Effective altruism, risk, and human extinction

Richard Pettigrew (University of Bristol)

Global Priorities Institute | August 2022

GPI Working Paper No . 2-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Should longtermists recommend hastening extinction rather than delaying it?[*]

Richard Pettigrew

August 4, 2022

Abstract

Longtermism is the view that the most urgent global priorities, and those to which we should devote the largest portion of our current resources, are those that focus on ensuring a long future for humanity, and perhaps sentient or intelligent life more generally, and improving the quality of those lives in that long future. The central argument for this conclusion is that, given a fixed amount of a resource that we are able to devote to global priorities, the longtermist's favoured interventions have greater expected goodness than each of the other available interventions, including those that focus on the health and well-being of the current population. In this paper, I argue that, even granting the longtermist's axiology and their consequentialist ethics, we are not morally required to choose whatever option maximises expected utility, and may not be permitted to do so. Instead, if their axiology and 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. And such decision theories do not recommend longtermist interventions. Indeed, sometimes, they recommend hastening human extinction. Many, though not all, will take this as a reductio of the longtermist's axiology or consequentialist ethics. I remain agnostic on the conclusion we should draw.

Longtermism is the view that the most urgent global priorities, and those to which we should devote the largest portion of our current resources, are those that focus on two things: (i) ensuring a long future for humanity, and perhaps sentient or intelligent life more generally, and (ii)

[*] I'd like to thank Teru Thomas, Andreas Mogensen, Hayden Wilkinson, Tim Williamson, and Christian Tarney for extraordinarily generous comments on an earlier draft of this material; Marina Moreno and Adriano Mannino for long and extremely illuminating discussions; and audiences in Bristol, Munich, Oxford, and the Varieties of Risk project for their insightful questions and challenges.

improving the quality of the lives led in that long future.[1] The central argument for the longtermist's conclusion is that, given a fixed amount of a resource that we are able to devote to global priorities, the longtermist's favoured interventions have greater expected goodness than each of the other available interventions, including those that focus on the health and well-being of the current population. Longtermists might disagree on what determines how much goodness different futures contain: for some, it might be the total amount of pleasure humans experience in that future less the total amount of pain they experience; for others, the pleasures and pains of non-human sentient beings might count as well; others still might also include non-hedonic sources of value. But they agree that, once that quantity is specified, we are morally required to do whatever maximises it in expectation. And, they claim, whichever of the plausible conceptions of goodness you use, the interventions favoured by the longtermists maximise that quantity in expectation. Indeed, Greaves & MacAskill (2021) claim that the expected goodness of these interventions so far outstrips the expected goodness of the alternatives that even putatively nonconsequentialist views, such as deontologism, should consider them the right priorities to fund and pursue.

In this paper, I object to longtermism's assumption that, once the correct account of goodness is fixed, we are morally required to do whatever maximises that quantity in expectation. The consequentialism that underpins longtermism does not require this. It says only that the best outcomes are those that contain the greatest goodness. But that only gives the *ends* of moral action; it does not specify how morality requires us to pick the *means to those ends*. Rational choice theory is the area in which we study how to pick the best means to whatever are our ends. And, while expected utility theory is certainly one candidate account of the best means to our ends, since the middle of the twentieth century, there have been compelling arguments that it is mistaken. According to those arguments, we are permitted to take risk into account when we choose in a way that expected utility theory prohibits. For instance, we are permitted to be risk-averse and give more weight to worst-case outcomes and less weight to best-case outcomes than expected utility theory demands, and we are permitted to be risk-inclined and give more weight to best-case outcomes and less to worst-case ones than expected utility theory demands. An example: Suppose each extra unit of some commodity—quality-adjusted life years, perhaps—adds the same amount to my utility, so that that my utility is a linear function of the amount of this commodity I possess. Then most will judge it permissible for me to prefer to take 30 units of that commodity for sure rather

[1] The idea has a long history, running through the Einstein-Russell Manifesto and a thought experiment described by Derek Parfit on the last few pages of *Reasons and Persons* (1984). It has been developed explicitly over the past decade by Beckstead (2013); Bostrom (2013); Ord (2020); Greaves & MacAskill (2021); MacAskill (2022).

than to take a gamble that gives me 50% chance of 100 units and a 50% chance of none, even though the expected utility of the gamble is greater. Risk-sensitive decision theories are designed to respect that judgment. I'll describe one of them in more detail in Section 2.

How does the argument for longtermism go if we use a risk-sensitive decision theory instead of expected utility theory to pick the best means to the ends that our version of consequentialism has specified? I'll provide a detailed account in Sections 1-2, but I'll give the broad picture here. Here, as there, I'll assume that our axiology is total human hedonism. That is, I'll assume the goodness of a state of the world is 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 ask whether the argument changes significantly if we specify a different axiology.

A future in which humanity does not go extinct in the coming century from something like a meteor strike or biological warfare might contain vast quantities of great happiness and human flourishing. But it might also contain vast quantities of great misery and wasted potential. Longtermists assume that their favoured interventions will increase the probability of the long happy future more than they will increase the probability of the long miserable future. There are a couple of routes to this conclusion. First, they might inductively infer that the historical trend towards greater total human well-being will continue, as a result of an increasing population as well as increasing average well-being, and so assume that the long happy future is currently more likely than the long miserable future; and then they might assume further that any intervention that reduces the probability of extinction will increase the probabilities of the long happy and long miserable futures in proportion to their current probabilities. Secondly, they always combine their attempts to prevent extinction with attempts to improve whatever future lives exist—that is, they might not only try to reduce the probability of extinction, but also try to decrease the probability of a miserable future conditional on there being a future at all. Either way, this ensures that, in expectation, the longtermist's intervention is better than the status quo.

However, by the lights of risk-sensitive decision theories, these considerations do not ensure that a longtermist intervention is the best means to the longtermist's goal. Indeed, for a mildly risk-averse decision theory, it is not. In fact, for a risk-averse decision theory coupled with the axiology the longtermist favours, the best means to their avowed end might be to hasten rather than prevent extinction. Since a risk-averse decision theory gives greater weight to the worst-case outcomes and less weight to the best-case outcomes than expected utility theory demands, and since the long miserable future is clearly the worst-case outcome and the long happy one the best-case, it can easily be that, when we use a risk-averse decision theory, the negative effect of the increase in the probability of the long miserable future is sufficient to swamp the positive effect of the increase in the probability of the long happy future. Such a risk-averse decision theory might declare that increasing the probability of extinction is a better means to our end than either preserving the status quo or devoting resources to the longtermist's intervention, since that reduces the probability of the long miserable future, even though it also reduces the probability of the long happy one.

So much for what a risk-averse decision theory tells us to do when it is combined with the longtermist's axiology. What reason have we for thinking that this combination determines the morally correct choice between different options? Even if we agree that such decision theories govern prudentially rational choice for some individuals, we may nonetheless think that none of them governs moral choice; we may think that some other decision theory does that, such as expected utility theory. There are at least two views on which the morally right choice for an individual is the one demanded by a risk-averse decision theory when combined with the longtermist's axiology.

