations from expected total utilitarianism. We consider some other deviations below, in the course of discussing cluelessness and fanaticism. Our conclusion is that the case for strong longtermism is at least *fairly* robust to variations in plausible axiological assumptions; we leave the investigation of other possible variations for future research.

<sup>19</sup> It is not immediately clear precisely what a person-affecting approach will say about the value of extinction risk mitigation, since the usual formulations of those theories do not specify how the theories deal with risk, and it is not immediately clear how to extend them to cases that do involve risk. Thomas (2019) explores a number of possibilities.

<sup>20</sup> 'Narrow' person-affecting approaches disagree, since they regard two states of affairs as incomparable whenever those states of affairs have non-identical populations (Heyd 1988). However, such approaches are implausible, for precisely this reason. Similarly, theories on which any two states of affairs with *non-equinumerous* populations are incomparable (Bader n.d.) are implausible. When comparing different sized populations, a 'wide' person-affecting approach will typically map the smaller population to a subset of the larger population, and compare well-being person by person according to that mapping (Meacham 2012); these theories will tend to agree with total utilitarianism on the evaluation of the AI catastrophes we discuss.

For similar reasons, we also do not consider here the incomparability that is introduced by a 'critical range' view (Blackorby, Bossert, and Donaldson 1996).

<sup>21</sup> A zero rate of pure time preference is endorsed by, inter alia, Sidgwick (1890), Ramsey (1928), Pigou (1932), Harrod (1948), Solow (1974), Cline (1992), Cowen (1992), Stern (2007), Broome (2008), Dasgupta (2008), Dietz, Hepburn, and Stern (2008), Buchholz and Schumacher (2010), and Gollier (2013). In a recent survey of academic economists with expertise on the topic of social discounting, 38% of respondents agreed with this 'Ramsey-Stern view' (Drupp et al. 2018:119). Greaves (2017) provides a survey of the arguments on both sides.

Even among philosophers, the consensus against discounting future well-being is not universal. In particular, some plausible models of partiality suggest assigning greater effective moral weight to one's own contemporaries than to far-future people (Setiya 2014; Mogensen 2019). However, even these models seem unlikely to recommend *sufficient* discounting to undermine the argument for longtermism (Mogensen 2019: sec. 6).

## 7 Cluelessness

Section 4 focused on worries about our *abilities to affect* the far future. A distinct family of worries is more directly epistemic, and involves the idea that we are *clueless* both about what the far future will be like, and about the differences that we might be able to make to that future.<sup>22</sup> Perhaps the beings that are around will be very unlike humans. Perhaps their societies, if they have anything that can be called a society at all, will be organised in enormously different ways. For these and other reasons, perhaps the kinds of things that are conducive to the well-being of far-future creatures are very different from the kinds of things that are conducive to our well-being. Given all of this, can we really have any clue about the far-future value of our actions *even in expectation*?

We take it for granted that we cannot *know* what the far future will be like. But, since the argument of sections 2–6 has already been conducted in terms of *expected* value, lack of knowledge cannot ground any objection to the argument. The objection must instead be something else.

In fact, there are several quite distinct possibilities in the vicinity of the ‘cluelessness’ worry. In the present section, we address five of these objections, relating to simple cluelessness, conscious unawareness, imprecision, arbitrariness, and ambiguity aversion.

### 7.1 Simple cluelessness

Our concern is with relatively weighty decisions, such as how to direct significant philanthropic funding. But it is illuminating to compare these to far more trivial decision situations, such as a choice of whether or where to go shopping on a given day.

Even in the latter cases, many have argued, we can be all but certain that our choice will have highly significant consequences *ex post*—far more significant than the more predictable nearer-term effects.<sup>23</sup> The reasons for this include the tendency for even trivial actions to affect the identities of future persons far into the future. However, when comparing quite trivial alternatives, we can have no idea *which* of the two will turn out to be superior vis-à-vis these deeply unpredictable very far-future effects.

Some have argued that these facts undermine any attempt to base decisions on considerations of the overall good *even in trivial everyday decision contexts* (e.g. Lenman 2000). We agree with Greaves (2016) that this concern is overblown: in the context of relatively trivial everyday decisions, at least, the deeply unpredictable far-future effects plausibly cancel out for the purpose of comparing actions in expected-value terms. Consequently, there is no objection here to basing *these* decisions on an expected-value assessment of nearer-future, more foreseeable effects.

As we have argued in section 4, however, decisions about how to spend philanthropic funding are disanalogous in this respect. We are not discussing the possibility that either funding AI safety research or not funding it might lead, as chance has it, to the birth of

<sup>22</sup> Since ‘washing-out’ concerns whether we are able to affect the far future *in expectation*, this too has an epistemic aspect, so that the distinction between the concerns of section 4 and those discussed here is not completely clear (Tarsney 2019). Nonetheless, the issues raised seem sufficiently different to warrant a separate treatment.

<sup>23</sup> See e.g. Lenman (2000), Greaves (2016). We agree with this claim, but our argument does not rely on it.

an additional unusually good or bad person several centuries hence. Rather, we are discussing the possibility that funding AI safety might have its intended effect of making AI safer. While there are certainly severe uncertainties in such work, it would be overly pessimistic to insist that success is no more likely than counterproductivity. Considerations of such ‘simple’ cluelessness therefore do nothing to undermine the argument for strong longtermism.

## 7.2 Conscious unawareness

The expected-value approach we assumed in section 3 is intended as a *subjective* decision theory: that is, it utilises only material that is accessible to the decision-maker at the time of decision. In particular, therefore, there is an implicit assumption that the agent herself is in a position to grasp the states, acts, and consequences that are involved in modelling her decision.

But perhaps this is not true. Consider, for example, would-be longtermists in the Middle Ages. It is plausible that the considerations most relevant to their decision—such as the benefits of science, and therefore the enormous value of efforts to help make the scientific and industrial revolutions happen sooner—would not have been on their radar. Rather, they might instead have backed attempts to spread Christianity, perhaps by violence: a putative route to value that, by our more enlightened lights today, looks wildly off the mark.

The suggestion, then, is that our current predicament is relevantly similar to that of our medieval would-be longtermists. Perhaps there are actions available to us that would, *if* we were able to think it all through in full detail, then deliver high expected benefits for the far future. But we know, if only by induction from history, that we have not thought things through in all relevant detail. Perhaps we thereby have good reason to reject subjective expected-value analysis, and use some quite different form of decision analysis to assess far-future effects—in which case, all bets are as yet off regarding what the conclusion will be.

This is the issue of *conscious unawareness*—knowing that one is unaware of many relevant considerations, mere awareness of which would influence one’s decision-making. Following much of the recent literature on this topic, however, our view is that conscious unawareness does not occasion any particularly significant revision of the Bayesian framework, for three reasons.

First, we know that we operate with coarse-grained models, and that the reasons for this include unawareness of some fine-grainings. Of course, failure to consider key fine-grainings might lead to different expected values and hence to different decisions, but this seems precisely analogous to the fact that failure to possess more information about which state in fact obtains similarly affects expected values (and hence decisions). Since our question is which actions are *ex ante* rational, both kinds of failure are beside the point.

Second, we know we are likely to be omitting some important possible states of nature from our model altogether. But consciousness of this can be modelled by inclusion of a ‘catchall’ state: ‘all the other possibilities I haven’t thought of’. Again, conceptualising parts of this state in more explicit terms might change some expected-value assessments, but

again this does nothing to undermine the *ex ante* rationality of decisions taken on the basis of one's existing assessments.<sup>24</sup>

Third, while the best options might well be ones that have not occurred to us, that does nothing to impugn the rationality of assessments of those possible options that *have* occurred to us. And our argument for strong longtermism, recall, requires only a lower bound on attainable far-future expected benefits.

