The case for strong longtermism

Hilary Greaves, William MacAskill

Global Priorities Institute | September 2019

GPI Working Paper No. 7-2019

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

The Case for Strong Longtermism

Hilary Greaves and William MacAskill

Work in Progress

1. Introduction

A striking fact about the history of civilisation is just how early we are in it. There are 5000 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 200,000 years to go; there are a further one billion years until the Earth is sterilized by the Sun; and trillions of years until the last conventional star formations. Even on the most conservative of these timelines, we have progressed through a tiny fraction of recorded history. If humanity's saga were a novel, we would still be on the very first page.

Normally, we pay scant attention to this fact. Political discussions are centered 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. We essentially never think about how our actions today might influence civilisation in hundreds of thousands of years hence.

We believe that this neglect of the very long-term future is a grave moral error.[1] An alternative perspective is given by a burgeoning view called *longtermism*,[2] on which we should be particularly concerned with ensuring that the long-run future goes well. In this article we accept this view but go further, arguing that impacts on the long run are the *most* important feature of our actions. More precisely, we argue for two claims.

> **Axiological strong longtermism (AL)**: In a wide class of decision situations, the option that is *ex ante* best is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.
>
> **Deontic strong longtermism (DL**): In a wide class of decision situations, the option one ought, ex ante, to choose is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.

By "the option whose effects on the very long-run future are best", what we mean is "the option whose effects on the future *from time t onwards* are best", where t is a surprisingly long time from now (say, 100 or even 1000 years). The idea, then, is that for the purposes of evaluating actions, we can in the first instance often *simply ignore* all the effects contained in the first 100 (or even 1000) years, focussing primarily on the further-future effects. Short-run effects act as little more than tie-breakers.

Note that both AL and DL are phrased in *ex ante* terms. AL concerns *ex ante* axiology. If expected value theory is the correct account of how to order uncertain prospects in terms of their betterness then, given AL, the *ex ante* best action would be one whose possible effects on the very long-run future do most (or nearly the most) to increase expected value.

---

[1] It is useful to consider a prudential analogue: if we were to live until the Earth were no longer habitable, how much attention would it be prudentially rational for us to pay to ensuring the very long-run future goes well? Presumably, far more than we would do now.

[2] See MacAskill (2019) for a discussion of this idea.

However, the longtermist claim does not essentially presuppose expected value theory; we briefly consider some alternatives in section 4. Similarly, for DL, the 'ought' in question is the 'subjective' ought: the one that is most relevant for action-guidance, and is relative, in some sense, to the beliefs that the decision-maker ought to have.[3]

Which decision situations fall within the scope of our claims? In the first instance, we argue that the following is one such case:[4]

> **The cause-neutral philanthropist.** Shivani has $10,000. Her aim is to spend this money in whatever way would most improve the world, and she is open to considering any project as a means to doing this.

The bulk of the paper is devoted to defending the claim that this situation is within the scope of axiological strong longtermism; in the final two sections we generalise this to a wider range of decision situations.

The structure of the paper is as follows. In section 2 we outline a plausibility argument for axiological strong longtermism. In our view, the most important respect in which the plausibility argument falls short of a proof is that it does not show that, as a matter of empirical fact, attempting to influence the course of the very long-run future is at all tractable. Section 3 is devoted to defending the crucial tractability claim.

In sections 2 and 3, we will at times help ourselves to some popular but controversial axiological and decision-theoretic assumptions (specifically, total utilitarianism and expected utility theory). This, however, is mainly for elegance of exposition. Section 4 conducts the corresponding sensitivity analyses, and argues that plausible ways of deviating from these assumptions are unlikely to undermine the argument. Section 5 argues that, while for concreteness we have focussed on the case of the cause-neutral philanthropist, if axiological strong longtermism is true of that decision context then it is also likely to be true of a fairly wide variety of other decision contexts (where cause-neutrality is absent, and/or where the decision is not one of how to spend money).

---

[3] It is widely agreed that either it is useful to distinguish between objective and subjective senses of 'ought' (Ewing 1948, pp.118-22; Brandt 1959, pp.360-7; Russell 1966; Parfit 1984, p.25; Portmore 2011; Dorsey 2012, Olsen 2017, Gibbard 2005, Parfit 2011), or 'ought' is univocal and subjective (Prichard 1932, Ross 1939 p.139, Howard-Snyder 2005, Zimmerman 2006, Zimmerman 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 1912 pp.88-9; Moore 1903 pp.199-200, 229-230; Ross 1930, p.32; Thomson 1986, pp. 177-9; Graham 2010; Bykvist 2011); according to this latter view, there is no coherent question of deontic strong longtermism in the vicinity of the thesis we attempt to discuss. Similarly (but less discussed), one might be skeptical of the notion of ex ante axiology; again, our discussion presupposes that any such skepticism is misguided.

[4] Note that Shivani need not be a *private* philanthropist. She could equally be in charge of some governmental or intergovernmental pot of resources, provided that the remit of that pot is cause-neutral, i.e. the remit is simply to maximise the good, rather than (say) to optimise the health or transport system. Given our stipulation about the content of Shivani's aim, it is almost trivially the case that if axiological strong longtermism is true of Shivani's decision situation, then so also is deontic strong longtermism. We discuss cases in which the connection between axiological and deontic strong longtermism is less direct in section 5.

Thus far, our discussion will have been exclusively focussed on axiological strong longtermism. Section 6 turns to the question of deontic strong longtermism. There, we argue that according to any plausible non-consequentialist moral theory, our discussion of axiological strong longtermism also suffices to establish deontic strong longtermism. Section 7 summarises.

The argument in this paper has some precedent in the literature. Nick Bostrom (2003) has argued, on the basis of the vast number of people who would live in the future if civilisation settles the stars, that increasing the probability that such settlement occurs should be the top priority for total utilitarians. Nick Beckstead (2013) argues from a somewhat broader set of assumptions to a similar conclusion.[5] Our aim in this paper is to expand on this prior work in four ways. First, whereas earlier work has focussed primarily on the examples of extinction risk mitigation and (sometimes) promotion of space settlement, we discuss a range of other "longtermist" interventions, and we argue that strong longtermism is true even if one sets aside the possibility of those (population-increasing) interventions. Second, we show that the argument goes through on a wide range of axiologies and decision theories, not only on the combination of total utilitarianism and expected utility theory. Third, we argue that insofar as strong longtermism is true of a decision context that involves allocating resources across cause areas, it is likely also to be true of various other decision contexts, including ones that do not involve cross-cause comparisons and ones that do not involve allocating money. Fourth, in addition to axiological strong longtermism, we also discuss the deontic claim: we argue that deontic strong longtermism is true, given any of a wide variety of plausible non-consequentialist theories.

We believe that axiological and deontic strong longtermism are of the utmost importance. If society came to adopt these views, much of what we would prioritise in the world today would change.

2. A plausibility argument for axiological strong longtermism

This section offers a plausibility argument for axiological strong longtermism. For the reasons discussed in subsequent sections, this argument does not by itself prove axiological strong longtermism. However, in our view it does show that, until and unless some objection proves damning, we should be predisposed towards believing axiological strong longtermism.

We start from two assumptions. One is empirical, while the other is evaluative.

[5] Beckstead's "Main Thesis" is: "From a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years" (ibid., p.1).

2.1 In expectation, the future is vast in size

It should be uncontroversial that there is a vast number of expected beings in the future of human civilisation.

A typical mammalian species's lifespan is around 500,000 years (Ceballos et al 2015); since *homo sapiens* has so far existed for only around 300,000 years (Schlebusch et al 2017), that comparison would suggest that we still have 200,000 years to go. However, we are not a typical mammalian species: we have 100 times the biomass of any large wild land animal that has ever lived (Wilson 2002:29), across a staggering diversity of environments, and, moreover, we have the technological power to avoid what would be extinction-level events for other animals, including the power to detect and deflect asteroids (NASA 2007). Of course, human civilisation itself introduces its own risks, such as from nuclear war and man-made pathogens. But it seems hard, given our state of knowledge, to be very confident that we will destroy ourselves, and so we should think there is at least a significant chance that we have a very large future ahead of ourselves.

For the purposes of this article, we will generally make the quantitative assumption that there are, in expectation, at least 1 quadrillion (10^{15}) people to come — 100,000 times as many people in the future as are alive today. This we be true if, for example, we assign at least a 1% chance to civilisation continuing until the Earth is no longer habitable, using an estimate of 1 billion years' time for that event[6] and assuming the same per-century population as today, of approximately 10 billion people per century.

Despite the magnitude of this number, we believe that it is conservative in the following ways. First, because of future technology, Earth could potentially host far greater per-century populations than is possible today, just as technology to date has enabled far greater per-century populations than were possible in hunter-gatherer times. Second, and even more importantly, civilisation in the future may spread to the stars. Having even a small credence in space settlement drastically increases the expected size of the future: there are around 10^{22} stars in the affectable universe (Ord 2016) and over 100 billion stars in the Milky Way alone (Howell 2018); the last of these will die in quintillions of years' time (Adams & Laughlin, 1997; Adams & Laughlin, 1999).[7] Even if just 0.01% of solar systems within the Milky Way were settled with the current population per century of Earth for just one billion years, there would be 10^{24} future people: one hundred trillion people for every person alive today. One would need to have a credence of less than one in one billion in this possibility in order for the expected number of future people to be fewer than one quadrillion.

[6] According to Adams (2008), the end of complex life on Earth may come in 0.9–1.5 billion years.

