Should longtermists recommend hastening extinction rather than delaying it?*

Richard Pettigrew

April 4, 2023

Abstract

Longtermists argue we should devote much of our resources to raising the probability of a long happy future for sentient beings. But most interventions that raise that probability also raise the probability of a long miserable future, even if they raise the latter by a smaller amount. If we choose by maximising expected utility, this isn't a problem; but, if we use a risk-averse decision rule, it is. I show that, with the same probabilities and utilities, a risk-averse decision theory tells us to hasten human extinction, not delay it. What's more, I argue that morality requires us to use a risk-averse decision theory. I present this not as an argument for hastening extinction, but as a challenge to longtermism.

Longtermism is the view that the most urgent global priorities, and those to which we should devote the largest portion of our resources, are

*I'd like to thank Teru Thomas, Andreas Mogensen, Hayden Wilkinson, Timothy Williamson, and Christian Tarney, Richard Chappell, and Alejandro Ortega for generous comments on earlier draft of this material; Marina Moreno and Adriano Mannino for long and extremely illuminating discussions; audiences in Bristol, Munich, Oxford, and the Varieties of Risk project for their insightful questions and challenges; and the referees for *The Monist* for suggestions that significantly improved the paper.

those that focus on (i) ensuring a long future for humanity, and perhaps sentient or intelligent life more generally, and (ii) improving the quality of the lives that inhabit that long future. While it is by no means the only one, the argument most commonly given for this conclusion is that these interventions have greater expected goodness per unit of resource devoted to them than each of the other available interventions, including those that focus on the health and well-being of the current population (Parfit, 1984; Beckstead, 2013; Greaves & MacAskill, 2021). In this paper, I argue that, even if we grant the consequentialist ethics upon which this argument depends, and even if we grant one of the axiologies that are typically paired with that ethics to give the argument, we are not morally required to choose an option that maximises expected utility; indeed, we might not even be permitted to do so. Instead, I will argue, if the argument's consequentialism is correct, we should choose using a decision theory that is sensitive to risk, and allows us to give greater weight to worse-case outcomes than expected utility theory does. And, I will show, such decision theories do not always recommend longtermist interventions. Indeed, sometimes, they recommend exactly the opposite: sometimes, they recommend hastening human extinction. Many, though not all, will take this as a reductio of the consequentialism or the axiology of the argument. I remain agnostic on the conclusion we should draw.

1 A simple model of the choice between interventions

I'll begin in this section by introducing a simple model of the decision problem we face when we must choose where to donate some substantial amount of money, and I'll assume a particular ethical account of how we

should make such a decision. Specifically, I will assume consequentialism, coupled with a total human hedonist utilitarian axiology, which I describe in greater detail below. This is not because I think all longtermists accept that account of morally right action. They do not. Rather, I want to use this straightforward account to explain the structure of my objection to the longtermist argument that I described in the introduction. In Sections 4.1 and 4.2, I will generalise my objection by showing that its conclusion is robust under changes to the axiological assumptions I make in this section. So, if upon reading the version of the objection I formulate in this section, you feel I am misrepresenting the view of many longtermists, hold that thought! I'll consider alternative views later, and we'll see that my objection goes through for them as well.

Let's suppose you have some substantial quantity of money at your disposal—perhaps you have a great deal of personal wealth, or perhaps you manage a large pot of philanthropic donations, or perhaps you make recommendations to wealthy philanthropists who tend to listen to your advice. And let's assume there are three options between which you must choose:

(*SQ*) You don't spend the money, and the status quo remains.

(*QEF*) You donate to the Quiet End Foundation, a charity that works to bring about a peaceful, painless end to humanity within the coming century.

(*HFF*) You donate to the Happy Future Fund, a charity that works to ensure a long happy future for the species by reducing extinction risks and improving the prospects for happy lives in the future.

We'll also assume that there are four possible ways the future might unfold,

and their probabilities will be affected in different ways by the different options you choose:

(*lh*) *The long, happy future*: This is the best-case scenario. Humanity survives for a billion years with a stable population of around 10 billion people at any given time. During that time, medical, technological, ethical, and societal advances ensure that the vast majority of people live lives of extraordinary pleasure and fulfilment.

(*mh*) *The long mediocre/medium-length happy future*: This is a sort of catch-all good-but-not-great option. It collects together many possible future states that share roughly the same goodness. In one, humanity survives the full billion years, some human lives are happy, some mediocre, some only just worth living, many are miserable. In another, humanity does not survive so long, but the average level of happiness is higher. And so on.

(*ext*) *The short mediocre future*. Humanity goes extinct in the next century with levels of happiness at a mediocre level.

(*lm*) *The long miserable future*. This is the worst-case scenario. Humanity survives for the full billion years with a stable population around 10 billion at any given time. During that time, the vast majority of people live lives of unremitting pain and suffering, perhaps because they are enslaved to serve the interests of a small oligarchy.

To complete our model, we must assign utilities to each of the possible states of the world, *lh*, *mh*, *ext*, and *lm*; and, for each of the three alternatives, *SQ*, *HFF*, and *QEF*, we must assign probabilities to each of the states conditional on choosing that intervention.

First, the utilities. On the consequentialist view that underpins the argument for longtermism that interests me in this paper, the utilities measure the goodness of the state of the world. But which account of goodness? As I mentioned above, in this section, and for the purpose of introducing the objection, I will assume the axiology of total human hedonist utilitarianism. That is, I will take the goodness of a state of the world to be its total human hedonic value, which weighs the amount, intensity, and nature of the human pleasure it contains against the amount, intensity, and nature of the human pain it contains. In Section 4.1, I will ask what happens if we use different accounts of the goodness of a state of the world.

To specify utilities, we must specify a unit. Let's say that each human life year lived with the sort of constant extraordinary pleasure envisaged in the long happy future scenario (*lh*) adds one unit of utility, or utile, to the goodness of the states of the future, regardless of how many such life years already exist there. Then the utility of *lh* is 10^{19} utiles, since it contains 10^{19} human life years at the very high level of pleasure. We'll assume that the utility of the catch-all short-and-very-happy or long-and-mediocre scenario (*mh*) is 10^{11} utiles, the equivalent of a decade of human existence at the current population levels and in which each life is lived at the extremely high level of pleasure envisaged in *lh*; or, of course, a much longer period of existence at this population level with a much more mediocre level of pleasure. The utility of the near-extinction scenario (*ext*) is 10^4 utiles, since it contains one hundred years lived at the same mediocre average level that, in scenario *mh*, when lived for a billion years, resulted in 10^{11} utiles. And finally the long miserable scenario (*lm*). Here, we assume that some lives contain such pain and suffering that they are genuinely not worth living; that is, they contribute negatively to the utility of the world. Indeed,

I'll assume it is possible to experience pain that is as bad as the greatest pleasure is good. That is, the utility of the worst case scenario is simply the negative of the utility of the best case scenario, where we are taking our zero point to be the utility of non-existence. So the utility of *lm* is -10^{19}.

	lh	*mh*	*ext*	*lm*
$U(-)$	10^{19}	10^{11}	10^{4}	-10^{19}

Second, the probabilities of each state of the world given each of the three options, *SQ*, *QEF*, and *HFF*. Again, I will give specific quantities here, but in Section 4.2, I will ask how the argument works if we change these numbers.