We do not claim (nor do we believe) that issues of conscious unawareness have no effect on what the reasonable credences and values in a given decision situation are. The point is rather that these issues need not occasion any deep structural change to the analysis. Our further claim is that the numbers we have suggested in section 4 are reasonable *after* taking issues of conscious unawareness into account.

### 7.3 Arbitrariness

An obvious and potentially troubling feature of our discussion in section 4 is the paucity of objective guidance for the key values and probabilities. This seems to contrast starkly with, for instance, the usual impact evaluations for the short-term benefits of bednet distribution, which can be guided by relatively hard evidence (GiveWell 2020b).

This gives rise to three distinct, though related, concerns with the standard Bayesian approach that we have used. The first is simply that the probabilities and/or values in this case are too arbitrary for our argument to carry any weight. The second is that in cases where any precise assignments would be this arbitrary, it is inappropriate to have precise credences and values at all. The third is that in such cases, the appropriate decision theory is ambiguity averse, and that this might undermine the argument for strong longtermism. We address these concerns in turn.

The 'arbitrariness' objection is that even if a rational agent must have some precise credence and value functions, there is so little by way of rational restriction on *which* precise functions are permissible that the argument for strong longtermism is little more than an assertion that the authors' own subjective probabilities are ones relative to which this thesis is true.

We have some sympathy with this objection. However, there is a distinction between there being no watertight argument against some credence function on the one hand, and that credence function being *reasonable* on the other. Even in the present state of information, in our view credence-value pairs such that the argument for strong longtermism fails are unreasonable. If, for instance, one had credences such that the expected number of future people was only  $10^{14}$ , the status quo probability of catastrophe from AI was only 0.001%, *and* the proportion by which \$1 billion of careful spending would reduce this risk was also only 0.001%, then one would judge spending on AI safety equivalent to saving only 0.001 lives per \$100—less than the near-future benefits of bednets. But this constellation of conditions seems unreasonable.

However, we note that this issue is contentious. We regard the quantitative assessment of the crucial far-future-related variables as a particularly important topic for further research.

<sup>24</sup> The first type of unawareness is unawareness of possible *refinements*, the second is unawareness of possible *expansions* (Bradley 2017: sec. 12.3; Stefánsson and Steele 2021: sec. 3.2).

## 7.4 Imprecision

*Imprecise* approaches represent an agent by a *class* of pairs of probability and value functions—a *representor*—rather than a single such pair. The natural interpretation is that these correspond to incomplete orderings of options: one option is better than another, for instance, if and only if the first has higher expected value than the second on *all* probability-value pairs in the representor.<sup>25</sup>

ASL involves comparing *ex ante* far-future benefits with *ex ante* near-future or total benefits. If imprecision is a feature of rational evaluation at all, it is plausibly a particularly prominent feature of evaluation of far-future consequences. So perhaps, for any option (including the ones we have discussed above), any reasonable representor contains at least some elements according to which the far-future benefits of this option are no higher than the near-future benefits of bednet mitigation.

It is somewhat complex to say how one should evaluate ASL in the context of such imprecision. (For instance: Should we simply evaluate ASL itself relative to each element of the representor in turn, and superevaluate to arrive at an overall verdict? Or should we seek to define subsentential terms like ‘near-best’ in the context of representors? If the latter, how exactly?) The general idea, though, is that one way or another, if the possibility in the last sentence of the preceding paragraph is realised, then ASL is at least not determinately true.

Our reply to the imprecision critique is very similar to our reply to the arbitrariness critique. While we do not take a stand on whether or not any imprecision of valuation is either rationally permissible or rationally required (Elga 2010), we don’t ourselves think that any plausible degree of imprecision in the case at hand will undermine the argument for strong longtermism. For example, we don’t think any reasonable representor even *contains* a probability function according to which efforts to mitigate AI risk save only 0.001 lives per \$100 in expectation. This does seem less clear, however, than the claim that this is not a reasonable precise credence function.

## 7.5 Ambiguity aversion

In employing the standard Bayesian machinery, we have been assuming *ambiguity neutrality*. In contrast, an *ambiguity-averse* decision theory favours gambles that involve more rather than less objectively specified probabilities, other things being equal (Machina and Siniscalchi 2014).

Empirically, people commonly demonstrate ambiguity aversion. Suppose, for example, that one urn contains 50 red balls and 50 black balls, and a second urn contains both red and black balls in unknown proportion (Ellsberg 1961). If one is ambiguity averse, one might strictly prefer to bet on the risky urn, where one knows the probability of winning, regardless of which colour one is betting on. This preference seems inconsistent with expected utility theory, but is widespread (Trautmann and Kuilen 2015).

It might seem at first sight that ambiguity aversion would undermine the case for strong longtermism. In contemplating options like those discussed in section 4, one needs to settle

<sup>25</sup> Bewley (2002), Dubra, Maccheroni, and Ok (2004), and Galaabaatar and Karni (2013) provide representation theorems linking such representations to incomplete orderings.

one's credence that some given intervention to reduce extinction risk, or to increase the safety of ASI, would lead to a large positive payoff in the far future. But again, there seems significant arbitrariness here. In contrast, impact evaluations for the near-future benefits of bednet distribution seem to involve much more precisely bounded probabilities. Might an ambiguity-averse decision theory, then, take a substantially dimmer view of the far-future benefits of existential risk mitigation, and hence of strong longtermism?

Our answer is 'no', for two reasons.

First, whether or not ambiguity aversion has any prospect of undermining the argument for strong longtermism depends, in the first instance, on whether the agent in question is ambiguity averse with respect to the state of the world, or instead with respect to the difference one makes oneself to that state. The above argument-sketch implicitly assumed the latter.<sup>26</sup> But, if one is going to be ambiguity averse at all, it seems more appropriate for an altruist to be ambiguity averse in the former sense (Greaves et al, 2024). And it is far from clear that actions seeking to improve the far future increase ambiguity with respect to the state of the world. It is already extremely ambiguous, for instance, how much near-term extinction risk humanity faces.<sup>27</sup> We see no reason to think that this latter ambiguity is increased, rather than decreasing or remaining the same, by, for example, funding pandemic preparedness.<sup>28</sup>

Second, although it is psychologically natural, and correspondingly widespread, ambiguity aversion is anyway irrational. Here we agree with a fairly widespread consensus; we have nothing to add to the existing debate on this question.<sup>29</sup>

We conclude that the possibility of ambiguity aversion does not undermine the argument for strong longtermism.

## 8 Fanaticism

One obvious point of contrast between the paradigm examples of ways to attain high near-future vs. far-future expected benefits is that the former tend to involve high probabilities of relatively modest benefits, whereas the latter tend to involve tiny probabilities of enormous benefits. In discussing actions aimed at mitigating extinction risk, for instance, we conceded that it is *very unlikely* that any such action makes any significant difference; the argument for prioritising such actions nonetheless is characteristically that *if they do* make a significant difference, they might make a truly enormous one.

Even among those who are sympathetic in general to expected utility theory, many balk at its apparent implications for cases of this latter type. Suppose you are choosing between a 'safe option' of saving a thousand lives for sure and a 'risky option' that gives a 1 in 1

<sup>26</sup> To see the distinction in Ellsberg's two-urns setting, suppose that in the status quo, one is set to receive \$100 iff the ambiguous urn delivers a red ball. Suppose one's choice is between whether to *add to that background gamble* a bet on a black ball being drawn from the risky urn, or instead from the ambiguous urn. Pretty clearly, ambiguity aversion in the standard sense will recommend the latter (since one then faces zero ambiguity overall), notwithstanding the fact that the benefit *delivered by one's action* is more ambiguous in this case.