[7] Though conventional star formation will cease in about 1 to 100 trillion years, there are many proto-stars (called brown dwarfs) that are too small to ignite on their own, and their collisions will create a small but steady stream of new stars that will keep going for at least a million times as long as conventional star formation. Our thanks to Toby Ord for pointing this to our attention.

2.2 All consequences matter equally

Our second assumption is that, for the purposes of moral decision-making and evaluation, *all* the consequences of one's actions matter, and (once we control for the degree to which the consequence in question is predictable at the time of action) they all matter *equally*.

In particular, this assumption rules out a positive rate of pure time preference. Such a positive rate would mean that we should intrinsically prefer a good thing to come at an earlier time rather than a later time. If we endorsed this idea, our argument would not get off the ground.

To see this, suppose that future well-being is discounted at a modest but significant positive rate – say, 1% per annum.[8] Consider a simplified model in which the future certainly contains some constant number of people throughout the whole of an infinitely long future, and assume for simplicity that lifetime well-being is simply the time-integral of momentary well-being. Suppose further that average momentary well-being (averaged, that is, across people at a time) is constant in time. Then, with a well-being discount rate of 1% per annum, the amount of discounted well-being even in the whole of the infinite future from 100 years onwards is only about one third of the amount of discounted well-being in the next 100 years. While this calculation concerns total well-being rather than differences one could make to well-being, similar considerations will apply to the latter.

For present purposes, we take the assumption of a zero rate of pure time preference to be fairly uncontroversial. We know of no moral philosophers and few theoretical economists who would defend a non-zero rate of pure time preference.[9]

We also assume that, at least in decision contexts like Shivani's, there is no morally relevant distinction between the "direct" vs. the "indirect" effects of one's actions. This superficially seems to go against some popular views in medical ethics, where it is often held that it would be inappropriate for (say) a doctor or health service to prioritise treatment of one patient over another in a medically identical situation, on the grounds that the former patient occupies a more useful role in society, so that restoring her to full health sooner would have greater indirect benefits for other people (Kamm 1993, Brock 2003). It seems clear, however, that insofar as there is any plausible case for ignoring indirect effects in the medical context, that case will crucially hinge on features that are quite specific to that decision context, and that have no analogs in Shivani's decision context. There is no analog in Shivani's case, for instance, of the concern that taking indirect effects into account would undermine the trust that is important to the doctor-patient relationship (Angell 1993, Pellegrino 1997), or the

[8] In the survey by Drupp et al (2015), this was the median rate of pure time preference amongst those experts other than those who favoured a value in the "Ramsey-Stern" range 0-0.1% p.a.

[9] See Greaves (2017) for a survey of discounting in public policy, including a survey of the arguments for and against a positive rate of pure time preference. 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, & Stern, 2008; Buchholz & Schumacher, 2010; Gollier, 2013. In a recent survey of academic experts on the topic of social discounting, 38% of respondents agreed with this "Ramsey-Stern view" (Drupp et al 2018, p.119).

concern that doing so in a publicly funded health service might express odious views on the part of society regarding the relative moral worth of citizens (Mogensen, manuscript).

2.3 The plausibility argument

It is of course usually the case that the far-future effects of one's actions are harder to predict than their near-future effects. And this is of course relevant to *ex ante* evaluations. However, putting together the assumption that the expected size of the future is vast and the assumption that all consequences matter equally, it becomes at least plausible that the amount of *ex ante* good we can generate by influencing the expected course of the very long-run future exceeds the amount of *ex ante* good we can generate via influencing the expected course of short-run events, *even after* taking into account the greater uncertainty of further-future effects.

We can flesh this line of thought out more precisely by temporarily making some additional assumptions that are (however) inessential to the core argument. Let us temporarily assume (i) expected utility theory, (ii) separability of value with respect to time, and (iii) a finite time horizon T (whose precise value will not matter). Then (by temporal separability) there is some natural way of assigning values to temporally localised states of affairs, such that the goodness of a whole history of the universe (from the big bang to the heat death) is given by adding up these instantaneous values across all time; and, for the purposes of evaluation under uncertainty, the relevant quantity is the expectation value of this quantity. When comparing two actions, let Δ_S ("short-term delta") and Δ_L ("long-term delta") be the average amount by which the two actions' expected value differs across the duration of the short-run future (i.e. the next 100 or 1000 years) and the very long-run future (i.e. the remainder of the future beyond that), respectively. Then, the expected value difference between the two actions is given by $t \cdot \Delta_S + (T - t) \cdot \Delta_L$.

The point now is that, since the duration of the very long-run future ($T - t$) is *vastly* greater than the duration t of the short-run future, it is quite plausible that the sign of the expected value difference will be driven by the sign of Δ_L. Suppose that we define the 'short run' as the next 100 years, and use our assumption that there is a 1% chance of civilisation lasting for 1 billion years, at constant population levels. Given these choices of parameters, in order for strong longtermism to be true, the expected value, per millenium, of the best long-run action would need to be greater than 1/100000th of the expected value of the best short-run action for the next millennium. That's a very low bar.

In the next section we will argue that, for many interesting pairs of actions that one could consider, Δ_L (while small) will be sufficiently large as to meet this bar. This means that, of the two actions under comparison, the action that is better overall will be the one that has the more beneficial effects on the very long-run future, as axiological strong longtermism would lead us to expect.

A complementary reason for suspecting that axiological strong longtermism is true concerns the behaviour of other actors. In general, there are diminishing marginal returns to the

expenditure of resources towards a given aim, because the lower-hanging fruit are usually picked first. Shivani's question is which of her options is optimal in a sense that is completely impartial between times. But Shivani is not alone in the universe, and the vast majority of other actors have different priorities: because of greater concern for themselves and those they know than for future strangers, because of institutional incentives like election cycles and quarterly earnings reports, and because of an intrinsic human tendency towards impatience, they exhibit a significant amounts of preference for near-term positive effects over long-term positive effects (Frederick, Loewenstein and O'Donoghue 2002). Shivani should therefore expect that most other actors have been selectively funding projects that deliver high short-run benefits, and leaving unfunded projects that are better by Shivani's lights, but whose most significant benefits occur over the course of the very long run. This means that Shivani should expect to find axiological strong longtermism true at the current margin — *provided* (which we have not yet argued) that there were any projects with significantly beneficial *ex ante* effects on the very long-run future to begin with.

To summarise our plausibility argument: because of the vast expected number of people in the future, it is quite plausible that for options that are appropriately chosen from a sufficiently large choice set, effects on the very long-run future dominate *ex ante* evaluations, even after taking into account the fact that further-future effects tend to be more uncertain. Therefore, the best options ex ante will tend to be the ones that have the most beneficial effects ex ante on the very long-run future. Further, because of the near-term bias exhibited by the majority of existing actors, we should expect these options to be systematically under-exploited at the current margin. This renders axiological strong longtermism plausible.

The bulk of the remainder of the paper is devoted to articulating and assessing various lines of resistance to the claim that the above considerations lead to the truth of axiological strong longtermism: empirical, axiological and decision-theoretic.

3. The intractability objection

One key component of our claim is that as a matter of empirical fact, in an interestingly large class of cases, the ratio $\frac{\Delta_L}{\Delta_S}$ is large enough to render axiological strong longtermism true. However, we certainly don't think this is obvious. One might well worry that it is essentially impossible to significantly influence the long-term future: perhaps, for example, the magnitude of the effects of one's actions (in value-difference terms) decays with time from the point of action, and sufficiently fast that in fact the short-term effects tend to dominate the

all-time integral V. Call this *the washing-out hypothesis*.[10] We ourselves regard this as the most serious objection to axiological strong longtermism.

We agree that the washing-out hypothesis is true of some decision contexts: in particular, for many relatively trivial decision contexts, such as a decision about whether or not to click one's fingers. However, we claim that it is also false of many decision situations, and in particular of Shivani's. If Shivani is specifically looking for options whose effects do not wash out, we claim she can find some.

It will not always be so, but in Shivani's case there is a fairly natural status quo situation: a state of affairs in which Shivani "does nothing" with her money (that is, she just keeps the money in her bank account indefinitely). Our argument will make use of comparisons to this status quo ("business as usual", or "BAU") situation, although our conclusion is not dependent on the choice of status quo.

To argue for axiological strong longtermism, we claim, it suffices to establish that there exists at least one example of an option available to Shivani with the property that its far-future expected value (relative to BAU) is significantly greater than the best available short-term expected value (again, relative to BAU).

This claim suffices to ground an argument for axiological strong longtermism. To see why it suffices, we reason as follows.

Let SR* be the option whose effects on the short run are best. There are two cases to consider: SR* is also contained in the small subset of long-run-best options, or it isn't.

In the first case, axiological strong longtermism follows almost trivially. It is fairly likely in this case that SR* itself is also overall best. If SR* isn't overall best, then the overall best option must be one that is better than SR* in the long-run future. But in that case, it too must be in the small subset of options that are overall long-run best. So axiological strong longtermism is true.

Consider now the second case. Note that according to the numbers in our examples to come, the highest attainable long-term improvement is not just a bit bigger, but bigger by an order of magnitude or more, than the highest attainable short-run improvement (relative to BAU). In that case, the option that is overall-best must have at least 9/10 of the long-run value (relative to BAU) that the long-run-best option has (otherwise there is no way for the short-run improvement to suffice to render the option in question best overall, despite its being far-future-suboptimal; we assume here, as seems plausible, that none of the options under consideration is significantly net *detrimental,* relative to BAU, in the short term). This

[10] It is important here to distinguish between *ex ante* and *ex post* versions of the washing-out claim. The ex post version is certainly false. The reason for this is familiar from the philosophical literature on cluelessness: even our most trivial actions (such as decisions about whether and when to cross the road) have (for instance) identity-affecting consequences, and therefore in *ex post* terms have effects into the very long-term future that are almost certain to outweigh the action's short-term effects (Moore 1903 §93, Smart 1973:33, Parfit 1984:351, Greaves 2016)**.** However, it is the *ex ante* version that is relevant to the arguments of this paper.

establishes that the overall-best option is contained in "a small subset of options whose ex ante effects on the very long-term future are best". (The "small subset" is: the set whose far-future improvement, relative to BAU, is at least 9/10 that of the best available far-future improvement.) That is, it establishes axiological strong longtermism.