First, let's specify the status quo. It seems clear that the long mediocre or short happy future (i.e. *mh*) is by far the most likely, absent any intervention, since it can be realised in so many different ways. I'll use a conservative estimate for the probability of extinction (*ext*) in the next century, namely, one in a hundred ($\frac{1}{10^2}$). And I'll say that the long happy future, while very unlikely, is nonetheless much much more likely than the long miserable one. I'll say the long happy future (i.e. *lh*) is a thousand times less likely than extinction, so one in a hundred thousand ($\frac{1}{10^5}$); and the long miserable future (i.e. *lm*) is a hundred times less likely than that, so one in ten million ($\frac{1}{10^7}$). This discrepancy between the long happy future and the long miserable one is a popular assumption among longtermists. They justify it by pointing to the great increases in average well-being that have been achieved in the past thousand years; they assume that this trend is very likely to continue, and I'll grant them that assumption here. So, conditional on a long future that is either happy or miserable, a happy one is 99% certain, while a miserable one has a probability of only 1%. And, finally, I'll say that the long mediocre or short happy future (i.e. *mh*) mops

up the rest of the probability $(1 - \frac{1}{10^3} - \frac{1}{10^5} - \frac{1}{10^7})$.

Next, suppose you donate to the Quiet End Foundation (QEF) or to the Happy Future Fund (HFF). I'll assume that both change the probability of extinction by the same amount, namely, one in ten thousand ($\frac{1}{10^5}$). Donating to QEF increases the probability of extinction (*ext*) by that amount, while donating to HFF decreases it by the same. Then the probabilities of the other possible outcomes (*lh*, *mh*, *lm*) change in proportion to their prior probability.

So here are the probabilities, where k^- and k^+ take the values required to ensure the probabilities of the four possible states of the world sum to 1 in each row:[1]

	lh	*mh*	*ext*	*lm*
$P(-\|SQ)$	$\frac{1}{10^5}$	$1 - \frac{1}{10^2} - \frac{1}{10^5} - \frac{1}{10^7}$	$\frac{1}{10^2}$	$\frac{1}{10^7}$
$P(-\|QEF)$	$\frac{1}{10^5}k^-$	$\left(1 - \frac{1}{10^2} - \frac{1}{10^5} - \frac{1}{10^7}\right)k^-$	$\frac{1}{10^2} + \frac{1}{10^5}$	$\frac{1}{10^7}k^-$
$P(-\|HFF)$	$\frac{1}{10^5}k^+$	$\left(1 - \frac{1}{10^2} - \frac{1}{10^5} - \frac{1}{10^7}\right)k^+$	$\frac{1}{10^2} - \frac{1}{10^5}$	$\frac{1}{10^7}k^+$

Now, this is a forest of numbers, many of which seem so small as to be negligible. But it's reasonably easy to see that the expected utility of donating to the Happy Future Fund (*HFF*, illustrated in Figure 2) is greater than the expected utility of the status quo (*SQ*, illustrated in Figure 1), which is greater than the expected utility of donating to the Quiet End Foundation

[1] That is,

- $k^+ = 1 + \frac{\frac{1}{10^5}}{1-\frac{1}{10^2}}$ and
- $k^- = 1 - \frac{\frac{1}{10^5}}{1-\frac{1}{10^2}}$

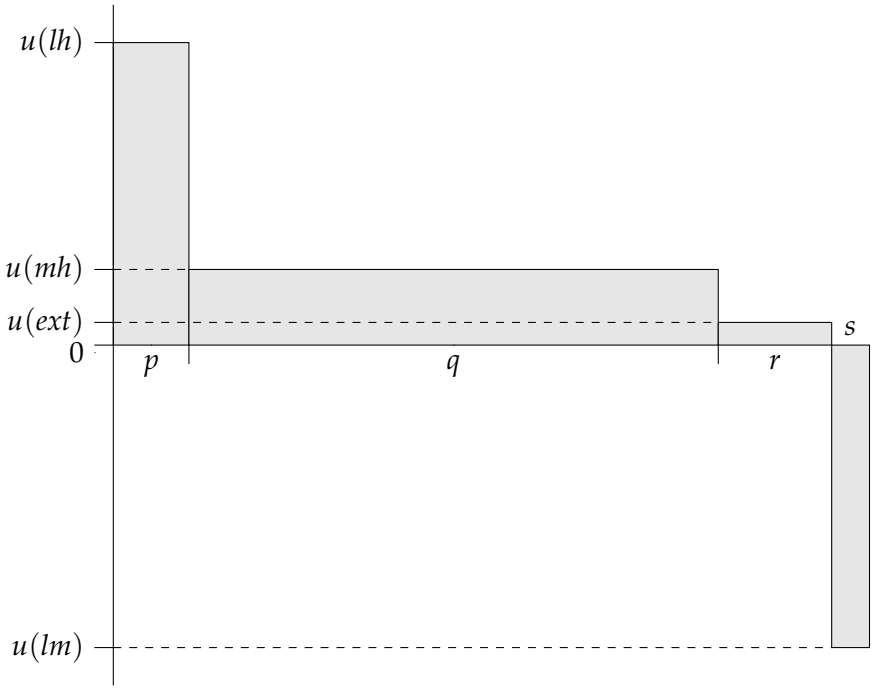

Figure 1: The expected utility of the status quo *SQ* is given by the grey area, where *p* is the probability of *lh*, *q* is the probability of *mh*, *r* is the probability of *ext*, and *s* is the probability of *lm* (any area below the zero line counts negatively). Not to scale!

(*QEF*, illustrated in Figure 3). After all, the Quiet End Foundation takes away more probability from the best outcome (*lh*) than it takes away from the worst outcome (*lm*); and it takes away probability from the second-best outcome (*mh*) while adding it to the second-worst outcome (*ext*). So it has a negative effect in expectation. The Happy Future Fund, in contrast, adds more probability to the best outcome than to the worst outcome, and it adds to the second-best while taking away from the second-worst. So it has a positive effect in expectation.