<sup>27</sup> Beard, Row, and Fox (2020, Appendix A) and Sandberg and Bostrom (2008a) both present a wide range of estimates from around 1% to 50%, from (respectively) a literature review and a conference participant survey.

<sup>28</sup> We investigate the issues outlined in this paragraph in more depth in (Greaves et al, 2024).

<sup>29</sup> See e.g. (Al-Najjar and Weinstein 2009) for a survey of arguments that ambiguity aversion is irrational. Rowe and Voorhoeve (2018) and Stefánsson and Bradley (2019) defend its rationality.



trillion chance of saving a quintillion lives. The expected number of lives saved is a thousand times greater for the risky option. Unless the utility function is very non-linear as a function of lives saved, correspondingly, the expected utility of the latter option is also likely to be greater. Yet, if you choose the risky gamble, it is overwhelmingly likely that a thousand people will die, for no gain.<sup>30</sup>

Intuitively, it seems at least permissible to save the thousand in this case. If so, this might suggest that while expected utility theory is a good approach to choice under uncertainty in more ordinary cases, it fails in cases involving extremely low probabilities of extremely large values. One might, then, seek a ‘non-fanatical’ decision theory—one that does not require the agent to sacrifice arbitrarily much, with probability arbitrarily close to 1 in ‘fanatical’ pursuit of an extremely unlikely but enormously larger payoff. Might a non-fanatical decision theory undermine the case for strong longtermism?

We regard this as one of the most plausible ways in which the argument for strong longtermism might fail. Our view is that at present, the question cannot be confidently settled, since research into the possibility of a non-fanatical decision theory is currently embryonic. However, initial results suggest that avoiding fanaticism might come at too high a price.

Beckstead and Thomas (2020), for instance, consider a sequence of gambles. The first gamble delivers a large but relatively modest benefit with certainty. The last gamble delivers an enormously large benefit with extremely small probability, and zero benefit otherwise. These two gambles are linked by a sequence in which each gamble offers only a very *slightly* lower probability of winning than the previous gamble, and involves a *much* better benefit if one does win. This sequence-schema illustrates that any transitive theory that is not fanatical must instead be worryingly ‘timid’: in at least one pairwise comparison of adjacent gambles, even an arbitrarily large increase in the value of a positive payoff fails to compensate for any arbitrarily small decrease in its probability. As Beckstead and Thomas go on to show, such timidity in turn leads to implausibly extreme forms of risk aversion in some cases, and to particularly implausible forms of dependence of option-assessments on assessments of causally isolated aspects of the state of affairs.

A complementary reply is that in any case, the probabilities involved in the argument for longtermism might not be sufficiently extreme for any plausible degree of resistance to ‘fanaticism’ to overturn the verdicts of an expected-value approach, at least at the societal level. For example, it would not seem ‘fanatical’ to take action to reduce a 1 in 1 million risk of dying, as one incurs from cycling 35 miles or driving 500 miles (respectively, by wearing a helmet or wearing a seat belt (Department of Transport 2020)). But it seems that society can positively affect the very long-term future with probabilities well above this threshold. For instance, in subsection 4.3, we suggested a lower bound of 1 in 100,000 on a plausible credence that \$1 billion of carefully targeted spending would avert an existential catastrophe from artificial intelligence.

Things are less clear on the individual level. If, for example, \$10 billion can reduce the risk of extinction (or a comparably bad outcome) by 1 in 100,000, and an individual philanthropist makes a \$10,000 contribution with effects proportional to that, then the philanthropist would reduce extinction risk by 1 in 10 billion. At this level, we are unlikely to find commonplace decisions relying on that probability that we would regard as

<sup>30</sup> A similar example is that of Pascal’s Mugging (Bostrom 2009).



non-fanatical.<sup>31</sup> So, if one is inclined to take seriously the fanaticism worry, despite the problems with ‘timidity’, it may be that the probabilities in question are problematically small on the individual level, but not at the social level.

Our inclination is to think that our intuitions on the societal level are correct, and that our intuitions around how to handle very low probabilities are unreliable. The latter has some support from the psychological literature (Kahneman and Tversky 1979:282–283; Erev, Glozman, and Hertwig 2008).

We therefore tentatively conclude that considerations of fanaticism do not undermine the argument for strong longtermism.

## 9 Deontic strong longtermism

In section 2, we distinguished between axiological and deontic versions of strong longtermism. So far, our discussion has focused exclusively on the case for the axiological claim.

The deontic analogue to ASL is

**Deontic strong longtermism (DSL):** In the most important decision situations facing agents today,

- (i) One ought to choose an option that is near-best for the far future.
- (ii) One ought to choose an option that delivers much larger benefits in the far future than in the near future.

Just as ASL concerns *ex ante* axiology, the ‘ought’ in DSL is the subjective ought: the one that is most relevant for action-guidance, and is relative to the credences that the decision-maker ought to have.<sup>32</sup>

Without assuming consequentialism, DSL does not immediately follow from ASL. We believe, however, that our argument for ASL naturally grounds a corresponding argument for DSL. This is because of the following *stakes-sensitivity argument*:

(P1) When the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose a near-best option.

(P2) In the most important decision situations facing agents today, the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor.

(C) So, in the most important decision situations facing agents today, one ought to choose a near-best option.

<sup>31</sup> One exception might be putting on a seatbelt for a 1-mile drive. If doing so decreases one’s chance of a fatal accident by a factor of one-third, then the seatbelt reduces one’s risk of death by about 1 in 1 billion. But perhaps this is not our reason for wearing seatbelts for short journeys.

<sup>32</sup> It is widely agreed that either it is useful to distinguish between objective and subjective senses of ‘ought’ (Ewing 1948:118–122; Brandt 1959:360–367; Russell 1966; Parfit 1984:25; Gibbard 2005; Parfit 2011; Portmore 2011; Dorsey 2012; Olsen 2017), or ‘ought’ is univocal and subjective (Prichard 1932; Ross 1939:139; Howard-Snyder 2005; Zimmerman 2006; 2008; Mason 2013). Our discussion presupposes that one of these disjuncts is correct. A minority of authors holds that ‘ought’ is univocal and objective (Moore 1903:199–200, 229–230; Ross 1930:32; Thomson 1986:177–179; Graham 2010; Bykvist 2011). Similarly (but less discussed), one might be sceptical of the notion of *ex ante* axiology; again, our discussion of ASL has presupposed that any such scepticism is misguided.

DSL follows from the conjunction of (C) and ASL.

The stakes-sensitivity argument is obviously valid. Are its premises true?

(P1) appeals to only a very moderate form of stakes-sensitive non-consequentialism. It allows that there may be some actions that are always permissible or prohibited, no matter how great the axiological stakes: for example, perhaps one is always permitted to save the life of one's child; or perhaps one is always prohibited from torturing another person. And it only entails that comparatively minor prerogatives are overridden when the stakes are very high.<sup>33</sup>

It is highly plausible that there should be at least this much stakes-sensitivity. The lack of stakes-sensitivity is a common objection to Kant's notorious view that even if a friend's life depends on it, one should not tell a lie (Kant 1996). Turning to prerogatives, in 'emergency situations' like wartime, ordinary prerogatives—for instance, to consume luxuries, to live with one's family, and even to avoid significant risks to one's life—are quite plausibly overridden. Nagel (1978) observes that public morality tends to be more consequentialist in character than private morality; one natural partial explanation for this (though not the one emphasised by Nagel himself) is that in public contexts such as governmental policy decisions, the axiological stakes tend to be higher.