Our remaining task, then, is to show that there does indeed exist at least one option available to Shivani with the property that its far-future expected value (over BAU) is significantly greater than the best available short-term expected value (again relative to BAU). That is the task of the remainder of this section. We will proceed by considering several possible examples, in three categories: speeding up progress, mitigating extinction risk, and steering towards a better rather than a worse "attractor state" in contexts that do not involve a threat of extinction.

For the remainder of this section, we will *temporarily* assume both total utilitarianism as a matter of axiology, and expected utility theory for the treatment of uncertainty.[11] We will further assume risk neutrality with respect to total well-being, so that the *ex ante* value of an option is the corresponding expected total welfare (rather than the expectation value of any nonlinear transformation of total welfare). This, however, is mainly for elegance of exposition. In section 4, we will argue that plausible ways of deviating from these assumptions are unlikely to undermine the argument.

3.1 Advancing progress

Conditions for human welfare have been getting progressively better over time: we today enjoy vastly better living conditions than those of the Stone Age or of medieval times. More concretely and recently, GDP per capita is growing exponentially (Bolt et al 2018); lifespans are increasing (Roser 2019a); and an increasing proportion of the world lives in a democracy (Roser 2019b).

This suggests that if there are actions we could take that would bring this march of progress forward in time, even by a small amount, this could have long-lasting beneficial effects. Suppose, for instance, we bring it about that the progress level that would otherwise have been realised in 2030 is instead realised in 2029 (say, by hastening the advent of some beneficial new technology), and that progress then continues from that point on just as it would have if the point in question had been reached one year later. Then, for as long as the progress curve retains a positive slope, people living at every future time will be a little bit better off than they would have been without the intervention. In principle, these small benefits at each of an enormous number of future times could add up to a very large aggregate benefit.

Just how much of an improvement this amounts to depends, however, on the shape of the progress curve. In a discrete-time model, the benefit of advancing progress by one time

[11] This set of assumption is consistent with, but stronger than, the set of assumptions we made in section 2.3.

period (assuming that at the end of history, one thereby gets one additional time period spent in the "end state") is equal to the duration of that period multiplied by the difference between the amounts of value that are contained in the first and last periods. Therefore, if value per unit time is set to plateau off at a relatively modest level, then the gains from advancing progress are correspondingly modest. Similarly, if value per unit time eventually rises to a level enormously higher than that of today, then the gains from advancing progress are correspondingly enormous.

Which of these scenarios describes our actual situation? If we make the conservative assumptions that the number of people per century remains fixed, and that their capacities for wellbeing are similar to ours, then advancing technological is unlikely to have very large long-run effects. We should expect economic growth, or at the very least the well-being that it leads to, to plateau at some point in the next few thousand years. In that case, we are in the first scenario, and the benefits of advancing progress are relatively modest.

If, however, the value of the future, per century, is much higher in the far future than it is today — whether because the population per century is much larger (due to space settlement or otherwise) or because some form of enhancement renders future people capable of much higher levels of well-being, or both — then the case for advancing progress is significantly stronger. For example, borrowing our earlier numbers on settlement of the Milky Way, if at the end of civilisation there are 10^{17} people living lives 10 times as good as today, then enabling one extra year of civilisation in that best state amounts to more than one hundred thousand times the amount of value in the whole lives of everyone alive today.[12]

3.2 Influencing the choice among attractor states

Here is an abstract structure which, *insofar as* it is instantiated in the real world — a question we will return to shortly — 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 an *attractor state* be a subset A of S with the property that, given the dynamics of the universe, if the instantaneous state of the world once enters A, then it tends to remain in A for an extremely long time. Now suppose that there are two or more such attractor states, differing significantly from one another in terms of average goodness. Suppose further that the world is not yet in any of the states in question, but is fairly likely to settle into one or the other of the states in question in the

[12] This last argument assumes, of course, that the relevant "last period" for the purposes of this argument is the last period *before the decline of civilisation or humanity* (etc.), rather than a state of the universe after that eventual decline. If, on the other hand, the eventual causes of human extinction (etc.) are endogenous, in such a way that advancing progress by one year equally hastens humanity's eventual decline by one year, then advancing progress by one year in 2029 simply amounts to sacrificing the value that would otherwise have been contained in the year 2029.

foreseeable future. Finally, suppose that there is something we can now do that changes the probability that the world ends up in a better vs. a worse one of these attractor states. Then — as a result of the persistence that is built into the definition of "attractor state" — 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 to Shivani that instantiate the structure just described. We claim that there are.

3.3 Mitigating risks of premature human extinction

Human extinction is an attractor state *par excellence*. To state the obvious: the chances of the human race re-evolving, if we go extinct, are miniscule. Similarly, non-extinction is (to a lesser extent) an attractor state: while there are indeed grave risks of an extinction event, there is at least a strong tendency for human existence to persist.

These attractor states have unequal expected value. In particular, according to a total utilitarian population axiology, given the expected number of future people and assuming that on average these people would have lives of positive welfare, premature human extinction tends to be astronomically bad. Correspondingly, even an extremely small reduction in extinction risk tends to have very high expected value. The point is well put by Nick Bostrom:

> "Even if we use the most conservative [estimate of how many descendants present humans could have if we don't go prematurely extinct], we find that the expected loss of an existential catastrophe is greater than the value of 10^{16} human lives. This implies that the expected value of reducing existential risk by a mere one millionth of one percentage point is at least a hundred times the value of a million human lives… One might consequently argue that even the tiniest reduction of existential risk has an expected value greater than that of the definite provision of any 'ordinary' good, such as the direct benefit of saving 1 billion lives." (Bostrom 2013:18)

As an empirical matter of fact, as is increasingly recognised, there are things we (or Shivani) could do that would reduce the chance of premature human extinction by a non-negligible amount. As a result, the cost-effectiveness of some such interventions compares very favourably, by total utilitarian lights, to that of the best ways of improving the short run.

For instance, Matheny (2007) calculates that with a budget of $20 billion we could, in expectation, save 8 billion life-years via further improvements to defence systems against the possibility of a major asteroid colliding with Earth, giving an expected cost of $2.50 per life-year saved. In addition, this figure based on the cost-effectiveness of asteroid defences serves only as a quite extreme lower bound on the cost-effectiveness of appropriately directed efforts to mitigate extinction risk: it is highly likely that there are yet more cost-effective opportunities to mitigate other extinction threats, in particular from emerging technologies

such as artificial intelligence and synthetic biology. In addition, Matheny's estimate uses extremely conservative assumptions about the future of the human race.[13]

For what it's worth, we ourselves are sympathetic to this view. The crucial claim that premature human extinction would be astronomically bad, however, is relatively fragile to variations among reasonably plausible evaluative views.

Three dimensions of variation are relevant. Firstly, "person-affecting" approaches to population ethics tend to regard premature extinction as being of modest badness, possibly as neutral, and even (if the view in question also incorporates "the asymmetry") possibly as a good thing (Thomas, manuscript). Secondly, one might have a "totalist" view but asymmetrically weight negative welfare more strongly than positive welfare unit-for-unit (Arrhenius and Bykvist 1995, chapter 3); this might render the expected value of continued human survival less clear, and therefore push against the conclusion that extinction would be astronomically bad. Thirdly, even on a total utilitarian view, there is some scope for reasonable disagreement about whether the average well-being level in the future is positive (Althaus and Gloor, 2018).

Because of this fragility, it is important to explore the prospects for improving the *ex ante* value of the long run future conditional on survival, alongside this case for mitigating extinction risk.

3.4 Influencing the choice among non-extinction attractor states

Extinction is the most vivid and obvious example of a value-relevant attractor state whose probability we can influence, but it is far from the only one. Here are some other examples.

First, climate change could have indefinite impacts on the future of civilisation. A sufficiently warmer climate could result in a slower long-run growth rate (Pindyck 2013, Stern 2006), making future civilisation poorer indefinitely; or it could mean that the planet cannot in the future sustain as large a human population (Aral 2014); or it could cause unrecoverable ecosystem losses, such as species extinction and destruction of coral reefs (IPCC 2014, pp.1052-54).[14,15] Shivani could act to mitigate climate change. For example, she could route her resources to the Coalition for Rainforest Nations or the Clean Air Task Force (Halstead 2018) in order to expedite political efforts to reduce carbon emissions, or she could fund research into clean energy, or she could use her resources to organise grass-roots support for climate change mitigation that would make strong action by governments align more closely with ordinary political incentives.

---

[13] Similarly, Millett and Snyder-Beattie (2017) calculate the cost of interventions that aim to reduce extinction risk from biotechnology to be in the range of $0.10-$100 per life-year saved.

[14] The possibility of abrupt and persistent climate change is discussed by Alley et al. (2003).

[15] Whether or not this is an example of an attractor state in the currently relevant sense is partly a question of whether the damage from climate change is a matter of the changed climate being *permanently less conducive to well-being*, or whether the primary value-relevant thing is that significant climate change would necessitate a costly *but relatively time-limited* period of adaptation. This seems unclear.