On the first, for a given individual, the same decision theory governs prudential choice and moral choice. What distinguishes those sorts of choice is only the utility function you feed into the decision theory to obtain its judgment. Prudential choice requires an individual to use the decision theory that matches their attitudes to risk, and then apply it in combination with their subjective utility function. And moral choice requires them to use the same decision theory, but this time combined with a utility function that represents the correct moral axiology, such as, perhaps, a utility function that measures total human hedonic value. So, in conjunction with the conclusions of previous paragraphs, we see that, for any sufficiently risk-averse individual, the morally correct choice for that individual is not to devote resources to the longtermist intervention, but rather to hasten extinction.

On the second view, for all individuals, the same decision theory governs moral choice, and it may well be different from the one that governs any particular individual's prudential choices. This decision theory is the one that matches what we might think of as the aggregate of the attitudes to risk held by the population who will be affected by the choice in question, perhaps with particular weight given to the risk-averse members of that population. Since most populations are risk-averse on the whole, this aggregate of their attitudes to risk will likely be quite risk-averse. So, for any individual, whether they are themselves risk-averse, the morally correct choice is to hasten extinction rather than prevent it, since that is what is required by the risk-averse decision theory that matches the population's aggregate attitudes to risk.

In the remainder of the paper, I make these considerations more precise and answer objections to them.

1 A simple model of the choice between interventions

Let me begin by introducing a simple model of the decision problem we face when we choose how to commit some substantial amount of money to do good. I will begin by using this simple precise model to raise my concern about the recommendations currently made by longtermists. In Sections 4.1 and 4.2, I will consider ways in which we might change the assumptions it makes, and ask whether doing so allows us to evade my concern.

Let's assume 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.

(*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.[2] 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 lives are happy, some mediocre, some only just worth living, many are miserable. In another, they live less long, but at a higher average level of happiness. And so on.

[2] Throughout, I will take 1 billion to be 10^9.

(*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 interventions, *SQ*, *HFF*, and *QEF*, we must assign probabilities to each of the states conditional on choosing that intervention.

First, the utilities. They measure the goodness of the state of the world. For simplicity, I will assume a straightforward total human hedonist utilitarian account of this goodness. 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. Again, this is a specific assumption made in order to provide concrete numbers for our model. In Section 4.1, I will ask what happens if we use different accounts of the goodness of a state of the world, including accounts that includes non-human animals and non-hedonic sources of value; and I will ask what happens to my argument if we use different estimates for the quantities involved here.

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. 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*. 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 that 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 ($\frac{1}{10^7}$). As I mentioned above, 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

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

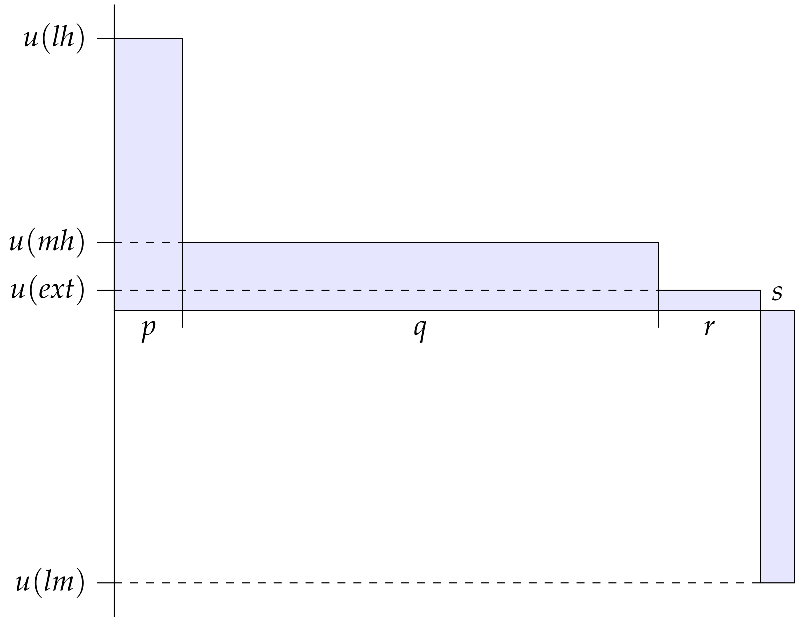

Figure 1: The expected utility of the status quo *SQ* is given by the grey area, where p is the probability of *lh*, q is the probability of *mh*, r is the probability of *ext*, and s is the probability of *lm* (any area below the zero line counts negatively). Not to scale!

	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*) is greater than the expected utility of the status quo (*SQ*), which is greater than the expected utility of donating to the Quiet End Foundation (*QEF*). 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 ex-

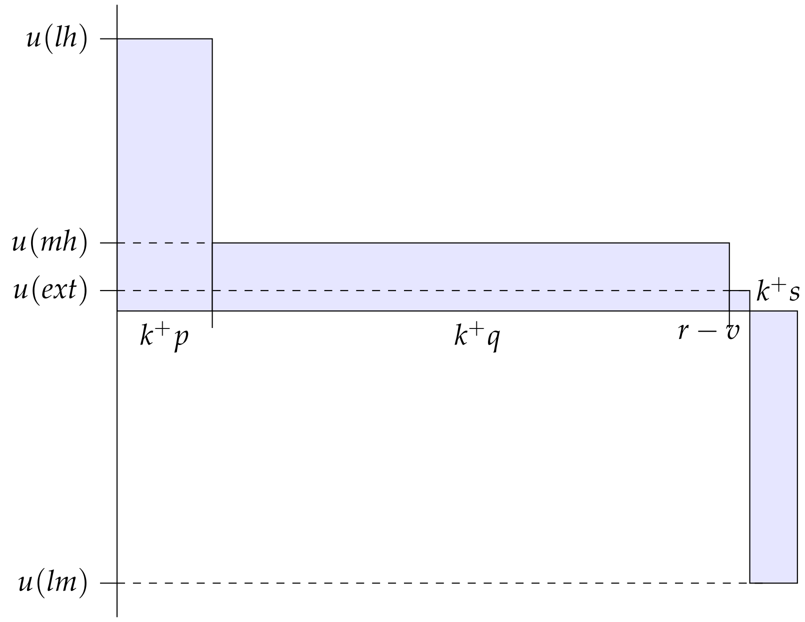

Figure 2: The expected utility of *HFF* is given by the grey area, where k^+p is the probability of *lh* given that you donate to the Happy Future Fund, k^+q is the probability of *mh* given that, $r-v$ is the probability of *ext* given that, and k^+s is the probability of *lm* given that. So v is the 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!