We foresee two lines of resistance to (P1). First, one could reject the idea of 'the good' altogether (Thomson 2008: sec. 1.4). On this view, there is simply no such thing as axiology. It's clear that our argument as stated would not be relevant to those who hold such views. But such a view must still be able to explain the fact that, in cases where there is a huge amount at stake, comparatively minor constraints and prerogatives get overridden. It seems likely that any such explanation will result in similar conclusions to those we have drawn, via similar arguments.

Second, and more plausibly, perhaps only some sorts of axiological considerations are relevant to determining what we ought to do. We consider two ways in which this idea might undermine our argument.

First, on a non-aggregationist view, comparatively small *ex ante* benefits to individuals are not relevant to determining what one ought to do, even if the benefits apply to an enormous number of people (Scanlon 1998:235; Voorhoeve 2014; Frick 2015).

Second, perhaps axiological considerations cannot outweigh non-consequentialist considerations when the axiological considerations involve altering the identities of who comes into existence (Parfit 1984: ch. 16).

However, both lines of thought risk proving too much. Let's first consider the non-aggregationist response. Consider a Briton, during WWII, deciding whether to fight; or someone debating whether to vote in their country's general election; or someone deciding whether to join an important political protest; or someone deciding whether to reduce their carbon footprint. In each case, the *ex ante* benefits to any particular other person are tiny. But in at least some such cases, it's clear that the agent is required to undertake the relevant action, and the most natural explanation of why is because the axiological stakes are so high.<sup>34</sup>

Second, consider the non-identity response. It's clear that governments ought to take significant action to fight climate change. But almost all of the expected damages from climate

<sup>33</sup> (P1) is very similar to Singer's claim that 'If it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do it' (Singer 1972:231).

<sup>34</sup> None of these examples, however, involves foregoing an opportunity to save many lives of identified people. In this respect, our examples are perhaps relevantly dissimilar to a decision between spending to benefit the far vs. the near future. We thank an anonymous referee for pressing this reply.

change come from its impacts on those who are yet to be born.<sup>35</sup> What's more, any policy designed to mitigate climate change will also affect the identities of those unborn people. Endorsing the non-identity response would therefore risk rejecting the idea that welfarist considerations generate any obligations for society today to fight climate change, even while accepting that climate change will significantly and avoidably reduce welfare in expectation for centuries to come. That position is clearly incorrect.

Turning now to (P2): The 'high-stakes' aspect of this premise is justified in part on the basis of the arguments of sections 3–4. At least on our main and low estimates of the expected size of the future, in the decision situations we've discussed, not only are the best options those that have the near-best far-future consequences, but they are *much* better than those options whose far-future consequences are nowhere near best.

At the same time, at least for most members of rich countries, the decision situations we've discussed are those where the personal prerogatives are arguably comparatively minor, and where there are no serious side-constraints. This is clearest in the cases of individual decisions about *where to direct* one's altruistic spending (holding fixed the total size of one's 'altruistic budget'), and about career choice. The decision to give to organisations that will positively influence the far future rather than organisations more geared towards improving the near future, or to work in a career that is particularly beneficial for the long-term future, might well involve some sacrifices.<sup>36</sup> But they are not close to the sorts of sacrifices where there might be absolute or near-absolute prerogatives. Similarly, these are not circumstances where one is required to violate side-constraints in order to achieve the near-best long-term outcome.

The slightly less clear cases are those involving individual decisions about the total *size* of one's 'altruistic budget' (vs. 'personal budget'), and societal decisions about how many resources to devote to improving the prospects for the far future (vs. the near future, including the lifetimes of present people). Here, it remains true that no serious side-constraints need be involved. One might worry, though, that here our argument will be too demanding: might it imply that we, individually or as a society, ought to devote *most* of our resources to improving the far future, at the large expense of our own prudential interests?

As in the discussion of demandingness in the context of global poverty, a range of responses to this concern is possible. We have nothing to add to the existing literature on demandingness (e.g. Kagan 1984; Mulgan 2001; Hooker 2009).<sup>37</sup> We will simply note that even if, for example, there is an absolute cap on the total sacrifice that can be morally required, it seems implausible that society today is currently anywhere near that cap. The same remark applies to at least the vast majority of individuals in rich countries. We ought to be doing a lot more for the far future than we currently are.<sup>38</sup>

<sup>35</sup> For example, the Stern Review predicts the vast majority of damages to occur after 2100 in both 'baseline' and 'high climate' scenarios (Stern 2007:178, fig. 6.5d).

<sup>36</sup> There are, however, reasons to think that these sacrifices are not as great as we might initially suppose (MacAskill, Mogensen, and Ord 2018).

<sup>37</sup> Mogensen (2021) discusses specifically the relationship between demandingness and longtermism.

<sup>38</sup> Might our arguments go further than this, and justify atrocities in the name of the long-term good? Perhaps the French Revolution had good long-term consequences, in terms of bringing about a more liberal and democratic world: does strong longtermism, if so, justify the guillotine?

We do not think so, for at least two reasons. The first is that, for such serious side-constraints, something closer to absolutism or near-absolutism becomes much more plausible (or, at least, it takes more than mere *ex ante* goodness to justify violation of those side-constraints). The second is that, in almost all cases, when there is some option available that promotes the long-term good while violating a serious side-constraint, there will be some alternative option available that achieves a similar amount of long-term good without violating that side-constraint. Liberal democracy could have been achieved in France without the Reign of Terror.

## 10 Summary and conclusions

The potential future of civilisation is vast. Once we appreciate this, it becomes plausible that impact on the far future is the most important feature of our actions today.

Strong longtermism would be false in a world that had sufficiently weak causal connections between the near and the distant future, such that it was too difficult to significantly influence the course of the very long-run future. However, we have argued, the world we find ourselves in today does not have this feature.

We presented our central case in terms of (i) a total utilitarian axiology and (ii) an expected-value treatment of decision-making under uncertainty. However, we argued, plausible deviations from either or both of these assumptions do not undermine the core argument.

This essay mainly focused on the decision situations of a society or individual considering how to spend money without constraints as to cause area, and of an individual's career choice. We argued that these are situations where we can in expectation significantly influence the far future. Precisely because of this, they are among the most important decision situations we face, and axiological strong longtermism follows.

In our own view, the weakest points in the case for axiological strong longtermism are the assessment of numbers for the cost-effectiveness of particular attempts to benefit the far future, the appropriate treatment of cluelessness, and the question of whether an expected-value approach to uncertainty is too 'fanatical' in this context. These issues in particular would benefit from further research.

In addition to axiological issues, we also discussed the counterpart deontic issues. We suggested that deontic strong longtermism might well be true even if consequentialism is false, on the grounds that (i) the stakes involved are very high, (ii) a plausible non-consequentialist theory has to be sensitive to the axiological stakes, becoming more consequentialist in output as the axiological stakes get higher, and (iii) in the key decision situations, any countervailing constraints and/or prerogatives are comparatively minor. Quite plausibly, in the world as it is today, the most important determinants of what we ought to do arise from our opportunities to affect the far future.

It is possible, but far from obvious, that far-future impacts are also more important than near-future impacts in a much wider class of decision situations: for instance, decisions about whether or not to have a child, and government policy decisions within a relatively narrow 'cause area'. Insofar as they are, strong longtermism could potentially set a methodology for further work in applied ethics and applied political philosophy: for each issue in these subfields, one could identify the potential far-future effects from different actions or policies, and then work through how these bear on the issue in question. The answers might sometimes be surprisingly revisionary.