Second, it is plausible that within the next century or two there might be developed strong international governance organisations, or even a world government. This could occur via a variety of means: it could be developed gradually out of existing international organisations like the UN; it could be put in place in the aftermath of a third world war, just as organisations like the UN were put in place after WWII; or it could come about because one country becomes dominant, globally, through military conquest, or through greater economic power (Rodrik 2000; Cabrera 2010; Cabrera 2012; Lu 2016; Yacoub 2018).

But once such institutions were created, they might persist indefinitely. Political institutions often change as a result of conflict or competition with other states. For strong world governments, this consideration would not apply (Caplan 2008). In the past, governments have also often changed as a result of civil war or internal revolution. However, advancing technology might make that far less likely for a future world government: modern and future surveillance technologies could prevent insurrection, and AI-controlled police and armies could be controlled by the leaders of the government, thereby removing the possibility of a military coup (Caplan 2008; Smith 2014).[16]

Further, there is more than one way an international governance organisation or a world government might be constituted, and some of the possible ways are significantly more conducive to well-being than others. (For example, constitutions that are strongly influenced by the parochial interests of a few nations who were powerful at the time of formation are likely to significantly underweight the interests of those in the disadvantaged countries, and so be significantly less conducive to overall global well-being. Criticisms of this form are often directed against the World Trade Organisation (Pogge 2008).) And there are things that Shivani could do to non-negligibly influence the probability that we end up with a more rather than a less well-being conducive set of institutions. For instance, she could fund the development of better methods of political decision-making such as via the Good Judgment Project[17] or the Center for Election Science,[18] or she could play a waiting game: establishing a fund with the express remit of exerting political pressure in beneficial directions over the process of world institution formation, whenever the day should come when such a fund could be put to good use.

Third, it is also plausible that within the next century we might develop advanced artificial general intelligence, with a higher intelligence level than that of humans (Bostrom 2014 chs. 1-2; Müller and Bostrom 2016; Grace et al 2018). Unlike humans, such artificial agents would not be mortal: even though any given piece of hardware would wear out, the underlying code that determined the agent's goals would be copyable (and the agent in question would have every incentive to see to it that it was copied), and could therefore

[16] For reasons like these, axiological strong longtermism is is less likely to be true of various past decision contexts (for example, the decision context of a cause-neutral philanthropist living 100 or 1000 years ago) than it is to be true of present decision contexts. We set aside the question of whether axiological longtermist is true of past decision contexts; our claim here is that axiological strong longtermism is (at least) highly plausible for present decision contexts.

[17] https://goodjudgment.com/

[18] https://www.electionscience.org/

persist indefinitely (Hanson 2016, pp.57-8). Furthermore, the current dominion of *Homo Sapiens* over the Earth is plausibly due to our being the most intelligent species; if an artificial agent or agents more intelligent than humans were created, then, in general we should expect them to exert very significant control over human affairs, and more generally over the conditions needed for the flourishing of sentient life, for as long as they do persist (Chalmers 2010; Bostrom 2014, pp.vii). What direction this influence pulls will depend on the goals that are embodied in the AI system. Those goals will probably be extremely difficult or impossible for humans to change once the AI system has started acting in the world, but there are things that Shivani can do now, prior to the creation of any such system, to improve the probability that such an AI system embodies goals that are more rather than less conducive to well-being. She could, for example, fund technical and policy work on ensuring the safe development of artificial intelligence at OpenAI or the Center for Security and Emerging Technology.[19]

Like any discussion of unprecedented future possibilities, this discussion is of course very speculative (and our extremely brief overview cannot hope to do justice to the complexity of the relevant issues). As such, it can tend to evoke a reaction that "this is just science fiction". However, given rapid recent progress both in hardware and in the performance of narrow AI systems, it would be extremely overconfident to believe that there is no chance of such advances occurring over the course of the next century; experts in the field in fact assign it quite a substantial probability.[20] There is also a wide consensus among diverse leading thinkers (both within and outside the AI Research community) to the effect that the risks we have just hinted at are indeed very serious ones, and that much more should be done to mitigate them.[21]

There are therefore a number of events that might occur over the next century that might have indefinitely long-lasting effects, and that are amenable to significant present influence through judicious choice of how to deploy resources. We have attempted a representative selection, while scratching only the surface of the possibilities. But we also have some ability to shape how these events progress, if they do occur, and different versions of these events would be at least somewhat predictable in the value of their effect on future civilisation. If we create sufficiently powerful agents of indefinite lifespan, then the aims of those agents will determine the course of the future. If we create a world government, then the values

[19] OpenAI (https://openai.com/) is directly attempting to build safe and beneficial artificial general intelligence. The Center for Security and Emerging Technology (https://cset.georgetown.edu/) provides policy advice to the US government on issues relating to the development of AI.

[20] In an expert survey of AI researchers, Grace et al (2018) found, for example, that the median (resp. mean) estimate for the number of years until full automation of labour would be achieved with 50% probability was 100 years (resp. 122 years). (It is unclear precisely how seriously to take these survey responses, however, as they exhibit apparent inconsistency across answers to closely related questions. Some survey participants were asked when AI would outperform humans in all tasks with 50% probability; to that question, the mean response was 45 years.)

[21] See for example the open letter on research priorities for robust and beneficial artificial intelligence (https://futureoflife.org/ai-open-letter) which was signed by several leading thinkers.

embodied in the constitution of that government will constrain future decision-makers indefinitely. If we let climate change continue unabated, we potentially lose goods that we can't get back. And there are concrete options for things we can do that have an influence on these possible long-lasting changes.

Now, the argument we are making is ultimately a quantitative one: that the expected impact one can have on the long-run future is greater than the expected impact one can have on the short run. It's not true, in general, that options that involve low probabilities of high stakes systematically lead to greater expected values than options that involve high probabilities of modest payoffs: everything depends on the numbers. (For instance, not all insurance contracts are worth buying.) So merely pointing out that one *might* be able to influence the long run, or that one can do so to a nonzero extent (in expectation), isn't enough for our argument. But, we will claim, any reasonable set of credences would allow that for at least one of these pathways, the expected impact is greater for the long-run.

Suppose, for instance, Shivani thinks there's a 1% probability of a transition to a world government in the next century, and that $1 billion of well-targeted grants — aimed (say) at decreasing the chance of great power war, and improving the state of knowledge on optimal institutional design — would increase the well-being in an average future life, under the world government, by 0.1%, with a 0.1% chance of that effect lasting until the end of civilisation, and that the impact of grants in this area is approximately linear with respect to the amount of spending. Then, using our figure of one quadrillion lives to come, the expected good done by Shivani contributing $10,000 to this goal would, by the lights of a utilitarian axiology, be 100 lives. In contrast, funding for Against Malaria Foundation, often regarded as the most cost-effective intervention in the area of short-term global health improvements, on average saves one life per $3500.[22]

Alternatively, consider artificial intelligence. Suppose that $1bn of well-targeted grants could reduce the probability of existential catastrophe from artificial intelligence by 0.001%. Again, for simplicity, assume that the impact of grants is approximately linear in amount spent. Then the expected good done by Shivani contributing $10,000 to AI safety would be equivalent, by the lights of our utilitarian axiology, and on the assumption of one quadrillion lives to come, to one hundred thousand lives saved.

Of course, in either case one could debate these numbers. But, to repeat, all we need is that there be *one* course of action such that one ought to have a non-minuscule credence in that action's having non-negligible long-lasting influence. Given the multitude of plausible ways by which one could have such influence, diverse points of view are likely to agree on this claim.

[22] This figure is taken from GiveWell's cost-effectiveness model of March 21 2019, using their median estimate of cost per death averted (after accounting for leverage and funging). The model is accessible here: https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models

3.5 A meta-option: Funding research into longtermist intervention prospects

Our list of examples of plausibly cost-effective first-order longtermist interventions is clearly quite speculative, and the evaluation of such options is currently extremely under-researched. In sections 3.3 and 3.4, we argued (for the cases of extinction risk mitigation and other "attractor state" structures respectively) that even despite this, the overall expected cost-effectiveness of the best longtermist interventions very significantly exceeds that of the best available ways of improving the short run, so that axiological strong longtermism is true.

Here we offer a complementary argument. To that end, let us suppose instead, for the sake of argument, that some reasonable credences do *not* assign higher expected cost-effectiveness to *any particular one* of the proposed longtermist interventions than they do to the best short-termist interventions, because of the thinness of the case in support of each such intervention. Suppose that the way in which those credences agree with the "existence claim" mentioned at the end of the preceding section is rather: they agree it is highly likely that *given sufficient additional information*, at least one of the proposed longtermist interventions (or another such intervention in a similar spirit) would come to have significantly higher expected value, relative to the updated credences, than the best short-termist options. However, before gathering that extra information, we cannot tell *which* longtermist intervention has that property.

It does not follow that the credences in question would recommend funding short-termist interventions. That is because Shivani also has what we might call a "second-order" longtermist option: *funding research into* the cost-effectiveness of various possible attempts to influence the very long run, such as those discussed above. Provided that subsequent philanthropists would take due note of the results of such research, this second-order option could easily have higher expected value (relative to Shivani's current probabilities) than the best short-termist option, since it could dramatically increase the expected effectiveness of future philanthropy (again, relative to Shivani's current probabilities).

Finally, here is another option that is somewhat similar in spirit: rather than spending now, Shivani could save her money for a later time. That is, she could set up a foundation or a donor-advised fund, with a constitutionally written longtermist mission. This fund would pay out whenever there comes a time when there is some action one could take that will, in expectation, sufficiently affect the value of the very long-run future.