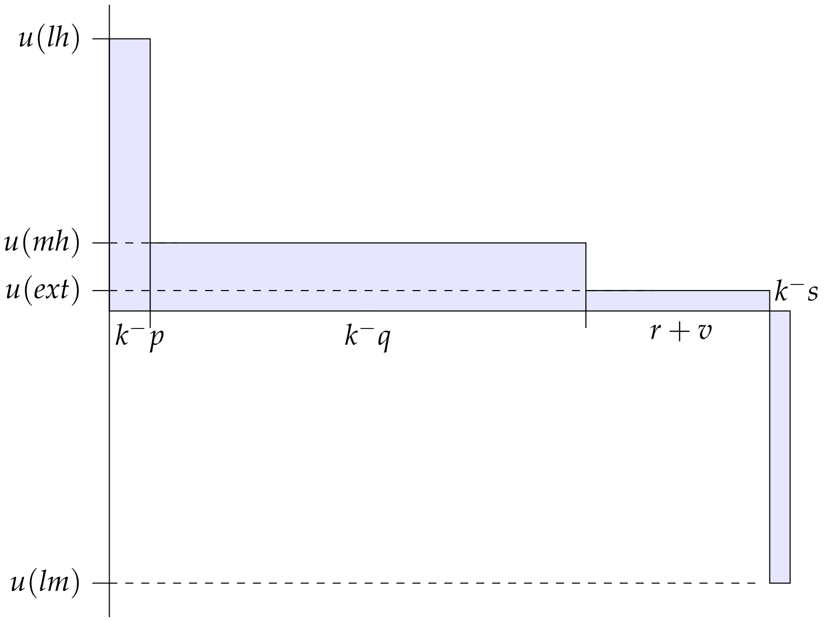

Figure 3: The expected utility of *QEF* is given by the grey area, where k^-p is the probability of *lh* given that you donate to the Quiet End Foundation, k^-q is the probability of *mh* given that, $r+v$ is the probability of *ext* given that, and k^-s is the probability of *lm* given that. So v is 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!

pected 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 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 longtermist's assumption 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, the Happy Future Fund is by far the best, then the status quo, and then the Quiet End Foundation. According to the longtermist, we should do whatever maximises expected utility. And so we should donate to the Happy Future Fund.

2 Rational choice theory and risk

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 classical utilitarian, who takes the goodness of a world to be the total human hedonic good that exists at that world, 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.

Utilitarianism, and indeed consequentialism more generally, supplies us with an axiology. It tells us how much goodness each possible state of affairs contains. And it tells 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. But it does not tell us what the morally *right* action is for an individual who is uncertain about what states their actions will bring about. To supply that, we must combine consequentialism with an account of decision-making under uncertainty. As I put it above, consequentialism provides the ends of moral action; but it says nothing about 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 longtermists, should follow their lead.

Consider the following example.[3] Sheila is a keen birdwatcher. Every time she sees a new species, it gives her great pleasure. What's more, the amount of extra pleasure each new species brings is the same no matter many she's seen before. Her first species—a blue tit in her grandparents' garden as a child—adds as much happiness to her stock as her two hundredth—a golden eagle high above Glenshee when she's thirty. And Sheila is a hedonist who cares only for pleasure. Now suppose she is planning a birding trip for her birthday, and she must choose between two nature reserves: in one, Shapwick Heath, she's sure to see 49 new species; in the other, Leighton Moss, she'll see 100 if the migration hasn't started and none if it has. And she's 50% confident that it has started. Here's the payoff table for her choice (with one utile per bird seen):

	Migration has started	Migration hasn't started
Shapwick Heath	49	49
Leighton Moss	0	100

According to expected utility theory, Sheila should choose to go to Leighton Moss, since, if each new species adds a single utile to an outcome, that option has an expected utility of 50 utiles, while Shapwick Heath has 49. And yet it seems quite rational for her to choose Shapwick Heath. In that way, she is assured of seeing some new species; indeed, she's assured of seeing quite a lot of new species; she does not risk seeing none, which she does risk if she goes to Leighton Moss. If Sheila chooses to go to Shapwick Heath, we might say that she is risk-averse, though perhaps only slightly. Leighton Moss is a risky option: it gives the possibility of the best outcome, namely, the one in which she sees 100 new species, but it also opens the possibility of the worst outcome, namely, the one in which she sees none. In contrast, Shapwick Heath is a risk-free option: it gives no possibility of the best outcome, but equally no possibility of the worst one either; it guarantees Sheila a middle-ranked option; its worst-case outcome, which is just its guaranteed outcome of 49 species is better than the worst-case outcome of Leighton Moss, which is seeing no species; but its best-case outcome, which is again its guaranteed outcome of seeing 49 species, is worse than the best-case outcome of Leighton Moss.

Standard expected utility theory says that the weight that each outcome receives before they are summed to give the expected utility of an option

[3] 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).

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 outcomes from best to worst. The risk-averse agent will wish to give greater weight to worse case outcomes than expected utility theory requires and less weight to the better case outcomes, while 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) and it is called *risk-weighted expected utility theory*. Whereas expected utility theory tells you to pick an option that maximises the expected utility from the point of view of your subjective probabilities and utilities, risk-weighted expected utility theory tells you to pick an option that maximises the risk-weighted expected utility from the point of view of your subjective probabilities, utilities, and attitudes to risk. Let's see how we represent these attitudes to risk and how we define risk-weighted utility in terms of them.

Your expected utility for an option is the sum of the utilities you assign to its outcome at different possible states of the world, each weighted by the probability you assign to that possible state on the supposition that you choose the option. Your risk-weighted expected utility of an option is also a weighted sum of your utilities for it given the different possible states of the world, but the weight assigned to its utility at a particular state of the world is determined not by your probability for that state of the world given you choose it, but by the probability you'll receive at least that much utility by choosing that option, the probability you'll receive more than that utility by choosing that option, and also your attitude to risk.

Here's how it works 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.

Now, to illustrate how risk-weighted expected utility theory works, 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, then S_2, 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

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

Easily the clearest way to understand how Buchak's theory works is by considering the following diagrams. In Figure 8, 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 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 triangle 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.

Roughly speaking, 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 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

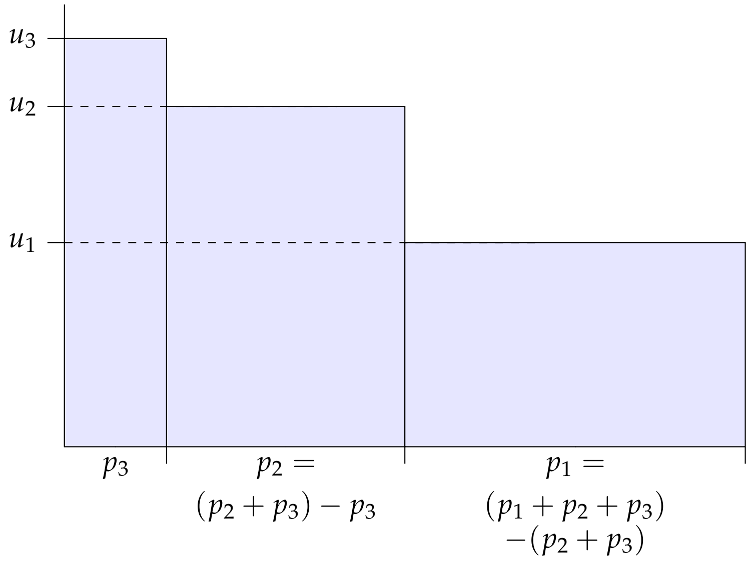

Figure 4: The expected utility of *O* is given by the grey area.