## Appendix

We claimed in the main text that (BR) entails:

- (a) that ASL(i) holds of a restriction of society's decision situation, obtained by removing any options involving net expected short-term harm from the choice set; and
- (b) that ASL(ii) holds of society's decision situation.

Here, we make these claims precise, and supply the proofs for them.

## Terminology and notation

For any option  $x$ , let  $N(x)$ ,  $F(x)$ ,  $V(x)$  respectively denote  $x$ 's near-future, far-future, and overall benefits. Let  $N^*$ ,  $F^*$ ,  $V^*$  respectively be the highest available near-future, far-future, and overall benefits. Let  $F'$  be the highest far-future benefit that is available without net short-term harm.

We interpret both 'near-best overall' and 'near-best for the far future' in terms of proportional distance from zero benefit to the maximum available benefit, and 'much larger' in terms of a multiplicative factor. There is, of course, flexibility on the precise values of the factors involved. We therefore consider the following precisifications of our key claims, carrying free parameters:

$BR(n)$ :  $F' \geq nN^*$ .

$ASL_i(\epsilon_o, \epsilon_f)$ : Every option that delivers overall benefits of at least  $(1 - \epsilon_o)V^*$  delivers far-future benefits of at least  $(1 - \epsilon_f)F^*$ .

$ASL_{ii}(\epsilon_o, r)$ : Every option that delivers overall benefits of at least  $(1 - \epsilon_o)V^*$  delivers far-future benefits that are at least  $r$  times its own near-future benefits.

In what follows, we prove claims (a) and (b) for specified relationships between the parameter values.

**Precisification of claim (a).** We claim (more precisely) that if  $BR(n)$  holds of a given decision situation, then for any  $\epsilon_o \in [0, 1]$ ,  $ASL_i(\epsilon_o, \epsilon_o + \frac{1}{n})$  holds of the restricted decision situation (with any options involving net short-term harm removed). For example, if  $n = 10$ , then every option that delivers at least 90% of available overall expected benefits delivers at least 80% of available far-future expected benefits, once any options involving net short-term harm are ruled out.

**Proof.** Suppose that  $BR(n)$  holds. Since far-future benefit  $F'$  is attainable without near-future net harm, the overall best option must deliver total benefits of at least  $F'$ ; so any near-best option must deliver total benefits of at least  $(1 - \epsilon_o)F'$ . But by  $BR(n)$ , the maximum attainable near-future benefit is at most  $\frac{F'}{n}$ . Therefore, any near-best option must deliver far-future benefits of at least  $(1 - \epsilon_o - \frac{1}{n})F'$ . But in this decision situation,  $F' = F^*$  (since near-future net harm is here ruled out).

**Precisification of claim (b).** We claim (more precisely) that if  $BR(n)$  holds then for any  $\epsilon_o \in [0, 1]$ ,  $ASL_{ii}(\epsilon_o, (1 - \epsilon_o)n - 1)$  also holds. For example, if  $n = 10$ , then every option that delivers at least 90% of available overall expected benefits delivers at least eight times as much far-future as near-future expected benefit.

**Proof.** Let  $x$  be any option that is near-best overall. Then

$$\begin{aligned} V(x) &\geq (1 - \epsilon_o)V^* && \text{by definition of near-best} \\ &\geq (1 - \epsilon_o)F' && \text{since, by hypothesis, } F' \text{ is achievable without short-term harm} \end{aligned}$$

But  $V(x) = N(x) + F(x)$ , so it follows that

$$\begin{aligned} F(x) &\geq (1 - \epsilon_o)F' - N(x) \\ &\geq (1 - \epsilon_o)F' - N^* \\ &\geq ((1 - \epsilon_o)n - 1)N^* && \text{by } BR(n) \\ &\geq ((1 - \epsilon_o)n - 1)N(x) \end{aligned}$$

## References

- Adams, F. C. (2008), 'Long-Term Astrophysical Processes', in N. Bostrom and M. Cirkovic (eds.), *Global Catastrophic Risks* (Oxford University Press), 33–47.
- Adams, F. C. and Laughlin, G. (1999), *The Five Ages of the Universe: Inside the Physics of Eternity* (Free Press).
- Al-Najjar, N. I. and Weinstein, J. (2009), 'The Ambiguity Aversion Literature: A Critical Assessment', *Economics & Philosophy* 25/3: 249–284.
- Althaus, D. and Gloor, L. (2018), 'Reducing Risks of Astronomical Suffering: A Neglected Priority', *Foundational Research Institute*, <https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/> (accessed 22 January 2025).
- Altman, S. (2015), 'Machine Intelligence, Part 1', <https://blog.samaltman.com/machine-intelligence-part-1>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016), 'Concrete Problems in AI Safety', <https://arxiv.org/abs/1606.06565>
- Armstrong, S. and Sandberg, A. (2013), 'Eternity in Six Hours: Intergalactic Spreading of Intelligent Life and Sharpening the Fermi Paradox', *Acta Astronautica* 89:1–13.
- Arnold, R., Scott Marcus, J., Petropoulos, G., and Schneider, A. (2018), 'Is Data the New Oil? Diminishing Returns to Scale', *29th European Regional Conference of the International Telecommunications Society*.
- Arrhenius, G. and Bykvist, K. (1995), 'Future Generations and Interpersonal Compensations: Moral Aspects of Energy Use', *Uppsala Prints and Preprints in Philosophy* 21 (Uppsala Universitet).
- Bader, R. (n.d.), 'Person-Affecting Population Ethics' (unpublished manuscript).
- Barnosky, A., Matzke, N., Tomiya, S., Wogan, G., Swartz B., Quental, T., Marshall, C., McGuire, J., Lindsey, E., Maguire, K., Mersey, B., and Ferrer E. (2011), 'Has the Earth's Sixth Mass Extinction Already Arrived?' *Nature* 471/7336: 51–57.
- Beard, S., Rowe, T., and Fox J. (2020), 'An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards', *Futures* 115: 102469.
- Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, PhD thesis, Rutgers University, <https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/>
- Beckstead, N. (2014), 'Will We Eventually Be Able to Colonize Other Stars? Notes from a Preliminary Review', <https://www.fhi.ox.ac.uk/will-we-eventually-be-able-to-colonize-other-stars-notes-from-a-preliminary-review/> (accessed 22 January 2025).
- Beckstead, N. and Thomas, T. (2020), 'A Paradox for Tiny Probabilities and Enormous Values', GPI Working Paper 10-2020 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/> (accessed 22 January 2025)
- Benatar, D. (2006), *Better Never to Have Been: The Harm of Coming into Existence* (Clarendon Press).
- Berger, A. (2019), 'GiveWell's Top Charities Are (Increasingly) Hard to Beat', *Open Philanthropy Project*, <https://www.openphilanthropy.org/blog/givewells-top-charities-are-increasingly-hard-beat> (accessed 26 January 2021).
- Bewley, T. F. (2002), 'Knightian Decision Theory. Part I', *Decisions in Economics and Finance*, 25/2: 79–110.
- Blackorby, C., Bossert, W., and Donaldson, D. (1996), 'Quasi-orderings and Population Ethics', *Social Choice and Welfare* 13: 129–150.
- Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020), 'Are Ideas Getting Harder to Find?', *American Economic Review* 110/4: 1104–1144.
- Bollard, L. (2016), 'Initial Grants to Support Corporate Cage-free Reforms', *Open Philanthropy Project*, [https://www.openphilanthropy.org/blog/initial-grants-support-corporate-cage-free-reforms#Corporate\\_cage-free\\_campaigns\\_are\\_extremely\\_cost-effective](https://www.openphilanthropy.org/blog/initial-grants-support-corporate-cage-free-reforms#Corporate_cage-free_campaigns_are_extremely_cost-effective) (accessed 26 January 2021).
- Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas* 15/3: 308–314.
- Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* 69/3: 443–445.
- Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', *Global Policy* 4/1: 15–31.
- Bostrom, N. (2014), *Superintelligence: Path, Dangers, Strategies* (Oxford University Press).
- Bradley, R. (2017), *Decision Theory with a Human Face* (Cambridge University Press).
- Brandt, R. (1959), *Ethical Theory* (Prentice-Hall).
- Bricker, D. and Ibbitson, J. (2019), *Empty Planet: The Shock of Global Population Decline* (Crown Publishing Group).
- Broome, J. (2008), 'The Ethics of Climate Change', *Scientific American* 298: 96–102.
- Bryson, S., Kunimoto, M., Kopparapu, R., and Zamudio, K. (2021), 'The Occurrence of Rocky Habitable-Zone Planets around Solar-Like Stars from Kepler Data', *The Astronomical Journal* 161/1: 36.
- Buchak, L. (2013), *Risk and Rationality* (Oxford University Press).