Indeed, if you donate to the Happy Future Fund, you increase the expected utility of the world by around one billion utiles. Recall, that's one billion human life years lived at an extraordinary level of well-being. If the

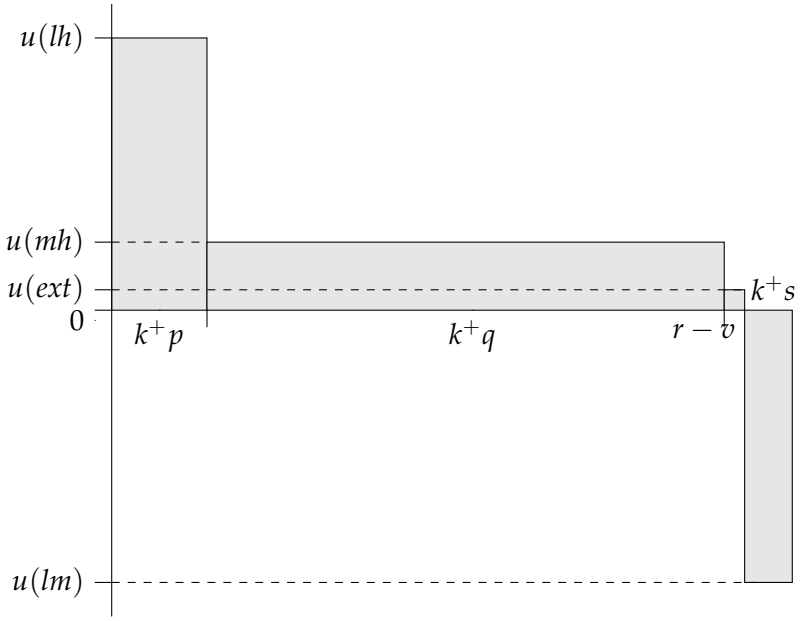

Figure 2: The expected utility of *HFF* is given by the grey area, where k^+p is the probability of *lh* given that you donate to the Happy Future Fund, k^+q is the probability of *mh* given that, $r - v$ is the probability of *ext* given that, and k^+s is the probability of *lm* given that. So v is the amount by which your donation decreases the probability of extinction and $k^+ = 1 + \frac{v}{1-r}$ is the factor by which the other probabilities are scaled. Not to scale!

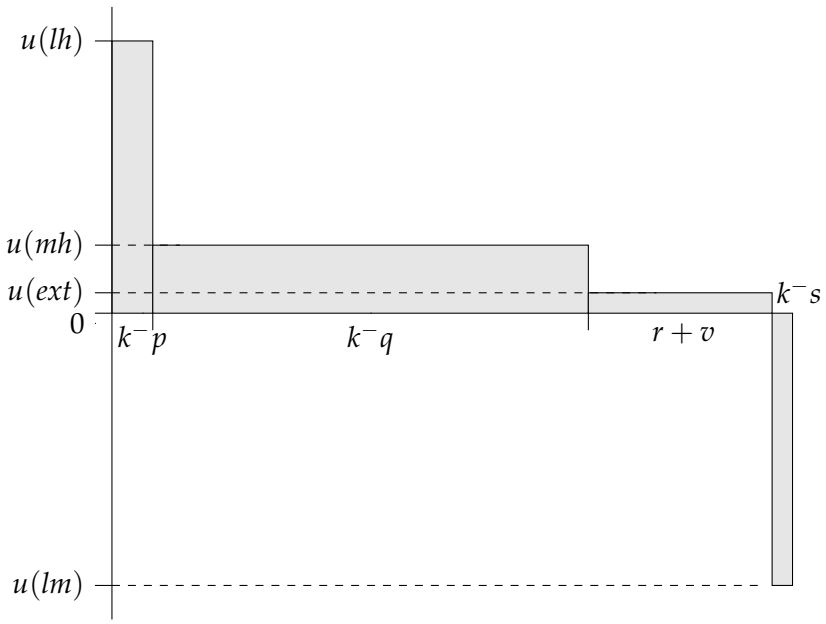

Figure 3: The expected utility of *QEF* is given by the grey area, where $k^- p$ is the probability of *lh* given that you donate to the Quiet End Foundation, $k^- q$ is the probability of *mh* given that, $r + v$ is the probability of *ext* given that, and $k^- s$ is the probability of *lm* given that. So v is the amount by which your donation increases the probability of extinction and $k^- = 1 - \frac{v}{1-r}$ is the factor by which the other probabilities are scaled. Not to scale!

same amount of money could, with certainty, have saved a hundred children under five years old from a fatal illness, that would only have added around seven thousand human life years, and they would not have been lived at this very high level of well-being. So, according to the assumption we are making throughout this section that we should do whatever maximises expected total human hedonic utility, we should donate to the Happy Future Fund instead of a charity that saves the lives of those vulnerable to preventable disease. And if you donate to the Quiet End Foundation, you decrease the expected utility of the world by around one billion utiles. Small shifts in probabilities can make an enormous difference when the utilities involved are so vast.

The upshot of this section is that, from the point of view of expected utility, when the utility is the total human hedonic utility, the Happy Future Fund is by far the best, then the status quo, and then the Quiet End Foundation. So, for the consequentialist who adopts that axiology, we should donate to the Happy Future Fund.

2 Rational choice theory and risk

The version of the longtermist argument sketched in the previous section concluded that we should donate to the Happy Future Fund instead of maintaining the status quo or donating to the Quiet End Foundation because doing so maximises expected goodness. In this section, I want to argue that even a consequentialist who adopts the total human hedonic utilitarian axiology that we've been assuming so far should not say that we are required to choose the option that maximises expected goodness. Rather, we are either permitted or required to take considerations of risk

into account.

Different versions of utilitarianism, and consequentialism more generally, supply us with an axiology. They tell us how much goodness each possible state of affairs contains. And they tell us that the morally *best* action is the one that maximises this goodness; it is the one that, if performed, will in fact bring about the greatest goodness. To derive from this an account of what the morally *right* action is for an individual who is uncertain about what states their actions will bring about, we must combine our axiology with an account of decision-making under uncertainty. A version of consequentialism provides the *ends* of moral action, and decision theory then tells us the *means*. Since orthodox decision theory tells you that prudential rationality requires you to choose by maximising expected utility, consequentialists often say that morality requires you to choose by maximising expected goodness. However, since the middle of the twentieth century, many decision theorists have concluded that prudential rationality requires no such thing. Instead, they say, you are permitted to make decisions in a way that is sensitive to risk. In this section, I want to argue that consequentialists, including those longtermists who subscribe to that view, should follow their lead.