These two considerations show that the bar for empirical objections to our argument to meet is very high. Not only would it need to be the case that, out of all the (millions) of actions available to an actor like Shivani, for none of them should one have non-negligible credence that one can positively affect the expected value of the long-run future by any non-negligible amount. It would also need to be the case that one should be virtually certain that there will be no such actions in the future, and that there is almost no hope of discovering any such actions through further research. This constellation of conditions seems highly unlikely.

4. Axiological and decision-theoretic objections

Our discussion above was conducted on the assumption of (i) a total utilitarian axiology and (ii) an expected-value approach to *ex ante* evaluation under uncertainty. Both of these assumptions are at least somewhat controversial. The present section examines the extent to which our arguments would be undermined by various ways of deviating from those assumptions. Broadly, the upshot will be that the case for strong longtermism is quite robust to plausible deviations from these starting axiological and decision-theoretic assumptions.

4.1 Population axiology

One of the categories of longtermist intervention we discussed in section 3 was that of mitigating extinction risk. As we noted there, the argument we gave for the claim that mitigating extinction risk has higher very long-run expected value (relative to BAU) than the best available short-run improvement (again over BAU) relied essentially on a controversial property of total utilitarianism. Many axiologies will not agree that premature extinction is *astronomically* bad, as is perhaps required for that argument to go through.

In particular, "person-affecting" approaches to population ethics tend to resist that claim.[23] According to the spirit of a person-affecting approach, perhaps, premature extinction is in itself at worst neutral: if humanity goes prematurely extinct, then there does not in fact exist any person who is worse off as a result of that extinction, and (according to the person-affecting ethos) this suffices to establish that the resulting state of affairs is not worse. Extinction risk mitigation may therefore beat the best short-termist options *only* conditional on a totalist population axiology.[24]

However, the other options for long-run influence we discussed (in section 3.4) are attempts to improve average future well-being, conditional on humanity *not* going prematurely extinct. While the precise numbers that are relevant will depend on the precise choice of axiology (and we will not explicitly crunch suggested numbers for any other axiologies), any plausible

[23] Anecdotally and imprecisely, person-affecting approaches seem to us to be the most popular family of alternatives to total utilitarianism, although it is difficult to formulate a complete person-affecting theory without falling into absurdity (Greaves 2017b). For an excellent exploration of how a person-affecting approach should handle issues of extinction risk see Thomas (manuscript). Similar remarks to those in the main text apply to the other possibilities we flagged in section 3.2, viz. the asymmetric weighting of negative over positive welfare, and the view that the zero level of well-being is sufficiently high relative to average future well-being to make premature extinction (at least) not astronomically bad.

[24] 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 (manuscript) formulates a number of possibilities.

axiology must agree that this is a valuable goal.[25] Therefore, the bulk of our argument is robust to plausible variations in population axiology.

4.2 Risk aversion with respect to welfare

One obvious point of contrast between the paradigm examples of interventions that have high short-term expected value (δ_{ST}) and those that (arguably) have high far-future expected value (δ_{LT}) is that the former tend to involve high probabilities of relatively modest welfare increases, whereas the latter tend to involve small probabilities of enormous welfare increases. One might well suspect, then, that risk aversion with respect to welfare would favour short-term over far-future welfare improvements.

There are (in our view strong) objections to the view that value is a risk-averse function of welfare. But here we set these aside. The more important points for present purposes are:

> (1) In this context it is crucial to distinguish between two senses of "risk aversion with respect to welfare", only one of which has any chance of favouring short-term over long-term welfare improvements.
>
> (2) Even on that sense of "risk aversion with respect to welfare", in order to undermine strong longtermism one would need fairly extreme risk aversion. Even if a little risk aversion of this type is reasonable, it is arguably unreasonable to be as risk averse as would be required.

We elaborate on these two points in turn.

First, we must distinguish between two senses of "risk aversion with respect to welfare".[26] The standard sense is risk aversion with respect to total welfare itself (that is, vNM value is a concave function of total welfare, w). But risk aversion in that sense tends to *increase* the importance of avoiding much lower welfare situations (such as near-future extinction), relative to the importance of increasing welfare from an already much higher baseline (as in the case of distributing bed nets in a world in which extinction is very far in the future).

A quite different case is risk aversion with respect to the *difference made by one's intervention*. (vNM value is a concave function of Δw.) This kind of risk aversion does indeed tend to favour actions with high-likelihood short-run benefits over attempts to improve far-future welfare. However, we should note that this is an unusual sense of "risk

[25] So-called "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, for this very reason, such approaches are implausible: this is far too much incomparability. For this reason, most person-affecting theorists themselves prefer a "wide" approach. 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 (Bader MS, Meacham 2012, Temkin 2012, Ross 2015). This type of theory will tend to agree that generally raising the well-being of future people is valuable, even if it is done in a way that does not preserve the identities of future persons.

[26] Here we discuss (two) types of "risk aversion" that are compatible with expected utility theory. Deviations from expected utility theory, including risk-weighted expected utility theory, are discussed in section 4.4.

aversion". Further, even if risk aversion with respect to total welfare is acceptable, it seems inappropriate for an altruistic agent to be risk averse in this second sense (Snowden 2015).[27]

Second, as the numbers in our discussion in section 3 suggested, if axiological strong longtermism is true at all, then it is likely to be true *by a large margin*. That is, it seems likely that (if strong longtermism is true at all) the intervention that is best by longtermist lights is not merely a bit better, but at least an order of magnitude better, than the option that is best by purely short-termist lights (all compared to BAU). If so, while risk aversion (with respect to the difference one makes oneself) will bring the option with the highest overall expected welfare improvement and the option with the highest short-term welfare improvement *closer together* in terms of expected value, only quite an extreme degree of risk aversion would actually *reverse the ranking* of these alternatives.

4.3 Non-aggregationism

One common objection to utilitarianism — here understood as the thesis that value is a linear function of total welfare *together with* a maximising consequentialist account of normative status — is that that view inappropriately favours interventions that deliver tiny benefits to huge numbers of people over interventions that deliver a very large benefit to a small number of people (perhaps to a single person). Arguably, for instance, it is inappropriate to favour giving a lollipop to n people over saving one person's life, regardless of how big n is.

To capture this intuition, many ethical theorists are sympathetic to a non-aggregationist view, according to which, when large benefits or harms to some are at stake, sufficiently trivial benefits or harms to others count for *nothing at all* from a moral point of view. (Scanlon 1998:235**,** Frick 2015**,** Voorhoeve 2014.) In the above example, such a view would indeed tend to hold that one ought to save a life rather than deliver even an arbitrarily large number of lollipop licks.

"Harm" and "benefit" here can be understood in either an *ex post* or an *ex ante* sense, leading to *ex post* and *ex ante* versions of non-aggregationism. This distinction makes no difference in the above example (the lollipop lick is a "trivial" benefit both in the *ex ante* and in the *ex post* sense). However, in other cases, *ex post* and *ex ante* non-aggregationism importantly come apart. For instance, in a choice between saving Alice's life for certain and holding a million-ticket lottery to decide which of one million other lives to save, *ex post* non-aggregationism would find nothing to choose between the two alternatives. But *ex ante* non-aggregationism would favour saving Alice, since the *ex ante* expected benefit to each of the one million other people is a millionfold smaller.

At first sight, it seems that *ex ante* non-aggregationism might undermine the argument for axiological strong longtermism. This is again because the ways in which typical longtermist interventions deliver high expected value is via a *very small probability of* a significant

[27] The issues in this subsection are investigated at greater length in Greaves, MacAskill, and Mogensen (manuscript).

benefit to each of an enormously large number of (possible) future persons.[28] Insofar as some good short-termist-motivated interventions instead involve *ex ante* much larger benefits to at least some of their beneficiaries, *ex ante* non-aggregationism will therefore tend to favour these over the longtermist interventions.[29]

However, this will amount to an argument against axiological strong longtermism only if non-aggregationism is itself an axiological view. But in fact, there are serious obstacles to interpreting non-aggregationism in axiological terms: non-aggregationist views generate cycles, and there is a widespread consensus (*pace* Temkin (2012)) that *axiology* cannot exhibit cycles (Broome 2004 pp.50-63; Voorhoeve 2013). Those who are sympathetic to non-aggregationism therefore tend themselves to interpret the view in purely deontological, rather than in axiological, terms (Voorhoeve 2014). (Many non-consequentialist theories do in any case posit that the "ought to choose rather than" relation is cyclic; see for example (Kamm 1996, chapter 12, esp. pp.339-44).) In the above example, on this view, *one ought to* simply save Alice rather than choose by lottery which of the other one million people to save, but that is not because the outcome of doing so (or anything else about the former action) is non-morally *better*; in fact the two available outcomes are equally good.

We conclude that the most plausible non-aggregationist views do nothing to undermine axiological strong longtermism. (We return to their implications for deontic strong longtermism in section 5.)

4.4. Giving extra weight to benefits to the badly off

A separate line of objection to utilitarianism is that it treats any given increase as being equally valuable, no matter how well off or poorly off future people are. As we noted earlier (section 3.1), generally speaking well-being has improved substantially over time. Continuing these trends into the future, we should perhaps expect people in the future to be far better off than they are today. Given a sufficiently large weighting of benefits to the worse-off compared to benefits to the better-off, we should therefore believe that the best actions

[28] This is true even if the identities of future persons are unaffected by the intervention in question. A further complication is that in general, such interventions (like all actions) lead to large-scale changes in the identities of far-future people. There are various, relevantly differing, ways a non-aggregationist might choose to treat non-identity cases, but if anything this complication will increase the extent to which non-aggregationism undermines long-termism.