- Buchholz, W. and Schumacher, J. (2010), 'Discounting and Welfare Analysis Over Time: Choosing the  $\eta$ ', *European Journal of Political Economy* 26/3: 372–385.
- Bykvist, K. (2011), 'How to Do Wrong Knowingly and Get Away with It', in F. Svensson and R. Śliwiński (eds.), *Neither/Nor. Philosophical Papers Dedicated to Erik Carlson on the Occasion of His Fiftieth Birthday*. Uppsala Philosophical Studies 58. (Department of Philosophy, Uppsala University), 31–47.
- Caplan, B. (2008), 'The Totalitarian Threat', in N. Bostrom and M. Cirkovic (eds.) *Global Catastrophic Risks* (Oxford University Press), 504–520.
- Cassman, K. G., Dobermann, A. R., and Walters, D. T. (2002), 'Agroecosystems, Nitrogen-Use Efficiency, and Nitrogen Management', *AMBIO: A Journal of the Human Environment* 31/2: 132–140.
- Chalmers, D. J. (1996), *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press).
- Christiano, P. (2014), 'We Can Probably Influence the Far Future', *Rational Altruist*, <https://rationalaltruist.com/2014/05/04/we-can-probably-influence-the-far-future/> (accessed 22 January 2025).
- Cline, W. R. (1992), *The Economics of Global Warming* (Institute for International Economics).
- Cohen, J. E. (1995), *How Many People Can the Earth Support?* (Norton).
- Cowen, T. (1992), 'Consequentialism Implies a Zero Rate of Intergenerational Discount', in P. Laslett and J. S. Fishkin (eds.), *Justice Between Age Groups and Generations* (Yale University Press), 162–168.
- Dasgupta, P. (2008), 'Discounting Climate Change', *Journal of Risk and Uncertainty* 37: 141–169.
- Department of Transport. (2020), 'Reported Road Casualties in Great Britain: Provisional Results 2019', [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/904698/rrcgb-provisional-results-2019.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/904698/rrcgb-provisional-results-2019.pdf) (accessed 28 May 2021).
- Dietz, S., Hepburn, C. J., and Stern, N. (2008), 'Economics, Ethics and Climate Change', in K. Basu and R. Kanbur (eds.), *Arguments for a Better World: Essays in Honour of Amartya Sen. Volume II: Society, Institutions and Development* (Oxford University Press), 365–386.
- Dorsey, D. (2012), 'Objective Morality, Subjective Morality and the Explanatory Question', *Journal of Ethics and Social Philosophy* 6/3: 1–24.
- Drupp, M. A., Freeman, M., Groom B., and Nesje F. (2018), 'Discounting Disentangled', *American Economic Journal: Economic Policy* 10: 109–134.
- Dubra, J., Maccheroni F., and Ok, E. A. (2004), 'Expected Utility Theory without the Completeness Axiom', *Journal of Economic Theory* 115/1: 118–133.
- Elga, A. (2010), 'Subjective Probabilities Should Be Sharp', *Philosopher's Imprint* 10/5: 1–11.
- Ellsberg, D. (1961), 'Risk, Ambiguity, and the Savage Axioms', *The Quarterly Journal of Economics* 75/4: 643–669.
- Erev, I., Glazman, I., and Hertwig, R. (2008), 'What Impacts the Impact of Rare Events', *Journal of Risk and Uncertainty* 36/2: 153–177.
- Ewing, A. C. (1948), *The Definition of Good* (Routledge & Kegan Paul).
- Finnveden, L., Riedel, J., and Shulman, C. (2022), 'AGI and Lock-in', *The Effective Altruism Forum*. <https://forum.effectivealtruism.org/posts/KqCybin8rtfP3qztq/agi-and-lock-in>
- Flyvbjerg, B. (2008), 'Curbing Optimism Bias and Strategic Misrepresentation in Planning: Reference Class Forecasting in Practice', *European Planning Studies* 16/1: 3–21.
- Frick, J. (2015), 'Contractualism and Social Risk', *Philosophy & Public Affairs* 43/3: 175–223.
- Galaabaatar, T. and Karni, E. (2013), 'Subjective Expected Utility with Incomplete Preferences', *Econometrica* 81/1: 255–284.
- Gibbard, A. (2005), 'Truth and Correct Belief', *Philosophical Issues* 15: 338–350.
- GiveWell. (2018), 'Estimating the Funding Gaps for Distribution of Antimalarial Nets and Seasonal Malaria Chemoprevention', [https://www.givewell.org/international/technical/programs/malaria-funding-gaps#What\\_would\\_it\\_cost\\_to\\_deliver\\_nets\\_to\\_everyone\\_who\\_needed\\_them](https://www.givewell.org/international/technical/programs/malaria-funding-gaps#What_would_it_cost_to_deliver_nets_to_everyone_who_needed_them) (accessed 9 February 2021).
- GiveWell. (2020a), 'GiveWell's Cost-Effectiveness Analyses', <https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models> (accessed 26 January 2021).
- GiveWell. (2020b), 'Our Criteria for Top Charities', <https://www.givewell.org/how-we-work/criteria#Criteria> (accessed 28 May 2021).
- Gollier, C. (2013), *Pricing the Planet's Future: The Economics of Discounting in an Uncertain World* (Princeton University Press).
- Good, I. J. (1966), 'Speculations Concerning the First Ultrainelligent Machine', in F. L. Alt and M. Rubinfeld (eds.), *Advances in Computers* 6. (Academic Press), 31–88.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans O. (2018), 'When Will AI Exceed Human Performance? Evidence from AI Experts', *Journal of Artificial Intelligence Research* 62: 729–754.
- Graham, P. A. (2010), 'In Defense of Objectivism about Moral Obligation', *Ethics* 121: 88–115.
- Greaves, H. (2016), 'Cluelessness', *Proceedings of the Aristotelian Society* 116/3: 311–339.
- Greaves, H. (2017), 'Discounting for Public Policy: A Survey', *Economics and Philosophy* 33/3: 391–439.