Consider the following example.[2] Sheila is a total hedonist utilitarian with some money to donate. Her philanthropic consultant offers two choices. If she chooses the first, then with certainty exactly 49 people will exist in the future after her death who wouldn't otherwise exist, and they'll

[2]For further motivations for risk-sensitive decision theories, see (Buchak, 2013, Chapters 1 and 2). The shortcomings of expected utility theory were first identified by Allais (1953). He presented four different options, and asked us to agree that we would prefer the first to the second and the fourth to the third. He then showed that there is no way to assign utilities to the outcomes of the options so that these preferences line up with the ordering of the options by their expected utility. For a good introduction, see (Steele & Stefánsson, 2020, Section 5.1).

live happy lives; if she chooses the second, then with 50% chance exactly 100 people will exist in the future after her death who wouldn't otherwise exist, and they'll live happy lives, and with 50% chance no extra people will exist in the future.

Here's the payoff table for her choice (with one utile per happy life lived, and N the amount of utility the universe will contain independently of her choice):

	World 1	World 2
Option 1	$N + 49$	$N + 49$
Option 2	N	$N + 100$

According to expected utility theory, Sheila should choose Option 2, since it has the higher utility in expectation ($N + 50$ utiles vs $N + 49$ utiles). And yet it seems quite rational for her to choose Option 1. After all, Option 1 is guaranteed to create goodness, and indeed quite a lot of it, while Option 2 has a substantial probability of creating none. If Sheila chooses Option 1, we might say that she is risk-averse in a way that is perfectly rational. Option 2 is a risky option: it gives the possibility of the best outcome, but it also opens the possibility of the worst outcome. In contrast, Option 1 is a risk-free option: it gives no possibility of the best outcome, but equally no possibility of the worst one either; it guarantees a middling outcome.[3]

Standard expected utility theory says that the weight that each outcome receives before they are summed to give the expected utility of an option is just the probability of that outcome given that you choose the option. But this ignores the risk-sensitive agent's desire to take into account not only the probability of the outcome but where it ranks in the ordering of

[3]By assuming that the lives created exist after her death, we rule out the possibility that the reason Sheila chooses Option 1 is that she factors in the regret she'll feel if she chooses Option 2 and no lives are created. Thanks to Alejandro Ortega for pushing me to consider regret aversion.

outcomes from best to worst. The risk-averse agent like Sheila wants to give greater weight to worse case outcomes—such as the outcome of Option 2 at World 1—than expected utility theory requires and less weight to the better case outcomes—such as the outcome of Option 2 at World 2. The risk-seeking agent will wish to give less weight to the worse cases and more to the better cases.

How might we capture this in our theory of rational choice? The most sophisticated and best developed way to amend expected utility theory to accommodate these considerations is due to Lara Buchak (2013), building on work by John Quiggin (1993), and it is called *risk-weighted expected utility theory*. In Buchak's theory, we model your attitudes to risk as a function R that takes numbers between 0 and 1 and returns a number between 0 and 1. We assume that R has three properties:

(i) $R(0) = 0$ and $R(1) = 1$,

(ii) R is strictly increasing, so that if $p < q$ then $R(p) < R(q)$, and

(iii) R is continuous.

Some examples: for any $0 < k < \infty$, let $R_k(x) = x^k$. So $R_2(x) = x^2$, $R_1(x) = x$, $R_{0.5} = \sqrt{x}$, and so on. Each R_k is a Buchakian risk function.

Now, to illustrate how risk-weighted expected utility theory incorporates these risk attitudes represented by a risk function R, suppose there are just three states of the world, S_1, S_2, and S_3. Suppose O is an option with the following utilities at those states:

	S_1	S_2	S_3
$U(- \,\&\, O)$	u_1	u_2	u_3

And, on the supposition that O is chosen, the probabilities of the states are these:

	S_1	S_2	S_3
$P(-\|O)$	p_1	p_2	p_3

And, suppose S_1 is the worst case outcome for O, S_2 is the second-worst (and also second-best), and S_3 is the best case. That is, $u_1 \leq u_2 \leq u_3$. Then the expected utility of O is

$$\mathrm{EU}(O) = p_1u_1 + p_2u_2 + p_3u_3$$

So the weight assigned to the utility u_i is the probability p_i. Now notice that, given O, the probability p_i of a state S_i is equal to the probability that O will obtain for you *at least utility* u_i less the probability that it will obtain for you *more than that utility*. So

$$\mathrm{EU}(O) = [(p_1 + p_2 + p_3) - (p_2 + p_3)]u_1 + [(p_2 + p_3) - p_3]u_2 + p_3u_3$$

Now, when we calculate the risk-weighted expected utility of O, the weight for utility u_i is the *risk-transformed* probability that O will obtain for you *at least utility* u_i less the *risk-transformed* probability that it will obtain for you *more than that utility*. So

$$\begin{aligned}\mathrm{REU}(O) = \\ &[R(p_1 + p_2 + p_3) - R(p_2 + p_3)]u_1 + \\ &\quad [R(p_2 + p_3) - R(p_3)]u_2 + \\ &\quad\quad R(p_3)u_3\end{aligned}$$

Easily the clearest way to understand how Buchak's theory works is by considering the following diagrams, where R is a convex risk function, such as $R_2(x) = x^2$. In Figure 4, the area of each rectangle gives the utility of each state of the world weighted by the weight that is applied to it in the

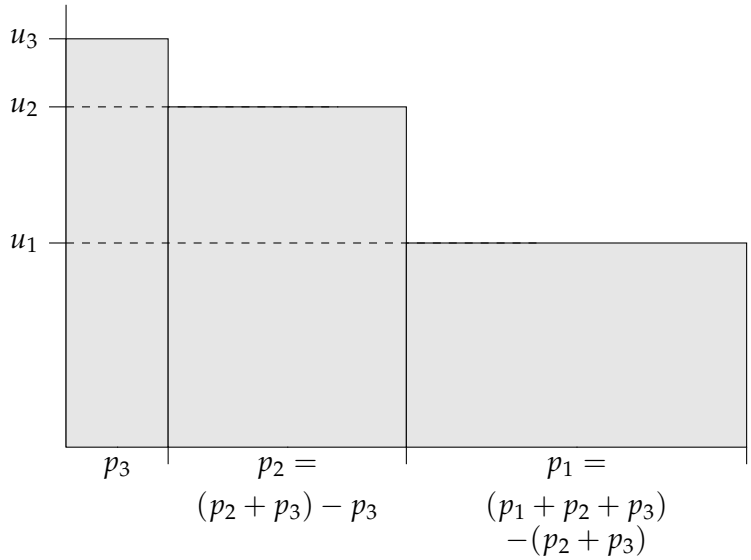

Figure 4: The expected utility of O is given by the grey area.

calculation of expected utility. For instance, the area of the right-most rectangle is the utility of state S_1 multiplied by the probability of state S_1 given the option is chosen: that is, it is p_1u_1, or $[(p_1 + p_2 + p_3) - (p_2 + p_3)]u_1$. So the total area of all the rectangles is the expected utility of the option. In Figure 5, the area of each rectangle gives the utility of each state weighted by the weight that is applied to it in the calculation of risk-weighted expected utility. For instance, the area of the right-most rectangle is the utility of state S_1 multiplied by the risk-transformed probability that O will obtain for you at least that utility less the risk-transformed probability that it will obtain for you more than that utility: that is, it is $[R(p_1 + p_2 + p_3) - R(p_2 + p_3)]u_1$. So the total area of all the rectangles is the risk-weighted expected utility of the option.