[29] We note in passing that some short-termist interventions share the feature that they deliver a very small ex ante benefit to a very large number of people. Bed net distribution might be one such intervention, since *for any given potential bednet recipient*, the chance that an increase in bed net distribution ends up saving that person's life is extremely small — amounting to only one life saved from many hundreds of bednets distributed. However, not all short-termist interventions have this feature. For example, a program of direct cash transfers (to sufficiently pre-identified beneficiaries) will not have the feature in question. *Ex ante* non-aggregationism would therefore lead to a substantial shift in prioritisation among short-termist interventions, in addition to favouring the then-best short-termist interventions over longtermist ones.

available to us are those with the best effects on the worst-off people. These, the objection continues, are people in extreme poverty alive today.

There are two problems with this objection. First, it is not clear that future people will be better-off than those in extreme poverty today. There are at least serious *possibilities* that future people will be even worse off — for example, because of the adverse influence of climate change, misaligned artificial general superintelligence, or domination by a repressive global political regime. In addition, many of the contenders for "longtermist interventions" that we discussed above are precisely aimed at improving the plight of these very badly off possible future people, or reducing the chance that they have terrible as opposed to flourishing lives.[30]

The second problem with the objection is that given the large margin by which (we have suggested) longtermist interventions deliver larger improvements to aggregate welfare than similarly costly shorttermist interventions, only quite an extreme priority weighting seems likely to lead to the result in question. Even if some degree of prioritarianism is plausible, the degree required might be too extreme to be plausible by any reasonable lights.

4.5. Decision-theoretic objections

Above, we assumed that the correct way to evaluate options in *ex ante* axiological terms, under conditions of uncertainty, is in terms of expected value. This is the orthodox account of rational decision-making under conditions of uncertainty. However, there are rival accounts. We must therefore consider whether any plausible alternative account tends to undermine the argument for axiological strong longtermism.

We briefly flag one rival account that we (with many others) consider implausible. This is the account according to which at least under conditions of 'Knightian uncertainty' — that is, when there is little objective guidance as to which probability distributions over possible outcomes are appropriate vs inappropriate — the best option *ex ante* is the one whose worst outcome is least bad. We are convinced by the usual objections to this "maximin" account. However, for present purposes, the more important point is that maximin in any case *supports*, rather than undermining, axiological strong longtermism. The reason is that the worst outcomes, from any option, are ones in which the vast majority of the long-run future is of highly negative value (or, at best, have zero or very little positive value). Therefore, according to maximin, the only consideration that is relevant to ex ante axiological option evaluation is the avoidance of these long-term catastrophic outcomes.

A second rival approach is risk-weighted expected utility theory. The kind of "risk aversion" that this theory permits interacts with strong longtermism in precisely the same ways that the

[30] A somewhat related argument is made by Fleurbaey and Zuber (2015) against the claim that expected growth of consumption justifies a high discount rate on future goods.

type of risk aversion discussed above (section 4.2) does; we will not repeat that discussion here.

Finally, it might seem at first sight that ambiguity aversion would undermine the case for strong longtermism. In contemplating options like those discussed in section 3, the first-order task is to assess what are the rational credences that some given intervention to (say) reduce extinction risk, or reduce the chance of major global conflict, or increase the safety of artificial intelligence, and so on, would lead to a large positive payoff in the long run.[31] The thing that is most striking about this task is that it is *hard*. There is very little data to guide credences; one has an uncomfortable feeling of picking numbers, for the purposes of guiding important decisions, somewhat arbitrarily. That is, such interventions generate significant ambiguity. However, on reflection, attempts to optimise the short run also generate significant ambiguity, since it is very unclear what might be the long run consequences of (say) bed net distribution (Greaves, 2016). In addition, we again face the issue of whether one should be ambiguity averse with respect to the state of the world, or instead with respect to the difference one makes oneself to that state. We explore these issues in a related paper (Greaves, MacAskill and Mogensen, manuscript).

5. The scope of strong longtermism

So far, we have discussed the decision context of a cause-neutral philanthropist. Two features of this decision context are potentially particularly relevant. Because of cause-neutrality, the assessment of strong longtermism so far has centred on cross-cause comparisons (for instance, whether Shivani could do more good, in expectation, by focussing on AI safety or instead on malaria prevention). In addition, our comparisons so far have been of different ways to spend money, rather than of other kinds of actions.

However, neither of these features seems likely to be essential to the arguments. Insofar as axiological strong longtermism is true of Shivani's decision context, it seems likely to be true also of a fairly wide variety of other decision contexts.

To see that cause-neutrality is inessential to strong longtermism, suppose that Sophie is a philanthropic grantmaker comparing two deworming programs.[32] The two programs, A and B, would operate in two different countries (respectively, α and β). Suppose that one can deworm a greater number of children per dollar spent in country α than in country β (say, because of greater population density). Thus, the *short-term* cost-effectiveness of program A is higher than that of program B: per dollar spent, more additional child-years would be spent in school, and so on. However, the benefits of deworming are not only of this short-term character: there are also effects on later life incomes, and thereby (presumably) on

[31] More precisely: what is the rational credence distribution over the spectrum of possible sizes of the payoff of such an intervention.

[32] Deworming is often regarded as one of the most cost-effective types of intervention in the field of global health (see e.g. (GiveWell, 2018)).

later-generation incomes. For many of these knock-on effects, the magnitude of the effects depends on country-specific features, such as the background rate of economic growth in the country in question. Suppose then that countries α and β have different rates of economic growth, so that the benefits of deworming a child in country β compound over time more than do the benefits of deworming a child in country α. Then it could easily be the case that attending to the longer-term benefits of deworming reverses the comparative cost-effectiveness estimate of A vs. B that one would reach based on consideration of short-term benefits alone. Granted, this particular example involves timescales of decades rather than millenia. However, this is mainly for simplicity; it seems likely that similar points also apply on the longer timescale (but with the details of the expected longer-term benefits becoming substantially more complicated on that longer timescale).

To see that philanthropy is inessential to strong longtermism, suppose that Adam is a young graduate choosing his career path. He can choose to train either as a development economist, or as an AI safety researcher. While there are differences between Adam's decision context and Shivani's, there are also important similarities. In particular, the considerations that might make it better (in expectation) for Shivani to fund AI safety rather than developing world poverty reduction similarly might make it better (in expectation) for Adam to train as an AI safety researcher rather than as a development economist.

6. Deontic strong longtermism

Let us return now to the decision context of a philanthropist who is 'cause-neutral' at least in the limited sense that the range of options available to her is not restricted as to cause area (whether or not *her aim* includes cause-neutrality).

In section 2.1, we distinguished between axiological strong longtermism and deontic strong longtermism. Recall:

> **Axiological strong longtermism (AL)**: In a wide class of decision situations, the option that is *ex ante* best is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.
>
> **Deontic strong longtermism (DL**): In a wide class of decision situations, the option one ought, ex ante, to choose is contained in a fairly small subset of option whose *ex ante* effects on the very long-run future are best.

So far, our discussion has focussed exclusively on the case for axiological strong longtermism. This suffices for the analysis of Shivani's decision context, since (by stipulation) *her aim* was simply to maximise the good. Given that aim, instrumental rationality requires Shivani to select the most long-term beneficial option, if axiological strong longtermism is true.

However, in other decision contexts, it could happen that (by the lights of a non-consequentialist moral theory) deontic strong longtermism is false, even if axiological strong longtermism is true. Most relevantly, it seems this could happen if there is an agent-relevant prerogative such that there is no moral obligation to choose a "longtermist option", on the grounds that a permissible personal point of view places significantly higher value on some short-termist option.[33] Is this seeming veridical?

An argument for deontic strong longtermism could be either indirect (going via axiological strong longtermism *en route* to a deontic longtermist conclusion), or direct. We will outline an argument of the indirect type. The investigation of whether or not there is any sound direct argument for deontic strong longtermism is beyond the scope of this paper.

The indirect argument we propose is *the stakes-sensitivity argument*:

> (P1) In decision context C, the options that are best for the very long-run future are *enormously* better overall, in purely axiological terms, than even the options that are very best for the short run. *(Large-margin axiological strong longtermism)*
>
> (P2) When the axiological stakes are very high, non-consequentialist constraints and prerogatives tend to be outweighed, so that what one ought to do is simply whichever option is best.
>
> (C) In C, one ought to prefer the options that are best for the long-run future over those that are best for the short run.

(P1) goes importantly beyond axiological strong longtermism as we have formulated the latter. However, depending on exactly which are the most plausible sets of numbers for our examples, (P1) is arguably supported by the same examples that we have used to support axiological strong longtermism (section 3). In the context of those examples, recall, we suggested that particular longtermist interventions plausibly generated X times as much good (relative to 'business as usual') as the best attainable short-term good.

(P2) deserves more discussion. Call non-consequentialist views that endorse P2 *stakes-sensitive non-consequentialism*. ((P2) is of course trivially true according to consequentialism.)

That the non-consequentialist view should be stakes-sensitive is very plausible, intuitively. The lack of stakes-sensitivity is a common objection to Kant's notorious view that that even if a friend's life depends on it, one should not tell a lie (Kant, 1996). 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.

[33] Similar considerations apply to the issue of constraints, but prerogatives are the more salient departure from consequentialism in the present context.