- Greaves, H., Thomas, T., Mogensen, A., and MacAskill, W. (2024), 'On The Desire to Make a Difference', *Philosophical Studies* 181: 1599–1626.
- Harrod, R. F. (1948), *Towards a Dynamic Economics* (Macmillan).
- Hawking, S. (2018), *Brief Answers to the Big Questions* (John Murray Press).
- Heyd, D. (1988), 'Procreation and Value: Can Ethics Deal with Futurity Problems?' *Philosophia* 18/2–3: 151–170.
- Hooker, B. (2009), 'The Demandingness Objection', in T. Chappell (ed.), *The Problem of Moral Demandingness* (Palgrave MacMillan), 148–162.
- House of Commons Public Administration Select Committee. (2013), 'The role of the Charity Commission and "public benefit": Post-legislative scrutiny of the Charities Act 2006'. Third Report of Session 2013–14. Volume I: Report, together with formal minutes, oral and written evidence, <https://publications.parliament.uk/pa/cm201314/cmselect/cmpubadm/76/7602.htm> (accessed 22 January 2025).
- Howard-Snyder, F. (2005), 'It's the Thought that Counts', *Utilitas* 17: 265–281.
- Isaacson, W. (2003), *Benjamin Franklin: An American Life* (Simon & Schuster).
- Jordà, Ò., Knoll, K., Kuvshinov, D., Schularick, M., and Taylor, A. M. (2019), 'The Rate of Return on Everything, 1870–2015', *The Quarterly Journal of Economics* 134/4: 1225–1298.
- Kagan, S. (1984), 'Does Consequentialism Demand Too Much? Recent Work on the Limits of Obligation', *Philosophy & Public Affairs* 13/3: 239–254.
- Kahneman, D. and Lovallo, D. (1993), 'Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking', *Management Science* 39/1: 17–31.
- Kahneman, D. and Tversky, A. (1979), 'Prospect Theory: An Analysis of Decision under Risk', *Econometrica* 47/2: 263–291.
- Kant, I. (1996), 'On a Supposed Right to Lie from Philanthropy', in M. Gregor (transl./ed.) *Immanuel Kant: Practical Philosophy* (Cambridge University Press), 605–616.
- Legg, S. (2008), *Machine Super Intelligence*, PhD thesis, University of Lugano.
- Lenman, J. (2000), 'Consequentialism and Cluelessness', *Philosophy & Public Affairs* 29/4: 342–370.
- Lewis, D. (1980), 'Mad Pain and Martian Pain', *Readings in the Philosophy of Psychology* 1: 216–222.
- Lewis, G. (2016), 'Beware Surprising and Suspicious Convergence', *The Effective Altruism Forum* <https://forum.effectivealtruism.org/posts/omoZDu8ScNbot6kXS/beware-surprising-and-suspicious-convergence> (accessed 22 January 2025).
- Liao, S. M. (2020), 'The Moral Status and Rights of Artificial Intelligence', in S. M. Liao (ed.), *Ethics of Artificial Intelligence* (Oxford University Press), 480–498.
- MacAskill, W. (2014), 'Replaceability, Career Choice, and Making a Difference', *Ethical Theory and Moral Practice* 17: 269–283.
- MacAskill, W. (2019), 'When Should an Effective Altruist Donate?', GPI Working Paper 8-2019 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/william-macaskill-when-should-an-effective-altruist-donate/> (accessed 22 January 2025).
- MacAskill, W. (2022), *What We Owe The Future* (Basic Books).
- MacAskill, W., Mogensen, A., and Ord, T. (2018), 'Giving Isn't Demanding', in P. Woodruff (ed.), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy* (Oxford University Press), 178–203.
- Machina, M. J. and Siniscalchi, M. (2014), 'Ambiguity and Ambiguity Aversion', in *Handbook of the Economics of Risk and Uncertainty, Volume 1* (North Holland), 729–807.
- Mason, E. (2013), 'Objectivism and Prospectivism about Rightness', *Journal of Ethics and Social Philosophy* 7/2: 1–21.
- Meacham, C. (2012), 'Person-Affecting Views and Saturating Counterpart Relations', *Philosophical Studies* 158/2: 257–287.
- Meng, F. and Ellis, T. (2020), 'The Second Decade of Synthetic Biology: 2010–2020', *Nature Communications* 11: 5174.
- Millett, P. and Snyder-Beattie, A. (2017), 'Existential Risk and Cost-Effective Biosecurity', *Health Security* 15/4: 373–383.
- Mogensen, A. (2019), 'The Only Ethical Argument for Positive ?' GPI Working Paper 5-2019 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/andreas-mogensen-the-only-ethical-argument-for-positive-delta-2/> (accessed 22 January 2025).
- Mogensen, A. (2021), 'Moral Demands and the Far Future', *Philosophy and Phenomenological Research* 101/3: 567–585.
- Moore, G. E. (1903), *Principia Ethica* (Cambridge University Press).
- Mulgan, T. (2001), *The Demands of Consequentialism* (Oxford University Press).

- Musk, E. (2014), 'One-on-one with Elon Musk. Talk at the MIT AeroAstro 1914-2014 Centennial Symposium', <https://aerastro.mit.edu/videos/centennial-symposium-one-one-one-elon-musk> (accessed 22 January 2025).
- Nagel, T. (1978), 'Ruthlessness and Public Life', in S. Hampshire (ed.), *Public and Private Morality* (Cambridge University Press), 75–91.
- Newberry, T. (2021a), 'How Many Lives Does the Future hold?', GPI Technical Report T2-2021 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/how-many-lives-does-the-future-hold-toby-newberry-future-of-humanity-institute-university-of-oxford/> (accessed 22 January 2025).
- Newberry, T. (2021b), 'How Cost-Effective Are Efforts to Detect Near-Earth-Objects?', GPI Technical Report T1-2021 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/how-cost-effective-are-efforts-to-detect-near-earth-objects-toby-newberry-future-of-humanity-institute-university-of-oxford/> (accessed 22 January 2025).
- Ngo, R. (2020), 'AGI Safety from First Principles', <https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view> (accessed 22 January 2025).
- O'Donoghue, T. and Somerville, J. (2018), 'Modeling Risk Aversion in Economics', *The Journal of Economic Perspectives* 32/2: 91–114.
- Ochs, A. (2019), 'What Is the Oldest Charitable Trust in the U.S., and What Does It Fund?', *Inside Philanthropy*, <https://www.insidephilanthropy.com/home/2019/4/21/what-is-the-oldest-charitable-trust-in-the-us-and-what-does-it-fund> (accessed 22 January 2025).
- Olsen, K. (2017), 'A Defense of the Objective/Subjective Moral Ought Distinction', *The Journal of Ethics* 21/4: 351–373.
- Open Philanthropy Project. (2020), Grants database, <https://www.openphilanthropy.org/giving/grants> (accessed 26 January 2021).
- Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury).
- Parfit, D. (1984), *Reasons and Persons* (Clarendon Press).
- Parfit, D. (2011), *On What Matters: Volume 1*, ed. S. Scheffler (Oxford University Press).
- Pierson, P. (2000), 'Increasing Returns, Path Dependence, and the Study of Politics', *The American Political Science Review* 94/2: 251–267.
- Pigou, A. C. (1932), *The Economics of Welfare: Volume 1*, 4th edition (Macmillan).
- Portmore, D. W. (2011), *Commonsense Consequentialism: Wherein Morality Meets Rationality* (Oxford University Press).
- Posner, R. (2004), *Catastrophe: Risk and Response* (Oxford University Press).
- Pratt, J. W. (1964), 'Risk Aversion in the Small and in the Large', *Econometrica* 32/1–2: 122–136.
- Prichard, H. A. (1932) [2002], 'Duty and Ignorance of Fact', in J. McAdam (ed.), *Moral Writings* (Oxford University Press), 85–110.
- Quiggin, J. (1982), 'A Theory of Anticipated Utility', *Journal of Economic Behavior and Organization* 3: 323–343.
- Quiggin, J. and Wakker, P. (1994), 'The Axiomatic Basis of Anticipated Utility: A Clarification', *Journal of Economic Theory* 64: 486–499.
- Rabin, M. (2000), 'Risk Aversion and Expected-Utility Theory: A Calibration Theorem', *Econometrica* 68/5: 1281–1292.
- Ramsey, F. P. (1928), 'A Mathematical Theory of Saving', *Economic Journal* 38: 543–559. (Reprinted in F. P. Ramsey, *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*, ed. D. H. Mellor.)
- Rees, M. (2018), *On the Future: Prospects for Humanity* (Princeton University Press).
- Reese, J. (2018), 'Why I Prioritize Moral Circle Expansion over Artificial Intelligence Alignment', *Effective Altruism Forum*, <https://forum.effectivealtruism.org/posts/BY8gXSpGijypbGitT/why-i-prioritize-moral-circle-expansion-over-artificial> (accessed 9 February 2021).
- Ross, W. D. (1930), *The Right and the Good* (Oxford University Press).
- Ross, W. D. (1939), *The Foundations of Ethics* (Clarendon Press).
- Rowe, T. and Voorhoeve, A. (2018), 'Egalitarianism under Severe Uncertainty', *Philosophy & Public Affairs* 46/3: 239–268.
- Russell, B. (1966), 'The Elements of Ethics', in *Philosophical Essays* (Simon and Schuster), 13–59.
- Russell, S. (2019), *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking).
- Sackmann, I.-J., Boothroyd, A. I., and Kraemer, K. E. (1993), 'Our Sun. III. Present and Future', *The Astrophysics Journal* 418: 457–468.
- Sandberg, A. (2013), 'Feasibility of Whole Brain Emulation', in V. C. Müller (ed.), *Theory and Philosophy of Artificial Intelligence* (Springer), 251–64.