If R is convex—e.g. $R(x) = x^k$, for $k > 1$—then the individual is risk-averse, for then the weights assigned to the worse case outcomes are greater than those that expected utility theory assigns, while the weights assigned

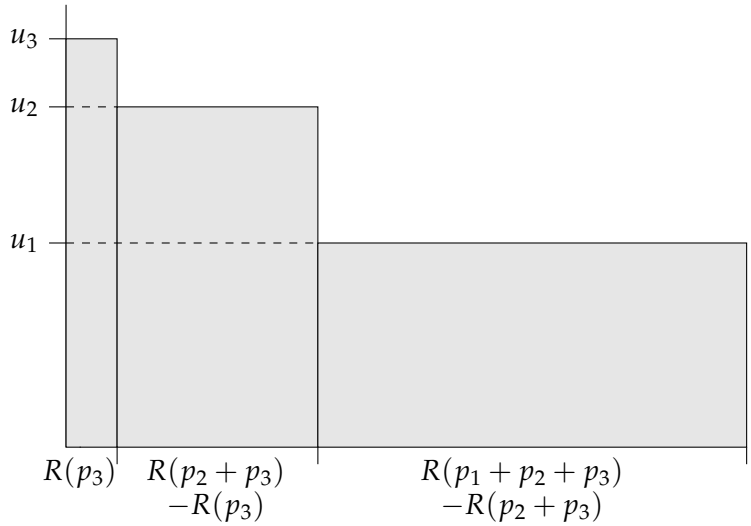

Figure 5: The risk-weighted expected utility of O is given by the grey area.

to the best case outcomes are less. If R is concave—e.g. $R(x) = x^k$, for $k < 1$—the individual is risk-inclined. And if R is linear—so that $R(x) = x$—then the risk-weighted expected utility of an option is just its expected utility, so the individual is risk-neutral.

Now let us apply this to the choice between doing nothing, donating to the Happy Future Fund, and donating to the Quiet End Foundation.

Suppose your risk function is $R_k(x) = x^k$ for $k > 1.5$. So you are risk averse. Indeed, for $k = 1.6$, you have the level of risk aversion that would lead you, in Sheila's situation, to prefer a guarantee of creating 35 happy lives to a 50% chance of creating 100 and a 50% chance of creating none, or a guarantee of creating 72 happy lives to an 80% chance of creating 100, or a guarantee of 4 to a 10% chance of 100. So you're risk-averse, but only moderately. Then

$$\text{REU}(HFF) < \text{REU}(SQ) < \text{REU}(QEF)$$

That is, the risk-weighted expected utility of donating to the Quiet End Foundation is greater than the risk-weighted expected utility of the status quo, which is itself greater than the risk-weighted expected utility of donating to the Happy Future Fund. So, you should not donate to the Happy Future Fund, and you should not do nothing; you should donate to the Quiet End Foundation.

The upshot: if we replace expected utility theory with risk-weighted expected utility theory in the argument for longtermism with which we began this paper, and retain the total human hedonic utilitarian axiology we've been assuming so far, then the conclusion will be that we should give different advice to individuals with different attitudes to risk; and indeed we should advise mildly risk-averse individuals like the one just described to donate to charities that work towards a peaceful end to humanity.

3 What we together risk

The conclusion of the previous section is a little alarming. If moral choice is just rational choice but with a utility function that measures total human hedonic value in place of one that measures our own subjective conception of goodness, even moderately risk-averse members of our society should focus their philanthropic actions on hastening the extinction of humanity. But I think things are worse than that. I think it follows not only that the risk-averse in our society should donate their money to such causes, but that everyone should, whether they are risk-averse, risk-inclined, or risk-neutral. In this section, I'll try to explain why.

Consider the following case. You're organising an indoor event during the COVID-19 pandemic. You have a certain budget that you could use

either to pay for better air filters for the venue, or better catering for the guests. Paying for better air filters is risk-free: regardless of whether anyone attending is infectious, they won't infect anyone else if the new filters are fitted, so this option has the same middling utility either way. Paying for the better catering is risky: it will improve the party for sure, but if someone's infectious, the increase to the utility provided by the better catering will be outweighed by the disutility of guests becoming ill, and all told it will have a low utility; on the other hand, if no-one is infectious, it will have a high utility.

You know that everyone in the group shares the same utility function—they assign the same high utility to better food and no infections, the same middling utility to mediocre food and no infections, and the same low utility to the better food and some infections. And you know they assign the same credence to whether anyone attending will be infectious. You're risk-neutral yourself, and so what's rational for you to choose is what maximises expected utility. When you do the calculation, it turns out that paying for the better catering has greater expected utility than paying for the air filters. However, you also know that most of the guests at the event are risk-averse to the extent that paying for the air filters maximises risk-weighted expected utility relative to their risk functions. What should you do?

It seems to me that you should pay for the air filters rather than the catering. This suggests that, when we make a decision that affects other people with different attitudes to risk, and when one of the possible outcomes of that decision involves harm to those people, we should give greater weight to the preferences of the risk-averse among them than to the risk-

neutral or risk-inclined.[4] If that's so, then it might be that the effective altruist should not only advise the risk-averse to donate to the Quiet End Foundation, but should advise everyone in this way. After all, this example suggests that the morally right choice is the rational choice when the utility function measures morally relevant goodness and the attitude to risk is determined by aggregating the risk attitudes of all the people who will be affected by the decision in some way that gives greater weight to the attitudes of the more risk-averse. And the empirical evidence suggests that most people are reasonably risk-averse (MacCrimmon & Larsson, 1979; Rabin & Thaler, 2001; Oliver, 2003).

What I have offered, then, is not a definitive argument that the longtermists must now focus their energies on bringing about the extinction of humanity and encouraging others to donate their resources to helping. But I hope to have made it pretty plausible that this is what their most popular argument in fact says they should do.

4 How should we respond to this argument?

How should we respond to these two arguments? The first is for the weaker conclusion: for many people who are risk averse, the morally correct choice is to donate to the Quiet End Foundation. The second is for the stronger conclusion: for everyone, regardless of their attitude to risk, the morally correct choice is to donate in that way.