Further, in 'emergency situation' situations like wartime, axiological considerations outweigh non-consequentialist considerations (at least for those fighting a just war). Consider, for example, the intuitions that one would have with respect to how one should act if one lived in Britain during World War II. It's very intuitive that, in that situation, that one is morally obligated to make significant sacrifices for the greater good that would not normally be required, such as by living far more frugally, separating oneself from one's family, and taking significant risks to one's own life — and this because the axiological stakes are so high.

We foresee four lines of resistance to (P2).

First, one could reject the idea of 'the good' altogether (Thomson 2008). On this view, there is simply no such thing as axiology. It's clear that our argument would not be relevant to those who hold such views. But such views have other problems, such as how to explain the fact that, in cases where there is a huge amount at stake, such as during wartime, ordinary prerogatives get overridden. It seems likely to us that any such explanation will result in similar conclusions to those we have drawn, via similar arguments.

Second, one might accept that the stakes could be outweighed by axiological considerations, but claim that, for decision-makers alive today, the stakes aren't high enough. However, though we can't rule this position out, we find it implausible. We argued above that the stakes in question are extremely large: that, even under quite conservative assumptions, by donating half their income, a middle-class member of an affluent country could do as much good as saving millions of lives. It would be unintuitive (and suspiciously convenient) if stakes this large were unable to outweigh the personal prerogative to spend on causes that are specially favoured by one's personal point of view.

Third, one might hold that some prerogatives are absolute: they cannot be overridden, no matter what the consequences. Absolutist views tend not to be very plausible, and have few adherents. (In the case of constraints as opposed to prerogatives, for instance, few people share Kant's view that even when an innocent life depends on it, one should not tell a lie even to an intending murderer.) However, for our purposes, even if the non-consequentialist is absolutist with respect to some prerogatives, our argument will most likely still go through for most decision situations. This is because, for most decision-makers, the case for strong longtermism does not involve or at least does not rely on the existence of extraordinarily demanding options. Perhaps, no matter how great the stakes, one is never required to give up one's own life, or that of one's own child, and perhaps one is never required to reduce oneself from a Western standard of living to an allowance of $2 per day. But, for the vast majority of decision-makers, in the vast majority of decision-situations, these will not be the choices at hand. Instead, the choice will be whether to switch career paths, or live somewhat more frugally, or where to donate a specified amount of non-necessary income, in order to try to positively influence the long-run future. Even if one is sympathetic to absolutism about some sacrifices, it's very implausible to be absolutist about these comparatively minor sorts of sacrifices (MacAskill, Mogensen, and Ord 2018).

Finally, and most plausibly, one might hold that only some sorts of axiological considerations are relevant to determining what we ought to do, and that once we make this distinction the core of our argument falters. We'll discuss two ways in which one could use this idea to reject our argument.

First, one might take a non-aggregationist view, and think that comparatively small benefits are not relevant to determining what one ought to do. This is the line of thought that we discussed in section 4.3 above, reappearing here in its proper place.

Second, one might think that axiological considerations cannot outweigh non-consequentialist considerations when (as here; cf. section 4.1) the axiological considerations involve altering the identities of who comes into existence.

However, both lines of response have significant problems, as they would prove too much.

Let's first consider the non-aggregationist response. Consider the example of someone alive in Britain during WWII, and considering whether or not to fight; or consider someone debating whether to vote in their country's general election; or someone who is deciding whether to join an important political protest; or someone who is reducing 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 person is question is obligated to undertake the relevant action. [34]

Second, consider the non-identity response. Climate change will again serve as a useful example. It's clear that governments (at least) ought to take significant action to fight climate change. But any policy designed to mitigate climate change will affect the identities of those who are to come. Therefore, the non-identity response would require rejecting the idea that the government is ever under an obligation to take significant action to fight climate change. That is clearly wrong.

In general, it seems to us that the 'common-sense' view on these matters is that we should care about the long-term future to a significant degree, but not to an overwhelming degree. Both responses to our argument avoid the latter implication, but only at the cost of telling us that we are essentially *never* obligated to ensure that the long-run future goes well. We find this implication to be a strong reason to reject the responses. We will not here take up the question of precisely how the views in question might avoid these implausible costs; we simply note that it is an adequacy constraint that they do so somehow, and that whatever moves they make in order to do so, the same moves are likely to render P2 true for the purposes of our argument.

[34] In addition, the small benefits response would create a stark division between what we as a society ought to do, and what we as individuals ought to do. Though any individual action might have a very small impact on the long-run future, as a society we could collectively have a significant influence. Those who endorse the small benefits response would therefore have to claim that, though as a society we ought to focus on improving the long-run future, this is not true for any individual constituent of that society. For some discussion of the general phenomenon of such "each-we dilemmas", see (Parfit 1984, pp.91-2). Temkin (2012, pp.85-95) discusses such dilemmas specifically in the context of non-aggregationist moral views.

The deontic longtermist claim is indeed surprising. However, we submit that this is because of surprising empirical facts (namely the sheer size of the future and the fact that we can, in expectation, take actions to significantly improve it), rather than some problem with the underlying normative motivation.

7. Summary and conclusions

Given the size of the future and the assumption that all consequences matter equally, it becomes at least plausible that the best options are generally best because of their effects on the course of the very long-run future, and not because of their more immediate effects. This paper has formulated and discussed a thesis — axiological strong longtermism — aimed at capturing that thought.

Axiological strong longtermism would be false in a world that had sufficiently weak causal connections between the near and the distant future, so that it was simply intractable to significantly influence the course of the very long-run future. However, we have argued, by adducing several examples, that the decision context we find ourselves in today (at least) does not have this feature.

We presented our central case in terms of (i) a total utilitarian axiology and (ii) an expected utility treatment of decision-making under uncertainty. However, we argued (in section 4) that plausible deviations from either or both of these theses do not undermine the core argument.

This paper mainly focussed on the decision context of a cause-neutral philanthropist. However, we also argued that insofar as strong longtermism (axiological or deontic) is true of that decision context, it is also plausible for decision contexts that do not involve cause-neutrality, and for decision contexts that involve how to spend (say) time rather than money.

In addition to axiological strong longtermism, we are also interested in the counterpart deontic question: roughly, that of whether and when *what one ought to do* is determined primarily by considerations of effects on the very long-term future. We argued that some such deontic longtermist claim will often be true, on the grounds that (1) axiological strong longtermism is true *by a large margin*, and (2) a plausible non-consequentialist theory has to be sensitive to the axiological stakes, becoming more consequentialist in output as the axiological stakes get higher.

References

Daron Acemoglu (2009). *Introduction to Modern Economic Growth*. Princeton: Princeton University Press.

Fred C. Adams (2008). Long-Term Astrophysical Processes. In Nick Bostrom and Milan Cirkovic (eds.) *Global Catastrophic Risks.* Oxford: Oxford University Press.

Fred C. Adams & G. Laughlin (1997). A dying universe: the long-term fate and evolution of astrophysical objects. Reviews of Modern Physics, 69(2), 337–72.

Fred C. Adams & G. Laughlin (1999). The Five Ages of the Universe: Inside the Physics of Eternity. Free Press.

R. B. Alley, J. Marotzke, W. D. Nordhaus, J. T. Overpeck, D. M. Peteet, R. A. Pielke Jr., R. T. Pierrehumbert, P. B. Rhines, T. F. Stocker, L. D. Talley, J. M. Wallace (2003). Abrupt Climate Change. *Science* **299**:2005–2010.

David Althaus and Lukas Gloor (2018). Reducing Risks of Astronomical Suffering: A Neglected Priority. *Foundational Research Institute*. Available at: https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/

Marcia Angell (1993). The doctor as double agent. *Journal of Kennedy Institute of Ethics* **3**:279–286.

Mustafa M. Aral (2014). Climate change and persistent high temperatures: does it matter? *Frontiers in Environmental Science* **2**(45).

Gustaf Arrhenius and Krister Bykvist (1995). Interpersonal Compensations and Moral Duties to Future Generations: Moral Aspects of Energy Use. *Uppsala Prints and Preprints in Philosophy* **21**, Uppsala Universitet.

Ralf Bader (manuscript) Neutrality and conditional goodness.

Nick Beckstead (2013). *On the Overwhelming Importance of Shaping the Far Future*. PhD thesis, Rutgers University.

Jutta Bolt, Robert Inklaar, Herman de Jong and Jan Luiten van Zanden (2018). Rebasing 'Maddison': new income comparisons and the shape of long-run economic development. *Maddison Project Working paper* **10**

Nick Bostrom (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas* **15**(3):308–314.

Nick Bostrom (2013). Existential Risk Prevention as Global Priority. *Global Policy* **4**(1):15–31.

Nick Bostrom (2014). *Superintelligence: Path, Dangers, Strategies*. Oxford: Oxford University Press.

Richard Brandt (1959). *Ethical Theory*. Englewood Cliffs, N.J.: Prentice-Hall.

John Broome (2004). *Weighing Lives*. OUP.

John Broome (2008). The ethics of climate change. *Scientific American* **298**:96-102.

Dan W. Brock (2003). Separate spheres and indirect benefits. *Cost Effectiveness and Resource Allocation* **1**(4).

Krister Bykvist (2011). How to Do Wrong Knowingly and Get Away with It. In *Neither/Nor. Philosophical Papers Dedicated to Erik Carlson on the Occasion of His Fiftieth Birthday*, 31–47. Uppsala Philosophical Studies 58. Uppsala: Department of Philosophy, Uppsala University

Luis Cabrera (2010). World government: Renewed debate, persistent challenges. *European Journal of International Relations* **16**(3) 511–530.

Luis Cabrera (ed.) (2012). *Global Governance, Global Government: Institutional Visions for an Evolving World System*. Albany: State University of New York Press.