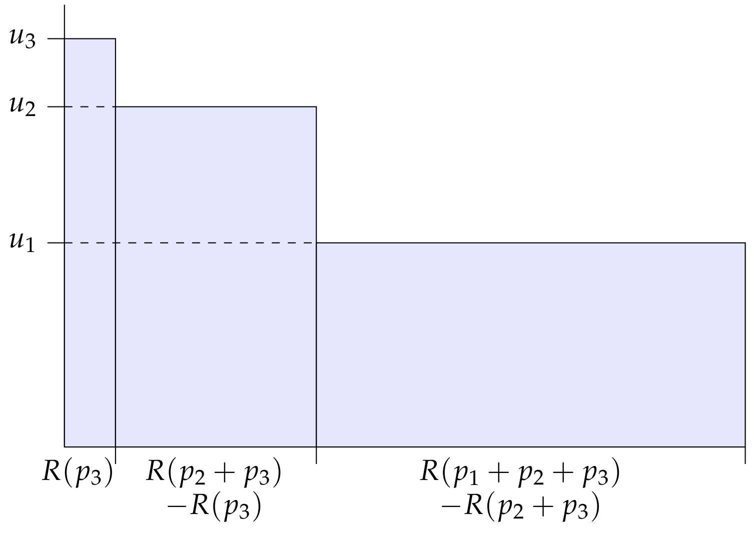

Figure 5: The risk-weighted expected utility of *O* is given by the grey area.

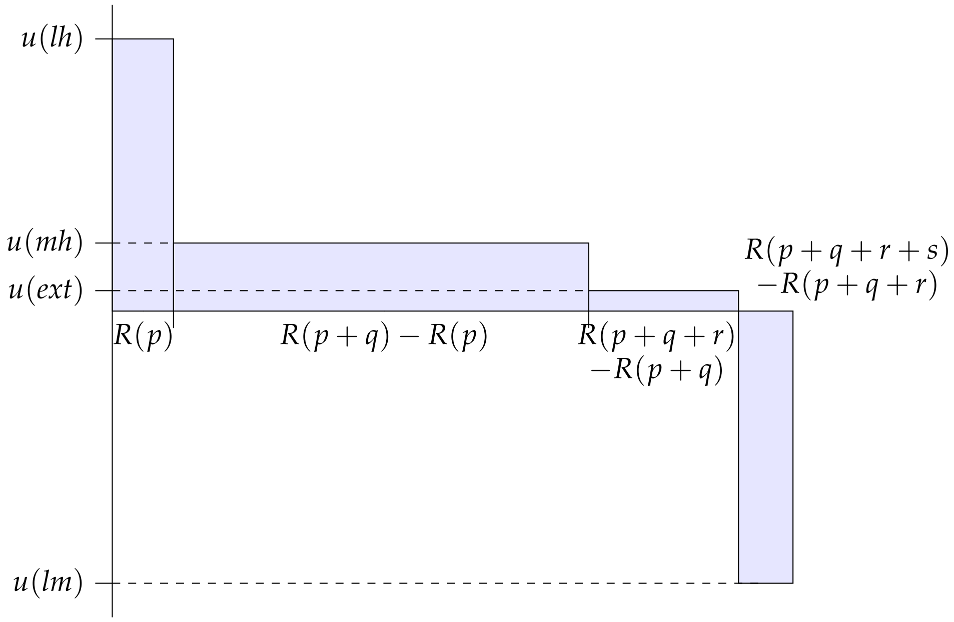

Figure 6: The risk-weighted expected utility of the status quo *SQ* is given by the grey area, where p is the probability of *lh*, q is the probability of *mh*, r is the probability of *ext*, and s is the probability of *lm*. Not to scale!

that $R(x) = x$—then the risk-weighted expected utility of an option is just its expected utility, so the individual is risk-neutral.

To see an example at work, consider Sheila's decision whether to go to Shapwick Heath or Leighton Moss. If R is Sheila's risk function, then the risk-weighted utility of going to Shapwick Heath is the sum of (i) its worst-case utility (i.e. 49) weighted by $R(1) - R(1/2)$ and (ii) its best-case utility (i.e. 49) weighted by $R(1/2)$, which is 49. And in general the risk-weighted utility of an option that has the same utility in every state of the world is just that utility. On the other hand, the risk-weighted utility of visiting Leighton Moss is the sum of (i) its worst-case utility (i.e. 0) weighted by $R(1) - R(1/2)$ and (ii) its best case utility (i.e. 100) weighted by $R(1/2)$, which is $R(1/2) \times 100$. So, providing $R(1/2) < 49/100$, the rational option for Sheila is the safe option, namely, Shapwick Heath. $R(x) = x^k$ for $k > 1.03$ will do the trick.

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, you have the level of risk aversion that would lead you, in Sheila's situation, to prefer a guarantee of seeing 35 new species of bird to a 50% chance of seeing 100 new species and a 50% chance of seeing none, or a guarantee of seeing 72 new species to an 80% chance of seeing 100, or

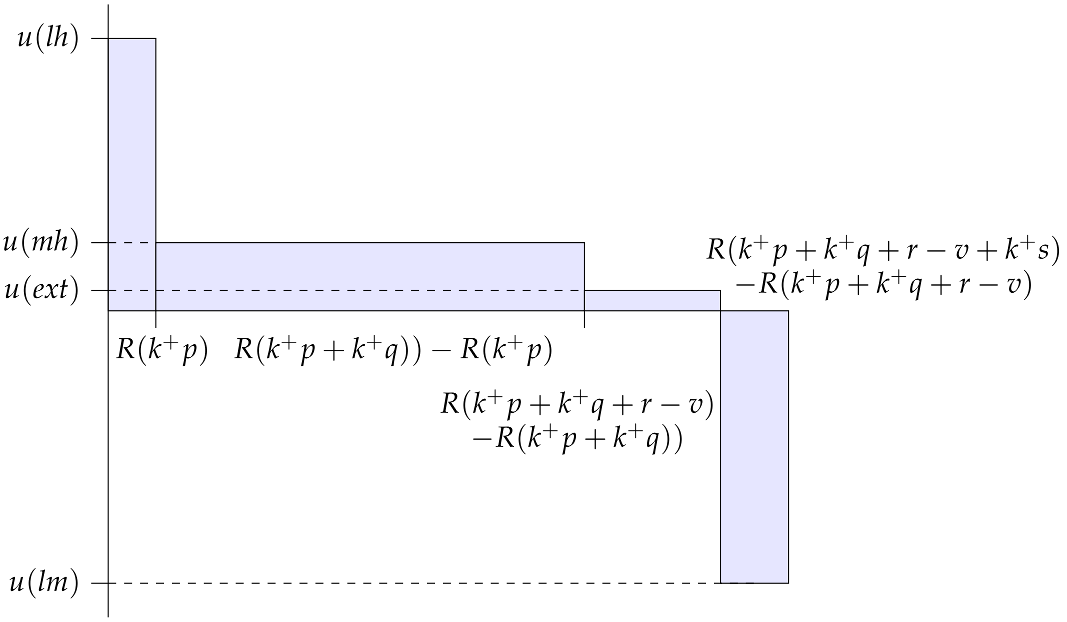

Figure 7: The risk-weighted expected utility of *HFF* is given by the grey area, where k^+p is the probability of *lh* given that you donate to the Happy Future Fund, k^+q is the probability of *mh* given that, $r - v$ is the probability of *ext* given that, and k^+s is the probability of *lm* given 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!