- Sandberg, A. (2014), 'Ethics of Brain Emulations', *Journal of Experimental and Theoretical Artificial Intelligence* 26/3: 439–457.
- Sandberg, A. and Bostrom, N. (2008a), 'Global Catastrophic Risks Survey', Future of Humanity Institute Technical Report #2008-1, <https://www.fhi.ox.ac.uk/reports/2008-1.pdf> (accessed 22 January 2025).
- Sandberg, A. and Bostrom, N. (2008b), 'Whole Brain Emulation: A Roadmap', Future of Humanity Institute Technical Report #2008-3, <https://www.fhi.ox.ac.uk/reports/2008-3.pdf> (accessed 22 January 2025).
- Sarbu, I. and Sebarchievici, C. (2017), *Solar Heating and Cooling Systems: Fundamentals, Experiments and Applications* (Elsevier).
- Scanlon, T. M. (1998), *What We Owe to Each Other* (Harvard University Press).
- Schröder, K.-P. and Smith, R. C. (2008), 'Distant Future of the Sun and Earth Revisited', *Monthly Notices of the Royal Astronomical Society* 386/1: 155–163.
- Setiya, K. (2014), 'The Ethics of Existence', *Philosophical Perspectives* 28/1: 291–301.
- Shulman, C. (2020), 'What Do Historical Statistics Teach Us about the Accidental Release of Pandemic Bioweapons?', *Reflective Disequilibrium*, <http://reflectivedisequilibrium.blogspot.com/2020/10/what-do-historical-statistics-teach-us.html> (accessed 22 January 2025).
- Sidgwick, H. (1890), *The Methods of Ethics* (Macmillan).
- Simon, J. (1998), *The Ultimate Resource 2*, revised edition (Princeton University Press).
- Singer, P. (1972), 'Famine, Affluence and Morality', *Philosophy & Public Affairs* 1/3: 229–243.
- Snyder-Beattie, A., Ord, T., and Bonsall M. B. (2019), 'An Upper Bound for the Background Rate of Human Extinction', *Scientific Reports* 9: 11054.
- Solow, R. M. (1974), 'The Economics of Resources or the Resources of Economics', *American Economic Review Papers and Proceedings* 64: 1–14.
- Statt, N. (2015), 'Bill Gates Is Worried about Artificial Intelligence Too', *CNET*, <https://www.cnet.com/news/bill-gates-is-worried-about-artificial-intelligence-too/> (accessed 22 January 2025).
- Steele, K. and Stefánsson, H. O. (2021), *Beyond Uncertainty: Reasoning with Unknown Possibilities* (Cambridge University Press).
- Stefánsson, H. O. and Bradley, R. (2019), 'What Is Risk Aversion?', *The British Journal for the Philosophy of Science* 70/1: 77–102.
- Stern, N. (2007), *The Economics of Climate Change* (Cambridge University Press).
- Stix, M. (2002), *The Sun: An Introduction* (Springer).
- Tarsney, C. (2019), 'The Epistemic Challenge to Longtermism', GPI Working Paper 10-2019 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/christian-tarsney-the-epistemic-challenge-to-longtermism/>
- Tegmark, M. (2017), *Life 3.0: Being Human in the Age of Artificial Intelligence* (Knopf).
- Texas Education Agency. (2020), Texas Permanent School Fund <https://tea.texas.gov/finance-and-grants/texas-permanent-school-fund> (accessed 22 January 2025).
- Thomas, T. (2019), 'The Asymmetry and the Long Term', GPI Working Paper 11-2019 (Global Priorities Institute, Oxford University), <https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/> (accessed 22 January 2025).
- Thomson, J. J. (1986), 'Imposing Risks', in W. Parent (ed.), *Rights, Restitution, and Risks* (Harvard University Press), 173–191.
- Thomson, J. J. (2008), *Normativity* (Open Court).
- Trammell, P. (2020), 'Dynamic Public Good Provision under Time Preference Heterogeneity: Theory and Applications to Philanthropy', <https://philiptrammell.com/static/PatienceAndPhilanthropy.pdf> (accessed 22 January 2025).
- Trautmann, S. T. and van de Kuilen, G. (2015), 'Ambiguity Attitudes', in G. Keren and G. Wu (eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making, I.* (Wiley), 89–116.
- Turing, A. (1951), 'Can Digital Computers Think?' *BBC Radio*, May 15 1951. <https://www.bbc.co.uk/vid/eos/cqv8n1d8dxo>
- United Nations, Department of Economic and Social Affairs, Population Division. (2019), *World Population Prospects 2019: Highlights*, World Population Prospects 2019 Highlights (accessed 22 January 2025).
- University of Texas System. (2021), 'The Permanent University Fund (PUF)', <https://www.utsystem.edu/puf> (accessed 22 January 2025).
- Van Den Bergh, J. C. J. M. and Rietveld, P. (2004), 'Reconsidering the Limits to World Population: Meta-analysis and Meta-prediction', *BioScience* 54/3: 195–204.
- Voorhoeve, A. (2014), 'How Should We Aggregate Competing Claims?' *Ethics* 125: 64–87.
- Wolf, E. and Toon, O. B. (2015), 'The Evolution of Habitable Climates under the Brightening Sun', *Journal of Geophysical Research: Atmospheres* 120: 5775–5794.

- Yudkowsky, E. (2013), 'Intelligence Explosion Microeconomics', Technical Report 2013-1 (Machine Intelligence Research Institute).
- Zimmerman, M. J. (2006), 'Is Moral Obligation Objective or Subjective?', *Utilitas* 18: 329–61.
- Zimmerman, M. J. (2008), *Living with Uncertainty: The Moral Significance of Ignorance* (Cambridge University Press).