[4] See (Thoma, 2023) for an overview of the literature on how we should act when our decisions affect others with different risk attitudes to our own. My own thinking owes much to Lara Buchak's discussion of her Risk Principle, though that covers a slightly different sort of situation from the one in question here (Buchak, 2017, 632).

4.1 Changing the utilities: conceptions of goodness

One natural place to look for the argument's weakness is in its axiology. Throughout, we have assumed the austere, monistic conception of morally relevant goodness offered by the hedonist utilitarian and restricted only to human pleasure and pain.

So first, we might expand the pale of moral consideration to include non-human animals and non-biological sentient beings, such as artificial intelligences, robots, and minds inside computer simulations. But, this is unlikely to change the problem significantly. It only means that there are more minds to contain great pleasure in the long happy future (*lh*), but also more to contain great suffering in the long miserable one (*lm*). And of course there is the risk that humanity continues to give non-human suffering less weight than we should, and as a result non-human animals and artificial intelligences are doomed to live miserable lives, just as factory-farmed animals currently do. While longtermists are surely right that the average well-being of humans has risen dramatically over the past few centuries, the average well-being of livestock has plummeted at the same time as their numbers have dramatically increased. If we simply multiply the utility of each outcome by the same factor to reflect the increase in morally relevant subjects, this will change nothing, since risk-weighted expected utility comparisons are invariant under positive linear transformations of utility—you can scale everything up by a factor and add some fixed amount and everything remains the same. And if we increase the utility of *lh* and *mh*, decrease the utility of *lm*, and leave the utility of *ext* untouched, on the grounds that the extra beings we wish to include within the moral pale are artificial intelligences that are yet to exist and so won't exist in significant numbers within future *ext*, then this in fact merely widens the gap

between the risk-weighted expected utility of donating to the Quiet End Foundation and the risk-weighted utility of donating to the Happy Future Fund. And the same happens if we entertain the more extreme estimates for the possible number of beings that might exist in the future, which arise because we colonise beyond Earth.

Second, we might change what contributes to the morally relevant goodness of a situation. For instance, we might say that there are features of a world that contains flourishing humans that add goodness, while there are no corresponding features of a world that contains miserable humans that add the same badness. One example might be the so-called higher goods of aesthetic and intellectual achievements. In situation *lh*, we might suppose, people will produce art, poetry, philosophy, music, science, mathematics, and so on. And we might think that the existence of such achievements adds goodness over and above the pleasure that people experience when they engage with them; they somehow have an intrinsic goodness as well as an instrumental goodness. This would boost the goodness of *lh*, but it leaves the badness of *lm* unchanged, since the absence of these goods is neutral, and there is nothing that exists in *lm* that adds further badness to *lm* in the way these higher goods add goodness to *lh*. If these higher goods add enough goodness to *lh* without changing the badness of *lm*, then it may well be that even the risk-averse will prefer the Happy Future Fund over the Quiet End Foundation. See Figure 6 for the effects of this on the difference between the risk-weighted expected utility of the Quiet End Foundation and the risk-weighted expected utility of the Happy Future Fund.[5]

Of course, the most obvious move in this direction is simply to assume

[5] A Mathematica notebook with all the calculations and figures included in this paper is available here: `https://drive.google.com/file/d/1JG7YLjExKISy-YGYkMLUbRy2W8jagYvJ/`

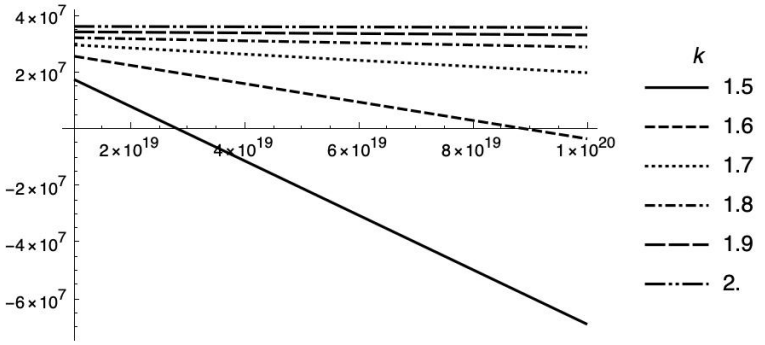

Figure 6: This plots $\text{REU}(QEF) - \text{REU}(HFF)$ as the utility of *lh* increases from 10^{19} to 10^{20}, for different risk functions of the form $R_k(x) = x^k$. So, for risk function $R_{1.5}$ (the solid line), the Quiet End Foundation is better than the Happy Future Fund providing whatever extra non-hedonic goodness *lh* includes does not increase its utility by a factor of more than 2.5. On the other hand, for $R_{1.6}$ (the dashed line immediately above the solid line), *QEF* is better than *HFF*, even if the non-hedonic goodness multiplies the utility of *lh* by a factor of up to around 8.5.

that the existence of humanity adds goodness beyond the pleasure or pain experienced by the humans who exist. Or perhaps it's not the existence of humans specifically that adds the value, but the existence of beings from some class to which humans belong, such as the class of intelligent beings or moral agents or beings capable of ascribing meaning to the world and finding value in it. Again, the idea is that the existence of these creatures is good independent of the work to which they put their special status. So, as for the case of the higher goods, this would add goodness to *lh*, which contains such creatures, but not only would it not add corresponding badness to *lm*; it would in fact add goodness to *lm*, since *lm* contains these beings who boast the special status. And it might add enough goodness to *lh* and *lm* that it would reverse the risk-averse person's preferences between the charities.

My own view is that it is better not to think of the existence of intelligent beings or moral agents as adding goodness regardless of how they

deploy that intelligence or moral agency. Rather, when we ascribe morally relevant value to the existence of humans, we do so because of their potential for doing things that are valuable, such as creating art and science, loving and caring for one another, making each other happy and fulfilled, and so on. But in scenario *lm*, the humans that exist do not fulfil that potential, and since that scenario specifies all aspects of the world's history—past, present, and future—there is no possibility that they will fulfil it, and so there is no value added to that scenario by the fact that beings exist in it that might have done something much better. And, at least if we suppose that the misery in scenario *lm* is the result of human cruelty or lack of moral care, we might think the fact that the misery is the result of human immorality makes it have lower moral utility.