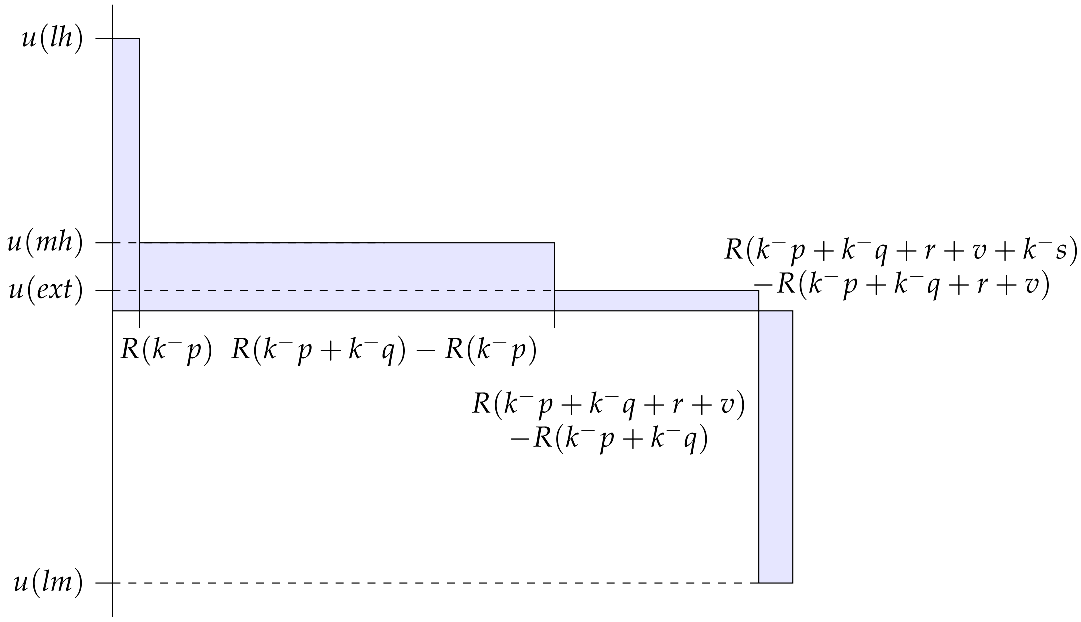
$u(lh)$
$u(mh)$
$u(ext)$
$u(lm)$
$R(k^- p)$
$R(k^- p + k^- q) - R(k^- p)$
$R(k^- p + k^- q + r + v)$
$-R(k^- p + k^- q)$
$R(k^- p + k^- q + r + v + k^- s)$
$-R(k^- p + k^- q + r + v)$

Figure 8: The risk-weighted 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!

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. So, if we replace expected utility theory with risk-weighted expected utility, then the effective altruist must give different advice to individuals with different attitudes to risk. And indeed the advice they give to mildly risk-averse individuals like the one just described will be 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 where the utility function measures the goodness that our moral axiology specifies rather than 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 this is not only what the longtermist should say to the risk-averse in our society, but what they should say to everyone, whether risk-averse, risk-inclined, or risk-neutral. In this section, I'll try to explain why.

To motivate the central principle used in the argument, consider the following case. A group of hikers make an attempt on the summit of a high, snow-covered mountain.[4] The route they have chosen is treacherous and they rope up, tying themselves to one another in a line so that, should one of them slip, the other will be able to prevent a dangerous fall. At one point in their ascent, the leader faces a choice. She is at the beginning of a particularly treacherous section—to climb up it is dangerous, but to climb down once you've started is nearly impossible. She also realises that she's at the point at which the rope will not provide much security, and will indeed endanger the others roped to her: if she falls while attempting this section, the whole group will fall with her, very badly injuring themselves. Due to changing weather, she must make the choice before she has a chance to consult with the group. Should she continue onwards and give the group the opportunity to reach the summit but also leave them vulnerable to serious injury, or should she begin the descent and lead the whole group down to the bottom safely?

[4] Thanks to Philip Ebert for helping me formulate this example!

She has climbed with this group for many years. She knows that each of them values getting to the summit just as much as she does; each disvalues severe injury just as much as she does; and each assigns the same middling value to descending now, not attaining the summit, but remaining uninjured. You might think, then, that each member of the group would favour the same option at this point—they'd all favour ascending or they'd all favour descending. But of course it's a consequence of Buchak's theory that they might all agree on the utilities and the probabilities, but disagree on what to do because they have different attitudes to risk. In fact, three out of the group of eight are risk-averse in a way that makes them wish to descend, while the remaining five wish to continue and accept the risk of injury in order to secure the possibility of attaining the summit. The leaders knows this. What should she do?

It seems to me that she should descend. 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 serious 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. 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 obtained by aggregating the risk attitude of all the people who will be affected by the decision in some way that gives most weight to the attitudes of the risk-averse.

To the best of my knowledge, Lara Buchak is the first to try to formulate a general principle that covers such situations. Here it is:[5]

> **Risk Principle** When making a decision for an individual, choose under the assumption that he has the most risk-avoidant attitude within reason unless we know that he has a different risk attitude, in which case, choose using his risk attitude. (Buchak, 2017, 632)

Here, as when she applies the principle to a different question about long-term global priorities in her 2019 Parfit Memorial Lecture, Buchak draws a distinction between rational attitudes to risk and reasonable ones. All reasonable attitudes will be rational, but not vice versa. So very extreme risk-aversion or extreme risk-inclination will count as rational, but perhaps not reasonable, just as being indifferent to pain if it occurs on a Tuesday, but not if it occurs at another time might be thought rational but not reasonable (Parfit, 1984, 124). Buchak then suggests that, when we do not know the risk attitudes of a person for whom we make a decision, we should make

[5] Cf. also (Rozen & Fiat, ms).

that decision using the most risk-averse attitudes that are reasonable, even if there are more risk-averse attitudes that are rational.

As the example of the climbers above illustrates, I think Buchak's principle has a kernel of truth. But I think we must amend it in various ways; and we must extend it to cover those cases in which (i) we choose not just for one individual but for many, and (ii) where our choice will affect different populations depending on how things turn out.

First, Buchak's principle divides the cases into only two sorts: those in which you know the person's risk attitudes and those in which you don't. It says: if you know them, use them; if you don't, use the most risk-averse among the reasonable attitudes. But of course you might know *something* about the other person's risk attitudes without knowing *everything*. For instance, you might know that they are risk-inclined, but you don't know to what extent; so you know that their risk function is concave, but you don't know which specific concave function it is. In this case, it seems wrong to use the most risk-averse reasonable risk attitudes to make your choice on their behalf. You know for sure the person on whose behalf you make the decision isn't so risk-averse as this, and indeed isn't risk-averse at all! So we might amend the principle so that we use the most risk-averse reasonable attitudes among those that our evidence doesn't rule out them having.

But even this seems too strong. You might have extremely strong but not conclusive evidence that the person affected by your action is risk-inclined; perhaps your evidence doesn't rule out that they have the most risk-averse reasonable risk attitudes, but it does make that very very unlikely. So you don't *know* that they are risk-inclined, but you've got very good reason for thinking they are; and you don't *know* that they do not have the most risk-averse reasonable risk attitudes, but you've got very good reason for thinking they don't. In this case, Buchak thinks you should nonetheless use the most risk-averse reasonable risk attitudes when you choose on their behalf. But this seems far too strong to me. It seems that you should certainly give greater weight to the more risk-averse attitudes among those you think they might have than the evidence seems to suggest; but you should not completely ignore your evidence. Having very strong evidence that they are risk-inclined, you should choose on their behalf using risk attitudes that are less risk-averse than those you'd use if your evidence strongly suggested that they are risk-averse, for instance, and less risk-averse than those you'd use if your evidence that they are risk-inclined was weaker. So evidence does make a difference, even when it's not conclusive.