Bryan Caplan (2008). The totalitarian threat. In Nick Bostrom and Milan Cirkovic (eds.) *Global Catastrophic Risks*. Oxford: Oxford University Press.

Gerardo Ceballos et al (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Science Advances* **1**(5).

David J. Chalmers (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* **17**(9–10):7–65

Diego Comin, William Easterly, and Erick Gong (2010). Was the Wealth of Nations Determined in 1000BC? *American Economic Journal: Macroeconomics* **2**:65–97.

Tyler Cowen (1992). Consequentialism implies a zero rate of intergenerational discount. In P. Laslett & J. S. Fishkin (eds.), *Justice between age groups and generations*. Yale University Press.

Partha Dasgupta (2008). Discounting climate change. *Journal of risk and uncertainty* **37**:141–169.

Dale Dorsey (2012). Objective Morality, Subjective Morality and the Explanatory Question. *Journal of Ethics and Social Philosophy* **6**(3):1–24.

Moritz A. Drupp, Mark Freeman, Ben Groom and Frikk Nesje. (2018). Discounting Disentangled. *American Economic Journal: Economic Policy* **10**:109-134. 10.1257/pol.20160240.

A.C. Ewing (1948). *The Definition of Good.* London: Routledge & Kegan Paul.

Marc Fleurbaey and Stephane Zuber (2013). Climate Policies Deserve a Negative Discount Rate. *Chicago Journal of International Law* **13**(2)

Shane Frederick, George Loewenstein, and Ted O'Donoghue (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature* **40**(2):351-401.

Johann Frick (2015). Contractualism and Social Risk. *Philosophy & Public Affairs* **43**(3):175-223.

Allan Gibbard (2005). Truth and Correct Belief. *Philosophical Issues* **15**:338–350.

GiveWell (2018). Combination Deworming (Mass Drug Administration Targeting Both Schistosomiasis and Soil-Transmitted Helminths). Available from <https://www.givewell.org/international/technical/programs/deworming>. Accessed 31 July 2019.

Katja Grace, John Salvatier, Allen Dafoe, Baobao Zhang, and Owain Evans (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research* **62**:729-754.

Peter A. Graham (2010). In Defense of Objectivism about Moral Obligation. *Ethics* **121**: 88-115.

Hilary Greaves (2016). Cluelessness. *Proceedings of the Aristotelian Society* **116**(3):311-339.

Hilary Greaves (2017a) Discounting for Public Policy: A Survey. *Economics and Philosophy* **33**(3):391–439.

Hilary Greaves (2017b). Population Axiology. *Philosophy Compass* **12**(1):

Hilary Greaves, William MacAskill, and Andreas Mogensen (manuscript). Existential risk and risk aversion.

John Halstead (2018). Climate Change. *Founders Pledge Cause Area Report*. Available at https://founderspledge.com/research/Cause%20Report%20-%20Climate%20Change.pdf

Robin Hanson (2016). *The Age of Em: Work, Love and Life when Robots Rule the Earth*. Oxford: Oxford University Press.

David Heyd (1988). Procreation and value: Can ethics deal with futurity problems? *Philosophia* **18**(2-3):151–170.

Frances Howard-Snyder (2005). It's the Thought that Counts. *Utilitas* **17**:265–281.

Elizabeth Howell (2018). How Many Stars Are in the Milky Way? *Space.com*. Available at https://www.space.com/25959-how-many-stars-are-in-the-milky-way.html

IPCC (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability*. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1132 pp.

Frances W. Kamm (1993). *Morality, Mortality Volume 1: Death and Whom To Save From It.* Oxford: Oxford University Press.

Frances W. Kamm (2001). *Morality, Mortality, Volume 2: Rights, Duties, and Status*. Oxford: Oxford University Press.

Immanuel Kant (1996). On a Supposed Right to Lie from Philanthropy. In Mary Gregor (transl./ed.) *Immanuel Kant: Practical Philosophy.* New York: Cambridge University Press.

Gregory Lewis (2016). Beware surprising and suspicious convergence. *The Effective Altruism Forum*, accessible at https://forum.effectivealtruism.org/posts/omoZDu8ScNbot6kXS/beware-surprising-and-suspicious-convergence

Catherine Lu (2016). World Government. *The Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/archives/win2016/entries/world-government/>.

William MacAskill, Andreas Mogensen, and Toby Ord (2018). Giving Isn't Demanding. In Paul Woodruff (ed.), *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*. Oxford: Oxford University Press, pp. 178–203.

William MacAskill (2019), 'Longtermism'. *The Effective Altruism Forum*, accessible at *https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism*

Elinor Mason (2013). Objectivism and Prospectivism About Rightness. *Journal of Ethics and Social Philosophy* **7**(2):1–21.

Jason G. Matheny (2007). Reducing the Risk of Human Extinction. *Risk Analysis* **27**(5):1335–1344.

Christopher Meacham (2012). Person-affecting views and saturating counterpart relations. *Philosophical Studies* **158(**2):257–287

Piers Millett and Andrew Snyder-Beattie (2017). Existential Risk and Cost-Effective Biosecurity. *Health Security* **15**(4):373–383.

Andreas Mogensen (manuscript). Meaning, medicine, and merit.

G.E. Moore (1903). *Principia Ethica.* Cambridge: Cambridge University Press.

G.E. Moore (1912). *Ethics*. London: Williams and Norgate.

Vincent C. Müller and Nick Bostrom (2016). Future progress in artificial intelligence: A survey of expert opinion. *Fundamental issues of artificial intelligence*. Springer, Cham. 555-572.

Thomas Nagel (1978). Ruthlessness and public life. In Stuart Hampshire (ed.), *Public and Private Morality*. Cambridge: Cambridge University Press.

NASA (2007). Near-Earth Object Survey and Deflection Analysis of Alternatives. https://www.nasa.gov/pdf/171331main_NEO_report_march07.pdf

Kristian Olsen (2011). A Defense of the Objective/Subjective Moral Ought Distinction. *The Journal of Ethics* **21**(4):351–373.

Toby Ord (2016). The Edges of Our Universe.

Derek Parfit (1984). *Reasons and Persons.* Oxford: Clarendon Press.

Derek Parfit (2011). *On What Matters*. Vol. 1. Ed. Samuel Scheffler. Oxford: Oxford University Press.

Edmund D. Pellegrino (1997). Managed care at the bedside: How do we look in the moral mirror? *Journal of Kennedy Institute of Ethics* **7**:321– 330.

Robert S. Pindyck (2013). Climate Change Policy: What Do the Models Tell Us? *NBER Working Paper* No. 19244.

Thomas W. Pogge (2008). *World Poverty and Human Rights*, second edition. Cambridge: Polity Press.

Douglas W. Portmore (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.

H. A. Prichard (1932) 2002. Duty and Ignorance of Fact. In *Moral Writings*, ed Jim MacAdam, 85–110. Oxford: Oxford University Press.

Frank P. Ramsey (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.)

Dani Rodrik (2000). How far will economic integration go? *Journal of Economic Perspectives* **14**(1): 177–186.

Max Roser (2019a). Life Expectancy. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/life-expectancy' [Online Resource]

Max Roser (2019b). Democracy. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/democracy' [Online Resource]

Jacob Ross (2015). Rethinking the Person-Affecting Principle. *Journal of Moral Philosophy* **12**(4):428–461.

W.D. Ross (1930). *The Right and the Good*. Oxford: Oxford University Press.

W. D. Ross (1939). *The Foundations of Ethics*. Oxford: Clarendon Press.

Bertrand Russell (1966). The Elements of Ethics. In *Philosophical Essays*, 13–59. New York: Simon and Schuster.

T. M. Scanlon (1998). *What We Owe to Each Other.* Cambridge, Mass.: Harvard University Press.

Carina M. Schlebusch et al (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**(6363):652–655.

J. J. C. Smart (1973). An outline of a system of utilitarian ethics. In J. J. C. Smart and Bernard Williams (eds.) *Utilitarianism: For and against*. Cambridge University Press.

Noah Smith (2014). Drones will cause an upheaval of society like we haven't seen in 700 years'. *Quartz*

James Snowden (2015). Does risk aversion give an agent with purely altruistic preferences a good reason to donate to multiple charities?

Nicholas Stern (2006). The Economics of Climate Change.

Larry S. Temkin (2012). *Rethinking The Good: Moral Ideals and the Nature of Practical Reasoning.* New York: Oxford University Press

Teruji Thomas (manuscript). The Asymmetry and the Long Term.

Judith Jarvis Thomson (1986). "Imposing Risks." In *Rights, Restitution, and Risks*, edited by William Parent, 173-91. Cambridge, MA: Harvard University Press.

Judith Jarvis Thomson (2008). *Normativity.* Chicago: Open Court.

Alex Voorhoeve (2013). Vaulting intuition: Temkin's critique of transitivity. *Economics & Philosophy* 29.3 (2013): 409-423

Alex Voorhoeve (2014). How Should We Aggregate Competing Claims? *Ethics* **125**: 64-87.

Edward O. Wilson (2002). *The Social Conquest of Earth.* New York: Liveright.

Amin Yacoub (2018). A World Government: A Critical Look into the Present, to Foresee the Future. *New York University Journal of International Law and Politics* **50**:4.

Michael J. Zimmerman (2006). Is Moral Obligation Objective or Subjective? *Utilitas* **18**: 329-61.

Michael J. Zimmerman (2008). *Living with Uncertainty: The Moral Significance of Ignorance.* Cambridge: Cambridge University Press.