For those who prefer an axiology on which it is not the hedonic features of a situation that determine its morally relevant goodness, but rather the degree to which the preferences of the individuals who exist in that situation are satisfied, you might hope to appeal to the fact that people have a strong preference for humanity to continue to exist, which gives a substantial boost to *lh* and *lm*, perhaps enough to make the Happy Future Fund the better option. But I think this only seems plausible because we've grained our preferences too coarsely. People do not have a preference for humanity to continue to exist *regardless of how humans behave and the quality of the lives they live*. They have a preference for humanity continuing in a way that is, on balance, positive. So adding the good of preference satisfaction to the hedonic good will likely boost the goodness of *lh*, since *lh* contains a lot of pleasure and also satisfies the preferences of nearly everyone, but it will also boost the badness of *lm*, since *lm* contains a lot of pain and also thwarts the preferences of nearly everyone.

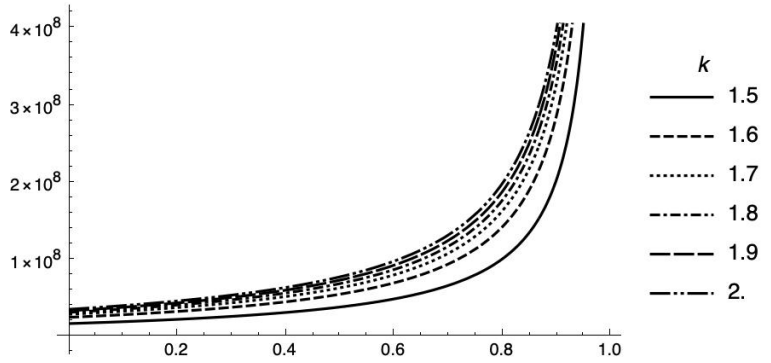

Figure 7: This plots REU(QEF) − REU(HFF) as the antecedent probability of *ext* ranges from 0 to 1, for different risk functions of the form $R_k(x) = x^k$. For each risk function, and for all probabilities for extinction, the Quiet End Foundation is better than the Happy Future Fund.

4.2 Changing the probabilities

In the previous section, we asked how different conceptions of goodness might change the utilities we've assigned to the four outcomes *lh*, *mh*, *ext*, and *lm* in our model. Now, we turn to the probabilities we've posited.

As I mentioned above, I used a conservative estimate of $\frac{1}{100}$ for the probability of near-term extinction. Toby Ord (2020) places the probability at $\frac{1}{6}$. At the time of writing, users of the opinion aggregator *Metaculus* place it at $\frac{1}{50}$.[6] How do these alternative probabilities affect our calculation? Figure 7 gives the results. The answer is that, for any risk function $R_k(x) = x^k$ with $k \geq 1.5$, our conclusion that donating to the Quiet End Foundation (QEF) is better than donating to the Happy Future Fun (HFF) is robust under any change in the probability of extinction.

Next, consider the change in the probabilities that we can affect by donating either to the Happy Future Fund or the Quiet End Foundation. I assumed that, either way, we'd change the probability of extinction by $\frac{1}{10^5}$—

[6]https://www.metaculus.com/questions/578/human-extinction-by-2100/. Retrieved 2nd August 2022.

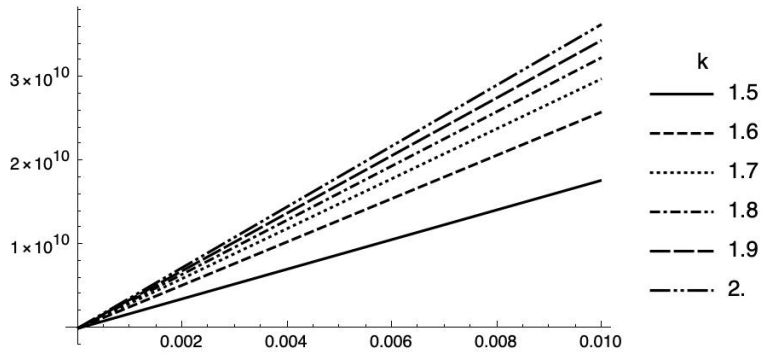

Figure 8: This plots REU(QEF) − REU(HFF) as the change in the probability of extinction that our intervention can achieve ranges from $\frac{1}{10^7}$ to $\frac{1}{10^2}$, for different risk functions of the form $R_k(x) = x^k$.

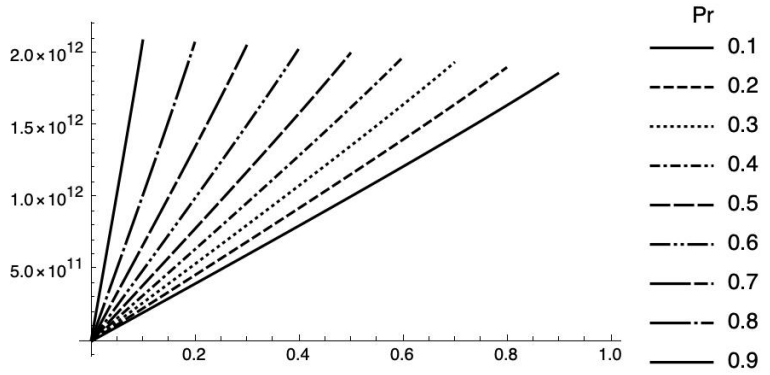

Figure 9: This plots REU(QEF) − REU(HFF) as the change in probability our intervention can achieve ranges from $\frac{1}{10{,}000}$ to 1, for risk function $R_{1.5}$ and different value Pr for the prior probability of extinction.

the Happy Future Fund decreases it by that amount; the Quiet End Foundation increases it by the same. But perhaps our intervention would have a larger or smaller effect than that. Figures 8 and 9 illustrate the effects. Again, our conclusion is robust: donating to the Quiet End Foundation is better than donating to the Happy Future Fund.

Finally, in our original model we assume that, after the intervention, the conditional probabilities of the three non-extinction options conditional on extinction not happening remained unchanged. The probability we re-

move from *ext* by donating to the Happy Future Fund is distributed to *lh*, *mh*, and *lm* in proportion to their prior probabilities. But we might think that, as well as reducing the probability of extinction, some of our donation might go to improving the probability of the better futures conditional on there being any future at all. But of course, if that's what the Happy Future Fund is going to do with our money, the Quiet End Foundation can do the same with the same amount of money. As well as working to increase the probability of extinction, some of our donation to the Quiet End Foundation might go towards improving the probability of the better futures conditional on there being any future at all. Above, we assumed that the probabilities of *lh*, *mh*, and *lm*, given that you donate to the Quiet End Foundation, are just their prior probabilities multiplied by the same factor k^-. And similarly the probabilities of *lh*, *mh*, and *lm* given that you donate to the Happy Future Fund are just their prior probabilities multiplied by the same factor k^+. But now suppose that the probability of *lh* given *QEF* is its prior probability multiplied by a factor $2nl^+$, the prob of *mh* given *QEF* is its prior probability multiplied by factor nl^+, and the prob of *lm* given *QEF* is its prior probability multiplied by factor l^+. And similarly for *HFF*, but with $2nl^-$, nl^-, and l^-. Then for what values of n is *QEF* still better than *HFF*? Figure 10 answers the question. For our original risk function $R_{1.5}$, *HFF* quickly exceeds *QEF*. But for an only slightly more risk-averse individual, with risk function $R_{1.7}$, *QEF* beats *HFF* for up to nearly $n = 5$.