Buchak objects to this approach as follows:

> When we make a decision for another person, we consider what no one could fault us for, so to speak [...] [F]inding out that

> a majority of people would prefer chocolate [ice cream] could give me reason to choose chocolate for my acquaintance, even if I know a sizable minority would prefer vanilla; but in the risk case, finding out a majority would take the risk could not give me strong enough reason to choose the risk for my acquaintance, if I knew a sizable minority would not take the risk. Different reasonable utility assignments are on a par in a way that different reasonable risk assignments are not: we default to risk avoidance, but there is nothing to single out any utility values as default. (Buchak, 2017, 631-2)

In fact, I think Buchak is right about the case she describes, but only because she specifies that there's a *sizable* minority that would not take the risk. But, as stated, her principle entails something much stronger than this. Even if there were only a one in a million chance that your acquaintance would reject the risk, the Risk Principle entails that you should not choose the risk on their behalf. But that seems too strong. And, in this situation, even if they did end up being that one-in-a-million person who is so risk-averse that they'd reject the risk, I don't think they could find fault with your decision. They would disagree with it, of course, and they'd prefer you chose differently, but if they know that you chose on their behalf by appealing to your very strong evidence that their risk-attitudes were not the most risk-averse reasonable ones, I think it would be strange for them to find fault with that decision. So I think Buchak is wrong to think that we *default* to risk-aversion; instead, the asymmetry between risk-aversion and risk-inclination is that more risk-averse possibilities and individuals should be given greater weight than more risk-inclined ones.

As it is stated, Buchak's principle applies only when you are making a decision for an individual, rather than for a group. And there again, I think the natural extension of the principle should be weakened. It seems that we do not consider immoral any decision on behalf of a group that goes against the preferences of the most risk-averse reasonable person possible, or even against preferences of the most risk-averse reasonable person in that group. For instance, it seems perfectly reasonable to be so risk-averse that you think the dangers of nuclear power outweigh the benefits, and yet it is morally permissible for a policymaker to pursue the project of building nuclear power stations because, while they give extra weight to the more risk-averse in the society affected, that isn't sufficient to outweigh the preferences of the vast majority who think the benefits outweigh the dangers.

Another crucial caveat to Buchak's principle is that it seems to apply only to decisions in which there is a risk of harm. Suppose that I must choose, on behalf of myself and my travelling companion, where we will go for a holiday. There are two options: Budapest and Bucharest. I know that going to Budapest will be very good, while I don't know whether Bucharest will be good or absolutely wonderful, but I know those are the possibilities. Then the risk-averse option is Budapest, and yet even if my travelling companion is risk-averse while I am risk-inclined, it seems that I do nothing morally wrong if I choose the risky option of Bucharest as our destination. And the reason that this is permissible is that none of the possible outcomes involves any harm—the worst that can happen is that our holiday is merely good rather than very good or absolutely wonderful.

Finally, in many decisions, there is a single population who will be affected by your actions regardless of how the world turns out, and in those cases, it is of course the risk attitudes of the people in that population you should aggregate to give the risk attitudes you'll use to make the decision on their behalf, weighting the more risk-averse more, as I've argued. But in some cases, and for instance in the choice between the Quiet End Foundation and the Happy Future Fund, different populations will be affected depending on how the world turns out: the world will contain different people in the four situations *lh*, *mh*, *ext*, and *lm*. How then are we to combine uncertainty about the population affected with information about the distribution of risk attitudes among those different possible populations? I think this is going to be a difficult question in general, just as it'll be difficult to formulate principles that govern situations in which there's substantial uncertainty about the distribution of risk attitudes in the population affected, but I think we can say one thing for certain: suppose it's the case that, for any of the possibly affected populations, were they the only population affected, you'd not choose the risky option; then, in that case, you shouldn't choose the risky option when there's uncertainty about which population will be affected.

Bringing all of this together, let's try to reformulate Buchak's risk principle:

Risk Principle*

(i) *Choosing on behalf of an individual when you're uncertain about their risk attitudes* When you make a decision on behalf of another person that might result in harm to that person, you should use a risk attitude obtained by aggregating the risk attitudes that your evidence says that person might have. And, when performing this aggregation, you should pay attention to how likely your evidence makes it that they have each possible risk attitude, but you should also give greater weight to the more risk-averse attitudes and less weight to the more risk-inclined ones than the evidence suggests.

(ii) *Choosing on behalf of a group when there's diversity of risk attitudes among its members* When you make a decision on behalf of a group of people that might result in harm to the people in that group, you should use a risk attitude obtained by aggregating the risk attitudes that those people have. And, when performing this aggregation, you should give greater weight to more risk-averse individuals in the group.

(iii) *Choosing on behalf of a group when there's uncertainty about the risk attitudes of its members either because a single population is affected but you don't know the distribution of risk attitudes within it, or because you don't know which population will be affected* When you make a decision on behalf of a group of people that might result in harm to the people in that group, and you are uncertain about the distribution of the risk attitudes in that group, then you should work through each of the possible populations with their distributions of risk attitudes in turn, perform the sort of aggregation we saw in (ii) above, then take each of those aggregates and aggregate those, this time paying attention to how likely your evidence makes each of the populations they aggregate, but also giving more weight to the more risk-averse aggregates and less weight to the more risk-inclined ones than the evidence suggests.

Like Buchak's, this version is not fully specified. For Buchak's, that was because the notion of reasonable risk attitudes remained unspecified. For this version, it's because we haven't said precisely how to aggregate risk attitudes nor how to determine exactly what extra weight an attitude receives in such an aggregation because it is risk-averse. I will leave the principle underspecified in this way, but let me quickly illustrate the sort of aggregation procedure we might use. Suppose we have a group of n individuals and their individual risk attitudes are represented by the Buchakian risk functions $R_1, \ldots, R_n$. Then we might aggregate those individual risk functions to give the aggregate risk function that represents the collective risk attitudes of the group by taking a weighted average of them: that is, the risk function R_G of the group is $R_G = \lambda_1 R_1 + \ldots \lambda_n R_n$ for some weights $\lambda_1, \ldots, \lambda_n$, each of which is non-negative and which together sum to 1. Then we might ensure that λ_i is greater the more risk averse (and thus convex) R_i is.

In any case, underspecified though the Risk Principle* is in various ways, I think it's determinate enough to pose the problem I want to pose. Many people are quite risk averse; indeed, the empirical evidence suggests that most are (MacCrimmon & Larsson, 1979; Rabin & Thaler, 2001; Oliver, 2003). We should expect that to continue into the future. So each of the possible populations affected by my choice of where to donate my money—that is, the populations that inhabit scenarios *lh*, *mh*, *ext*, and *lm*, respectively—are likely to include a large proportion of risk-averse individuals. And so the third clause of the Risk Principle*—that is, (iii)—might well say that I should choose on their behalf using an aggregated risk function that is pretty risk-averse, and perhaps sufficiently risk-averse that it demands we donate to the Quiet End Foundation rather than the Happy Future Fund or the Against Malaria Foundation or whatever other possibilities there are.

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 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 that, 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 that, for everyone regardless of attitudes to risk, the morally correct choice is to donate in that way. Here's the first in more detail:

(P1) The morally correct choice for you is the one required by the correct decision theory when that theory is applied using certain attitudes of yours and certain attitudes set by morality.