This is the first time we've anything less than robustness in our conclusion about the relative merits of *QEF* and *HFF*. It illustrates an important point. While the result that risk-averse individuals should prefer *QEF* to *HFF* is reasonably robust for risk-aversion represented by $R_{1.5}$, there are certain ways in which we might change our model so that this robustness

disappears, and the ordering of the two interventions becomes very sensitive to certain features, such as degrees of risk-aversion. For instance, you might think that the lesson of Figure 10 is that our longtermist interventions should balance more towards improving the future conditional on its existence and less towards ensuring the existence of the future. But we see that, for R_2, even $n = 9$ favours *QEF*. So, if this is the risk function we're using to make moral choices, the rebalancing will have to be very dramatic in order to favour *HFF*.

The results presented in this chapter go a long way to answering a question about the objection I'm raising against this popular longtermist argument. You might worry that I have assumed illegitimately that we should use precise probabilities to represent our uncertainty about how the future might develop given the three different interventions; you might think we should instead represent that uncertainty using imprecise probabilities. That is, we should represent our uncertainty concerning human extinction given the status quo, say, not by a single probability, but by a range of probabilities; and our uncertainties in all the relevant states of the world given the different possible interventions should be represented not by a single probability function, but by a set of them. Now, if we were to do that, we'd then have to explain how to make decisions using these imprecise probabilities, or sets of probabilities, and that's a notoriously hard task (Elga, 2010; Bradley & Steele, 2014; Bradley, 2016). But one thing that all the putative decision theories for imprecise probabilities agree upon is that, if every probability function in the set that represents your uncertainty agrees that one option is better than another, then the set agrees on that as well. But, in this section we've seen that, by the lights of each probability function in a wide range, *QEF* is better than *HFF*; so the set that collects all these

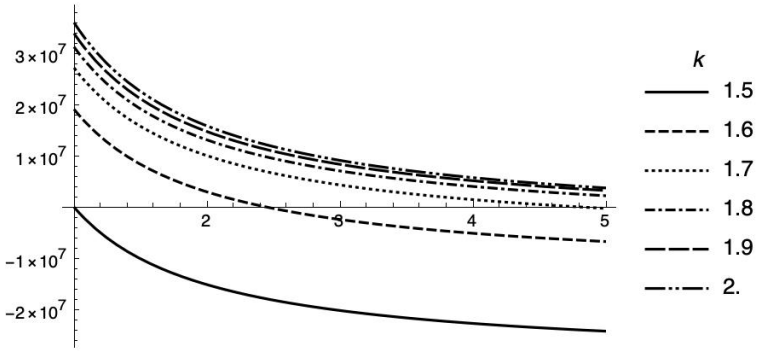

Figure 10: This plots $REU(QEF) - REU(HFF)$ as n ranges from 1 to 5, for risk functions of the form $R_k(x) = x^k$. Recall, the probability of *lh* given *QEF* is $2nl^-$, the probability of *mh* given *QEF* is nl^-, and the probability of *lm* given *QEF* is l^-. And their probabilities given *HFF* are the same, but with l^+ in place of l^-.

individual probability functions together itself agrees with that judgment.

5 Conclusion

I've presented arguments for two conclusions, one stronger than the other. Both are derived from the popular argument for longtermism that I presented in the introduction. The original version of that argument assumes that the morally right action is the one that maximises expected utility, where the utility is given by our axiology. My versions submit that the morally right action is, instead, the one that maximises risk-weighted expected utility, for a particular risk function. The first version claims you should use your own risk function. In this case, if you are sufficiently averse to risk, and you wish to donate some money to do good, then you should donate it to organisations working to hasten human extinction rather than ones working to prevent it. The second claims you should use a risk function that aggregates the risk attitudes of the people who will be affected by your decision, giving greater weight to the more risk-averse. In

this case, whether or not you are averse to risk yourself, you should donate to hasten the end of humanity.

What, then, is the overall conclusion? I confess, I don't know. Perhaps I have provided a reductio ad absurdum of a consequentialist approach that is often used to justify longtermism. Perhaps, but I don't claim that with any confidence. While it seems to me that the conclusions of my two versions of the argument must be wrong, I can't pinpoint where the argument itself goes astray. What I hope this paper will do is neither make you change the direction of your philanthropy nor lead you to reject the framework of longtermism for which the original argument is an important justification. Rather, I hope it will encourage you to think more carefully about how risk and our attitudes towards it should figure in our moral decision-making.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica*, *21*(4), 503–546.

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. Ph.D. thesis, Rutgers University, New Jersey.

Bradley, S. (2016). Imprecise Probabilities. In E. N. Zalta (Ed.) *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Bradley, S., & Steele, K. (2014). Should Subjective Probabilities Be Sharp? *Episteme*, *11*(3), 277–289.

Buchak, L. (2013). *Risk and Rationality*. Oxford, UK: Oxford University Press.

Buchak, L. (2017). Taking Risks Behind the Veil of Ignorance. *Ethics*, *127*(3), 610–644.

Elga, A. (2010). Subjective Probabilities Should Be Sharp. *Philosophers' Imprint*, *10*(5), 1–11.

Greaves, H., & MacAskill, W. (2021). The case for strong longtermism. GPI Working Paper No. 5, Global Priorities Institute, Oxford.

MacCrimmon, K. R., & Larsson, S. (1979). Utility Theory: Axioms versus 'Paradoxes'. In M. Allais, & O. Hagen (Eds.) *Expected Utility Hypotheses and the Allais Paradox*, vol. 21 of *Theory and Decision Library*. Dordrecht: Springer.

Oliver, A. (2003). A Quantitative and Qualitative Test of the Allais Paradox using Health Outcomes. *Journal of Economic Psychology*, *24*(1), 35–48.

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

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Quiggin, J. (1993). *Generalized Expected Utility Theory: The Rank-Dependent Model*. Kluwer Academic Publishers.

Rabin, M., & Thaler, R. H. (2001). Risk Aversion. *The Journal of Economic Perspectives*, *15*(1), 219–32.

Steele, K., & Stefánsson, H. O. (2020). Decision Theory. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 ed.

Thoma, J. (2023). Taking Risks on Behalf of Another. *Philosophy Compass*.