(P2) The correct decision theory is risk-weighted expected utility theory.

(P3) When you apply risk-weighted expected utility theory in ethics, you should use your own credences and risk attitudes, providing they're rational and reasonable, but you should use the moral utilities, which measure the morally relevant good.

(P4) Given your current evidential and historical situation, if you are moderately risk-averse, you maximise risk-weighted expected moral utility by choosing to donate to the Quiet End Foundation rather than by doing nothing or donating to the Happy Future Fund.

Therefore,

(C) If you are even mildly risk-averse, the morally correct choice for you is to donate to the Quiet End Foundation.

And the second:

(P1) The morally correct choice for you is the one required by the correct decision theory when that theory is applied using certain attitudes of yours and certain attitudes set by morality.

(P2) The correct decision theory is risk-weighted expected utility theory.

(P3′) When you apply risk-weighted expected utility theory in ethics, you should use your own credences, providing they're reasonable and rational, but you should use the utilities specified by moral axiology, and you should use risk attitudes obtained by aggregating actual and possible risk attitudes in the populations affected in line with the Risk Principle*.

(P4′) Given your current evidential and historical situation, you maximise the risk-weighted expected moral utility by choosing to donate to the Quiet End Foundation rather than by doing nothing or donating to the Happy Future Fund.

Therefore,

(C′) Whether you are risk-averse or not, the morally correct choice for you is to donate to the Quiet End Foundation.

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 9 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.

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

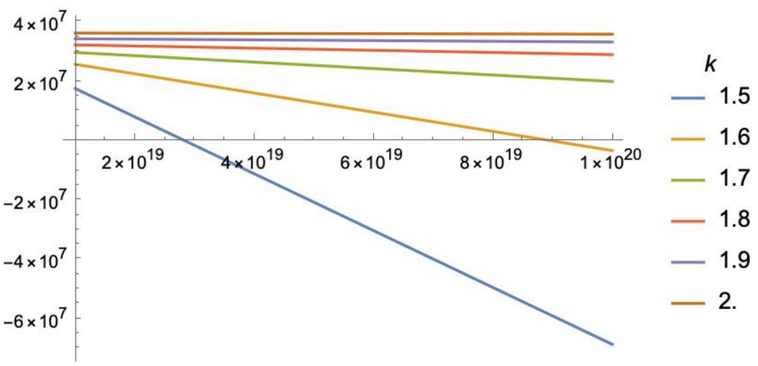

Figure 9: This plots $REU(QEF) - REU(HFF)$ as the utility of lh increases from 10^{19} to 10^{20}, for different risk functions of the form $R_k(x) = x^k$. So, for risk function $R_{1.5}$ (the blue line), the Quiet End Foundation is better than the Happy Future Fund providing whatever extra non-hedonic 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 yellow line), QEF is better than HFF, even if the non-hedonic goodness multiplies the utility of lh by a factor of 8.

ligent 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.

The same is true if we appeal to obligations that we have to those who have lived before us (Baier, 1981). At this point, we step outside the consequentialist framework in which longtermist arguments are usually presented, and into a deontological framework. But we might marry these two approaches and say that obligations rule out certain options from the outset and then consequentialist reasoning enters to pick between the remaining ones. Here, we might think that past generations created much of what they did and fought for what they did and bequeathed to us the fruits of their labours and their sacrifices on the understanding that humanity would continue to exist. And you might think that, by benefitting from what they bequeathed to us—those goods for which they laboured and which they made sacrifices to obtain—we take on an obligation not to go against their wishes and bring humanity to an end. But, as before, I think what they really wished was that humanity continue to exist *in a way they considered positive*. And so obligations to them don't rule prohibit ending humanity if by doing so you avoid a universally miserable human existence.

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 $\frac{1}{6}$. Users of the opinion aggregator *Metaculus* currently place it at $\frac{1}{50}$.[6] How do these alternative probabilities affect our calculation? Figure 10 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 $\frac{1}{10^5}$—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 11 and 12 illustrates the effects.

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

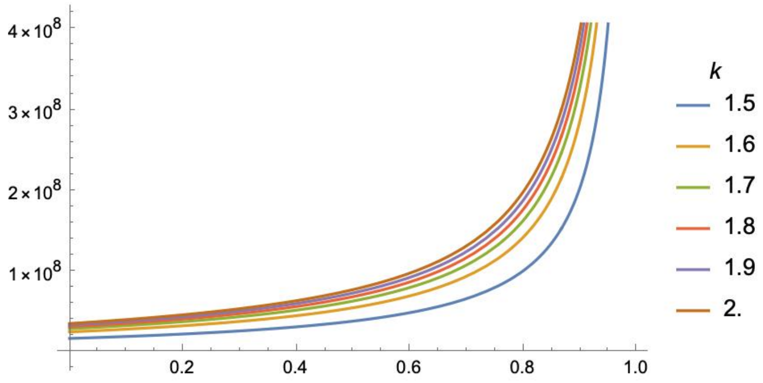

Figure 10: This plots $REU(QEF) - REU(HFF)$ as the probability of *ext* ranges from 0 to 1, for different risk functions of the form $R_k(x) = x^k$. For each risk function, and for all probabilities for extinction, the Quiet End Foundation is better than the Happy Future Fund.

Again, our conclusion is robust.

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 remove 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 are 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 13 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 out *HFF* for up to nearly $n = 5$.

This is the first time we've anything less than robustness in our conclu-

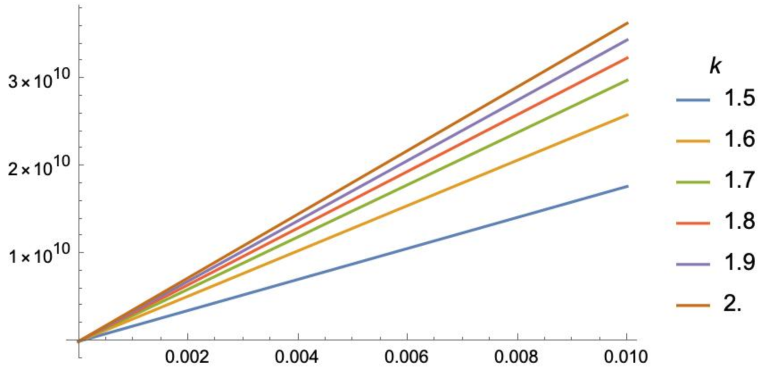

Figure 11: This plots $REU(QEF) - REU(HFF)$ as the change in the probability of extinction that our intervention can achieve ranges from $\frac{1}{10^7}$ to $\frac{1}{10^2}$, for different risk functions of the form $R_k(x) = x^k$.

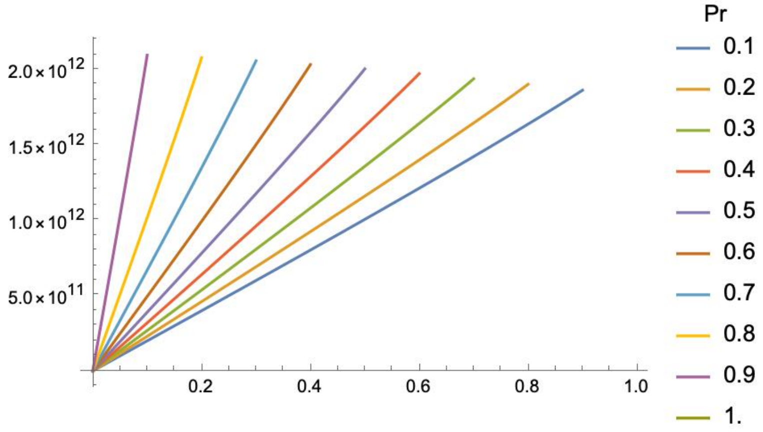

Figure 12: This plots $REU(QEF) - REU(HFF)$ as the change in probability our intervention can achieve ranges from $\frac{1}{10{,}000}$ to 1, for risk function $R_{1.5}$ and different value Pr for the prior probability of extinction.

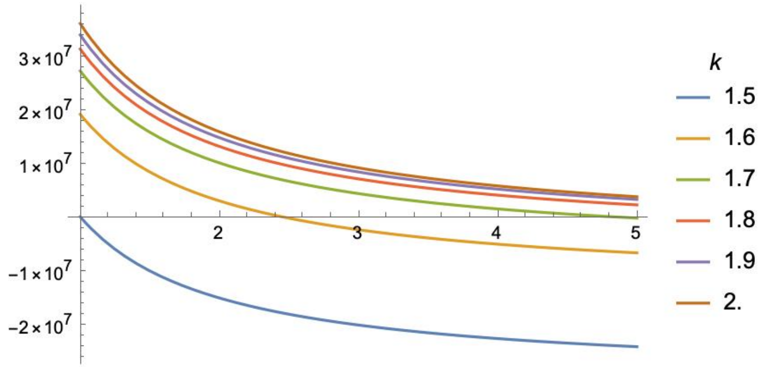

Figure 13: This plots $REU(QEF) - REU(HFF)$ as n ranges from 1 to 5, for risk functions of the form $R_k(x) = x^k$. Recall, the probability of lh given QEF is $2nl^-$, the probability of mh given QEF is nl^-, and the probability of lm given QEF is l^-. And their probabilities given HFF are the same, but with l^+ in place of l^-.

sion 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, since as degrees of risk-aversion. For instance, you might think that the lesson of Figure 13 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*.

4.3 Overturning unanimous preference

One apparent problem with letting risk-sensitive decision theory govern moral choice is that those choices may thereby end up violating the so-called *Ex Ante Pareto Principle*, which says that it's never morally right to choose one option over another when the second is unanimously preferred to the first by those affected by it.[7]

Suppose Ann is risk-neutral: that is, she values options at their expected utility. And suppose Bob is a little risk-averse: he value options at their risk-weighted expected utility with risk function $R_{1.4}(x) = x^{1.4}$. Both agree that the coin in my hand is fair, with a 50% chance of landing heads and a 50% chance of landing tails. I must make a decision on behalf of Ann and

[7] Thank you to Teru Thomas for pressing me to consider this objection.

Bob. Two options are available to me: A and B. If I choose A, Ann will face a gamble that leaves her with 12 utiles if the coin lands heads and 8 utiles if it lands tails, while Bob will be left with 0 utiles either way.[8] If I choose B, on the other hand, Bob will face the gamble instead, leaving him with 12 utiles if heads and 8 if tails, while Ann will be left with 1 utile either way. Their utilities are given in the following tables:

A	Heads	Tails
Ann	12	-8
Bob	0	0
Total	12	-8

B	Heads	Tails
Ann	1	1
Bob	12	-8
Total	13	-7

Then Ann prefers A to B, since A has an expected utility of 2 utiles for her, while B has an expected utility of 1 utile. And Bob prefers A to B, since A has risk-weighted utility of 0 utiles, while B has a risk-weighted expected utility of $(1 - ({}^{1}/{}_{2})^{1.4}) \times -8 + ({}^{1}/{}_{2})^{1.4} \times 12 \approx -0.42$. But note: the total utility of A is 12 is heads and -8 if tails, while the total utility of B is 13 if heads and -7 if tails. So, if we make our moral choices using risk-weighted expected utility theory combined with a utility function that measures the total utility obtained by summing Ann's utility and Bob's, we will choose B, which overturns their unanimous preferences.

At first sight, this seems worrying for our proposal that the morally correct choice is the one recommended by risk-weighted expected utility theory when combined with a utility function that measures morally relevant goodness. But, it's easy to see that there's nothing distinctive about our proposal that causes the worry. After all, *any* plausible decision theory, when coupled with an axiology that takes the value of an outcome to be the total utility present in the outcome, will choose B over A, for, in the jargon of decision theory, B strongly dominates A—that is, B is guaranteed to be better than A; however the world turns out, B is better than A; if the coin lands heads, B has greater total utility, and if the coin lands tails, B has the greater total utility. So it doesn't count against the Risk Principle* or the use of risk-weighted expected utility theory for moral choice that they lead to violations of the Ex Ante Pareto Principle. Any plausible decision theory will do likewise. Indeed, the longtermist's favoured theory, on which we should maximise expected total utility will favour B over A.

What is really responsible for the issue here is that we permit individuals to differ in their attitudes to risk. It is because Ann and Bob differ in these attitudes that we can find a decision problem in which they both prudentially prefer one option to another, while the latter option is guaranteed to give greater total utility than the former.[9] But we can't very well

[8] As always, we assume that the interpersonal utilities of Ann and Bob can be compared and measured on the same scale.

[9] Indeed, as Simon Blessenohl (2020) shows, all that is really required is that there is a pair

say that an individual's attitudes to risk are irrational or prohibited purely on the grounds that, by having them they ensure that maximising expected or risk-weighted expected total utility will lead to a violation of Ex Ante Pareto.

5 Conclusion

I've presented arguments for two conclusions, one stronger than the other: first, 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; secondly, whether or not you are averse to risk, this is what you should do. And I've considered some responses and I've tried to show that the arguments still stand in the light of them.

What, then, is the overall conclusion? I confess, I don't know. 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 effective altruism in which these arguments are given. 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.

Baier, A. (1981). The Rights of Past and Future Persons. In E. Partridge (Ed.) *Responsibilities to Future Generations: Environmental Ethics*, (pp. 171–83). New York: Prometheus Books.

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

Blessenohl, S. (2020). Risk Attitudes and Social Choice. *Ethics*, *130*, 485–513.

Bostrom, N. (2013). Existential Risk Prevention as Global Priority. Unpublished manuscript.

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

of options for which one individual prefers the first to the second while the other individual prefers the second to the first. And that might arise because of different attitudes to risk, but it might also arise because of different rationally permissible credences in the same proposition.

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

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

Hintze, A., Olson, R. S., Adami, C., & Hertwig, R. (2015). Risk sensitivity as an evolutionary adaptation. *Scientific Report*, *5*, 8242.

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

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.

Okasha, S. (2007). Rational Choice, Risk Aversion, and Evolution. *The Journal of Philosophy*, *104*(5), 217–35.

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.

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

Rozen, I. N., & Fiat, J. (ms). Attitudes to Risk when Choosing for Others. Unpublished manuscript.

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.

Zhang, R., Brennan, T. J., & Lo, A. W. (2014). The origin of risk aversion. *Proceedings of the National Academy of Science of the U.S.A.*, *111*(50), 17777–82.